Journal of Modern Applied Statistical Methods

Volume 4 | Issue 1 Article 34

5-1-2005

Vol. 4, No. 1 (Full Issue)

JMASM Editors

The easy way to find open access journals

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

www.doaj.org

The Directory of Open Access Journals covers free, full text, quality controlled scientific and scholarly journals. It aims to cover all subjects and languages.

Aims

- Increase visibility of open access journals
- Simplify use
- Promote increased usage leading to higher impact

Scope

The Directory aims to be comprehensive and cover all open access scientific and scholarly journals that use a quality control system to guarantee the content. All subject areas and languages will be covered.

In DOAJ browse by subject

Agriculture and Food Sciences
Biology and Life Sciences
Chemistry
General Works
History and Archaeology
Law and Political Science
Philosophy and Religion
Social Sciences
Arts and Architecture
Business and Economics
Earth and Environmental Sciences
Health Sciences
Languages and Literatures
Mathematics and statistics
Physics and Astronomy
Technology and Engineering

Contact
Lotte Jørgensen, Project Coordinator
Lund University Libraries, Head Office
E-mail: lotte.jorgensen@lub.lu.se
Tel: +46 46 222 34 31

Funded by
www.soros.org

Hosted by
LUND UNIVERSITY
www.lu.se

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

Editorial Board

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

Journal Of Modern Applied Statistical Methods

Invited Article

2 – 10 **Rand R. Wilcox** Within by Within Anova Based on Medians

Regular Articles

11 – 34 **Biao Zhang** Testing the Goodness of Fit of Multivariate Multiplicative-intercept Risk Models Based on Case-control Data

35 – 42 **Panagiotis Mantalos** Two Sides of the Same Coin: Bootstrapping the Restricted vs. Unrestricted Model

43 – 52 **John P. Wendell**, **Sharon P. Cox** Coverage Properties of Optimized Confidence Intervals for Proportions

53 – 62 **Rand R. Wilcox**, **Mitchell Earleywine** Inferences about Regression Interactions via a Robust Smoother with an Application to Cannabis Problems

63 – 74 **Stan Lipovetsky**, **Michael Conklin** Regression by Data Segments via Discriminant Analysis

75 – 80 **W. A. Abu-Dayyeh**, **Z. R. Al-Rawi**, **M. MA. Al-Momani** Local Power for Combining Independent Tests in the Presence of Nuisance Parameters for the Logistic Distribution

81 – 89 **B. Sango Otieno**, **C. Anderson-Cook** Effect of Position of an Outlier on the Influence Curve of the Measures of Preferred Direction for Circular Data

90 – 99 **Inger Persson**, **Harry Khamis** Bias of the Cox Model Hazard Ratio

100 – 105 **David A. Walker** Bias Affiliated with Two Variants of Cohen's d When Determining U_1 as A Measure of the Percent of Non-Overlap

106 – 119 **C. Anderson-Cook,** **Kathryn Prewitt** Some Guidelines for Using Nonparametric Methods for Modeling Data from Response Surface Designs

120 – 133 **Gibbs Y. Kanyongo** Determining the Correct Number of Components to Extract from a Principal Components Analysis: A Monte Carlo study of the Accuracy of the Scree Plot

134 – 139 **Abdullah Almasri**, **Ghazi Shukur** Testing the Casual Relation Between Sunspots and Temperature Using Wavelets Analysis

140 – 154 **Leming Qu** Bayesian Wavelet Estimation of Long Memory Parameter

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

155 – 162 **Kosei Fukuda** Model-Selection-Based Monitoring of Structural Change

163 – 171 **Lyle Broemeling**, **Dongfeng Wu** On the Power Function of Bayesian Tests with Application to Design of Clinical Trials: The Fixed-Sample Case

172 – 186 **Vincent Camara**, **Chris P. Tsokos** Bayesian Reliability Modeling Using Monte Carlo Integration

187 – 213 **Michael C. Long**, **Ping Sa** Right-tailed Testing of Variance for Non-Normal Distributions

214 – 226 **Hasan Hamdan**, **John Nolan**, **Melanie Wilson**, **Kristen Dardia** Using Scale Mixtures of Normals to Model Continuously Compounded Returns

227 – 239 **Michael B.C. Khoo**, **T. F. Ng** Enhancing the Performance of a Short Run Multivariate Control Chart for the Process Mean

240 – 250 **Paul A. Nakonezny**, **Joseph Lee Rodgers** An Empirical Evaluation of the Retrospective Pretest: Are There Advantages to Looking Back?

251 – 274 **Yonghong Jade Xu** An Exploration of Using Data Mining in Educational Research

275 – 282 **Bruno D. Zumbo**, **Kim H Koh** Manifestation of Differences in Item-Level Characteristics in Scale-Level Measurement Invariance Tests of Multi-Group Confirmatory Factor Analyses

Brief Report

283 – 287 **J. Thomas Kellow** Exploratory Factor Analysis in Two Measurement Journals: Hegemony by Default

Early Scholars

288 – 299 **Ling Chen**, **Mariana Drane**, **Robert F. Valois**, **J. Wanzer Drane** Multiple Imputation for Missing Ordinal Data

JMASM Algorithms and Code

300 – 311 **Hakan Demirtas** JMASM16: Pseudo-Random Number Generation in R for Some Univariate Distributions (R)

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

312 – 318 **Sikha Bagui**, **Subhash Bagui** JMASM17: An Algorithm and Code for Computing Exact Critical Values for Friedman's Nonparametric ANOVA (Visual Basic)

319 – 332 **J. I. Odiase**, **S. M. Ogbonmwan** JMASM18: An Algorithm for Generating Unconditional Exact Permutation Distribution for a Two-Sample Experiment (Visual Fortran)

333 – 342 **David A. Walker** JMASM19: A SPSS Matrix for Determining Effect Sizes From Three Categories: r and Functions of r, Differences Between Proportions, and Standardized Differences Between Means (SPSS)

Statistical Software Applications & Review

343 – 351 **Paul Mondragon**, **Brian Borchers** A Comparison of Nonlinear Regression Codes

Letter To The Editor

352 **Shlomo Sawilowsky** Abelson's Paradox And The Michelson-Morley Experiment

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/05/$95.00

JMASM is an independent print and electronic journal (http://tbf.coe.wayne.edu/jmasm) designed to provide an outlet for the scholarly works of applied nonparametric or parametric statisticians, data analysts, researchers, classical or modern psychometricians, quantitative or qualitative evaluators, and methodologists. Work appearing in *Regular Articles*, *Brief Reports*, and *Early Scholar*s are externally peer reviewed, with input from the Editorial Board; in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* are internally reviewed by the Editorial Board.

Three areas are appropriate for *JMASM*: (1) development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) development or study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods. Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Problems may arise from applied statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation. They should relate to the social and behavioral sciences, especially education and psychology. Applications from other traditions, such as actuarial statistics, biometrics or biostatistics, chemometrics, econometrics, environmetrics, jurimetrics, quality control, and sociometrics are welcome. Applied methods from other disciplines (e.g., astronomy, business, engineering, genetics, logic, nursing, marketing, medicine, oceanography, pharmacy, physics, political science) are acceptable if the demonstration holds promise for the social and behavioral sciences.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-10

1538 – 9472/05/$95.00

INVITED ARTICLE

Within By Within ANOVA Based On Medians

Rand R. Wilcox
Department of Psychology
University of Southern California, Los Angeles

This article considers a *J* by *K* ANOVA design where all *JK* groups are dependent and where groups are to be compared based on medians. Two general approaches are considered. The first is based on an omnibus test for no main effects and no interactions and the other tests each member of a collection of relevant linear contrasts. Based on an earlier paper dealing with multiple comparisons, an obvious speculation is that a particular bootstrap method should be used. One of the main points here is that, in general, this is not the case for the problem at hand. The second main result is that, in terms of Type I errors, the second approach, where multiple hypotheses are tested based on relevant linear contrasts, performs about as well or better than the omnibus method, and in some cases it offers a distinct advantage.

Keywords: Repeated measures designs, robust methods, kernel density estimators, bootstrap methods, linear contrasts, multiple comparisons, familywise error rate

Introduction

Consider a J by K ANOVA design where all JK groups are dependent. Let θ_{jk} (j =1,...J; k =1,...K) represent the (population) medians corresponding to these JK groups. This article is concerned with two strategies for dealing with main effects and interactions. The first is to perform an omnibus test for no main effects and no interactions by testing

$$Ho: C\theta = 0, \qquad (1)$$

where θ is a column vector containing the JK elements θ_{jk}, and C is an ℓ by JK matrix (having rank ℓ) that reflects the null hypothesis of interest. (The first K elements of θ are $\theta_{11},\ldots,\theta_{1K}$, the next K elements are $\theta_{21},\ldots,\theta_{2K}$, and so forth.) The second approach uses a collection of linear contrasts, rather than a single omnibus test, and now the goal is to control the probability of at least one Type I error.

A search of the literature indicates that there are very few results on comparing the medians of dependent groups using a direct estimate of the medians of the marginal distributions, and there are no results for the situation at hand. In an earlier article (Wilcox, 2004), two methods were considered for performing all pairwise comparisons among a collection of dependent groups. The first uses an estimate of the appropriate standard error stemming from the influence function of a single

Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology at the University of Southern California, Los Angeles.

order statistic. The second method uses the usual sample median in conjunction with a bootstrap estimate of the standard error. The bootstrap method performed quite well in simulations in terms of controlling the probability of at least one Type I error.

Recently, Dawson, Schell, Rissling and Wilcox (2004) dealt with an applied study where a two-way ANOVA design was used with all *JK* groups dependent. An issue is whether the results in Wilcox (2004) extend to this two-way design. One of the main results here is that the answer is no. The other main result deals with the choice between an omnibus test versus performing multiple comparisons where each hypothesis corresponding to a collection of relevant linear contrasts is to be tested. It is found that simply ignoring the omnibus test, and performing the relevant multiple comparisons, has practical value.

Some Preliminaries

For convenience, momentarily consider a single random sample $X_1,...,X_n$ and for any q, $0<q<1$, suppose the q[th] quantile, x_q, is estimated with $X_{(m)}$, where $m=[qn+.5]$ and [.] is the greatest integer function. Then, ignoring an error term, which goes to zero as $n \to \infty$,

$$X_{(m)} = x_q + \frac{1}{n}\sum IF_q(X_i), \qquad (2)$$

where

$$IF_q(x) = \begin{cases} \dfrac{q-1}{f(x_q)}, & \text{if } x < x_q \\ 0, & \text{if } x = x_q \\ \dfrac{q}{f(x_q)}, & \text{if } x > x_q, \end{cases}$$

(Bahadur, 1966; also see Staudte & Sheather, 1990).

Now consider the situation where sampling is from a bivariate distribution. Let X_{ik} (i=1,...,n; k=1, 2) be a random sample of n vectors. Let $X_{(1)k} \le \cdots \le X_{(n)k}$ be the observations associated with k[th] variable written in ascending order. Two estimates of the population median are relevant here. The first is

$$\hat{\theta}_k = X_{(m)k},$$

where again $m=[.5n+.5]$, and the other is $\hat{\theta}_j=M_k$, the usual sample median based on $X_{1k},...,X_{nk}$.

Although the focus is on estimating the median with q=.5, the results given here apply to any q, $0<q<1$. Let f_k be the marginal density of the k[th] variable and let

$$V_1 = (q-1)^2 P(X_1 \le x_{q1}, X_2 \le x_{q2}),$$
$$V_2 = q(q-1)P(X_1 \le x_{q1}, X_2 > x_{q2}),$$
$$V_3 = q(q-1)P(X_1 > x_{q1}, X_2 \le x_{q2}),$$

and

$$V_4 = q^2 P(X_1 > x_{q1}, X_2 > x_{q2}),$$

where x_{q1} and x_{q2} are the q[th] quantiles corresponding to the first and second marginal distributions, respectively. Then for the general case where $m=[qn+.5]$, a straightforward derivation based on equation (2) yields an expression for the covariance between $X_{(m)1}$ and $X_{(m)2}$:

$$\tau_{12}^2 = \frac{V_1+V_2+V_3+V_4}{nf_1(x_{q1})f_2(x_{q2})}. \qquad (3)$$

Also, (2) yields a well-known expression for the squared standard error of $X_{(m)1}$, namely,

$$\tau_{11}^2 = \frac{1}{n}\frac{q(1-q)}{f^2{}_1(x_{q1})}.$$

Using (3) to estimate τ_{12}^2 requires an estimate of the marginal densities. Here, a

variation of an adaptive kernel density estimator is used (e.g., Silverman, 1986), which is based in part on an initial estimate obtained via a so-called expected frequency curve (e.g., Wilcox, 2005; cf. Davies & Kovac, 2004). To elaborate, let MAD_k be the median absolute deviation associated with the kth marginal distribution, which is the median of the values $|X_{1k}-M_k|,...,|X_{nk}-M_k|$. For some constant κ to be determined, the point x is said to be close to X_{ik} if

$$|X_{ik}-x|\leq \kappa\times\frac{MAD_k}{.6745}.$$

Under normality, $\text{MADN}_k=MAD_k/.6745$ estimates the standard deviation, in which case x is close to X_{ik} if x is within κ standard deviations of X_{ik}. Let

$$N_k(x)=\{i:|X_{ik}-x|\leq \kappa\times MADN_k\}.$$

That is, $N_k(x)$ indexes the set of all X_{ik} values that are close to x. Then an initial estimate of $f_k(x)$ is taken to be

$$\tilde{f}_k(x)=\frac{1}{2\kappa MADN_k}\sum I_{i\in N_k(x)},$$

where I is the indicator function. Here, κ=.8 is used.

The adaptive kernel density estimate is computed as follows. Let

$$\log g=\frac{1}{n}\sum \log \tilde{f}_\kappa(X_i)$$

and

$$\lambda_i=(\tilde{f}_k(X_{ik})/g)^{-a},$$

where a is a sensitivity parameter satisfying $0\leq a\leq 1$. Based on comments by Silverman (1986), a=.5 is used. Then the adaptive kernel estimate of f_k is taken to be

$$\tilde{f}_\kappa(x)=\frac{1}{n}\sum \frac{1}{h\lambda_i}K\left\{h^{-1}\lambda_i^{-1}(x-X_i)\right\},$$

where

$$K(t) = \frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}, \quad |t|<\sqrt{5}$$
$$= 0, \quad \text{otherwise},$$

is the Epanechnikov kernel, and following Silverman (1986, p. 47 – 48), the span is

$$h=1.06\frac{A}{n^{1/5}},$$

$$A=min(s, IQR/1.34),$$

and where s is the standard deviation and IQR is the interquartile range based on $X_{1k},...,X_{nk}$.

Here, IQR is estimated via the ideal fourths. Let ℓ=[(n/4)+(5/12)]. That is, ℓ is (n/4)+(5/12) rounded down to the nearest integer. Let

$$h=\frac{n}{4}+\frac{5}{12}-l.$$

Then the estimate of the .25 quantile is given by

$$q_1=(1-h)X_{(\ell)}+hX_{(\ell+1)}. \qquad (4)$$

Letting $\ell'=n-\ell+1$, the estimate of the upper quartile, is

$$q_2=(1-h)X_{(\ell')}+hX_{(\ell-1)} \qquad (5)$$

and the estimate of the interquartile range is

$$IQR=q_2-q_1.$$

All that remains is estimating V_1, V_2, V_3 and V_4. An estimate of V_1 is obtained once an estimate of $P(X_1 \le x_{q1}, X_2 \le x_{q2})$ is available. The obvious estimate of this last quantity, and the one used here, is the proportion of times these inequalities are true among the sample of observations. That is, let A_{i12}=1 if simultaneously $X_{i1} \le X_{(m)1}$ and $X_{i2} \le X_{(m)2}$, otherwise A_{i12}=0. Then an estimate of V_1 is simply

$$\hat{V}_1 = \frac{(q-1)^2}{n} \sum \mathrm{A}_{i12}.$$

Estimates of V_2, V_3 and V_4 are obtained in a similar manner. The resulting estimate of the covariance between $X_{(m)1}$ and $X_{(m)2}$ is labeled $\hat{\tau}^2_{12}$. Of course, the squared standard error of $X_{(m)1}$ can be estimated in a similar fashion and is labeled $\hat{\tau}^2_{11}$.

An alternative approach is to use a bootstrap method, a possible appeal of which is that the usual sample median can be used when *n* is even. Generate a bootstrap sample by resampling with replacement *n* pairs of values from X_{ik} yielding X^*_{ik} (*i*=1,...,*n*; *k*=1, 2). For fixed *k*, let M^*_k be the usual sample median based on the bootstrap sample and corresponding to the k[th] marginal distribution. Repeat this *B* times yielding M^*_{bk}, *b*=1,...,*B*. Then an estimate of the covariance between M_1 and M_2 is

$$\hat{\xi}_{12} = \frac{1}{B-1} \sum (M^*_{i1} - \bar{M}_1)(M^*_{i2} - \bar{M}_2),$$

where $\bar{M}_k = \sum M^*_{bk} / B.$

Methodology

Now consider the more general case of a *J* by *K* design and suppose (1) is to be tested. Based on the results in the previous section, two test statistics are considered. The first estimates the population medians with a single order statistic, $X_{(m)}$, and the second uses the usual sample median, *M*.

Let X_{ijk} be a random sample of n_j vectors of observations from the j[th] group ($i = 1, \ldots, n_j$; *j*=1,...,*J*; *k*=1,...,*K*).

Let $\hat{\theta}_{jk} = X_{(m)jk}$ be the estimate of the median for the j[th] level of first factor and the k[th] level of the second. Then a test statistic for (1) can be developed along the lines used to derive the test statistic based on trimmed means, which is described in Wilcox (2003, section 11.9).

For convenience, let $\hat{\Theta}' = (\hat{\theta}_{11}, \ldots, \hat{\theta}_{JK})$. For fixed *j*, *k* and ℓ, $k \neq \ell$, let $v_{jk\ell}$ be the estimated covariance between θ_{jk} and $\theta_{j\ell}$. That is, $v_{jk\ell}$ is computed in the same manner as $\hat{\tau}^2_{12}$, only now use the data X_{ijk} and $X_{\imath jl}$, $i = 1, \ldots, n_j$. When $k=\ell$, $v_{jk\ell}$ is the estimated squared standard error of $\hat{\theta}_{jk}$. Let $\boldsymbol{V}$ be the K by K matrix where the element in the K[th] row and ℓ[th] column is given by $v_{jk\ell}$. The test statistic is

$$Q = \Theta' \mathbf{C}' (\mathbf{C}\mathbf{V}\mathbf{C}')^{-1} \mathbf{C}\Theta. \qquad (6)$$

As is well known, the usual choices for $\boldsymbol{C}$ for main effects for Factor A, main effects for Factor B, and for interactions are $\boldsymbol{C}=\boldsymbol{C}_J \otimes \mathbf{j}'_K$, $\mathbf{C} = \mathbf{j}'_J \otimes \mathbf{C}_K$ and $\boldsymbol{C}=\boldsymbol{C}_J \otimes \boldsymbol{C}_K$, respectively, where $\boldsymbol{C}_J$ is a *J*-1 by *J* matrix having the form

$$\begin{pmatrix} 1 & -1 & 0 & 0 & ... & 0 \\ 0 & 1 & -1 & 0 & ... & 0 \\ & & & \vdots & & \\ 0 & 0 & ... & 0 & 1 & -1 \end{pmatrix},$$

$\boldsymbol{j}_J'$ is a $1\times J$ matrix of ones and $\otimes$ is the (right) Kronecker product.

There remains the problem of approximating the null distribution of Q. Based on results in Wilcox (2003, chapter 11) when comparing groups using a 20% trimmed mean, an obvious speculation is that Q has, approximately, an F distribution with ν_1 and ν_2 degrees of freedom. For main effects for Factor A, main effects for Factor B, and for interactions, ν_1 is equal to J-1, K-1 and (J-1)(K-1), respectively. As for ν_2, it is estimated based on the data, but an analog of this method for medians was not quite satisfactory in simulations; the actual probability of a Type I error was too far below the nominal level. A better approach was simply to take $\nu_2 = \infty$, which will be assumed henceforth. This will be called method A.

An alternative approach is to proceed exactly as in method A, only estimate the .5 quantiles with the usual sample median and replace $\boldsymbol{V}_j$ with the bootstrap estimate described in section 2. (Here, B=100 is used.) This will be called method B.

An Approached Based on Linear Contrasts

Another approach to analyzing the two-way ANOVA design under consideration is to test hypotheses about a collection of linear contrasts appropriate for studying main effects and interactions. Consider, for example,

$$\hat{\Psi}_j = \sum \hat{\theta}_{jk}.$$

$j = 1, \ldots, J$. Then when dealing with main effects for Factor A, one could perform all pairwise comparisons among the $\Psi_{j.}$ This is for every $j < j'$,

$$H_0 : \Psi_j = \Psi_{j'}.$$

There is the problem of controlling the probability of at least one Type I error among the $(J^2 - J)/2$ hypotheses to be tested, and here this is done with a method derived by Rom (1990). Interactions can be studied by testing hypotheses about all of the relevant $(J^2 - J)(K^2 - K)/4$ tetrad differences, and of course, main effects for Factor B can be handled in a similar manner.

For convenience, attention is focused on Factor A (the first factor). Here, Ψ_j is simply estimated with

$$\hat{\Psi}_j = \sum \hat{\theta}_{jk}.$$

Writing

$$\hat{\Psi}_j - \hat{\Psi}_{j'} = \sum \sum c_{jk} \hat{\theta}_{jk}$$

for appropriately chosen contrast coefficents $c_{jk,}$ then of course an estimate of the squared standard error of $\hat{\Psi}_j - \hat{\Psi}_{j'}$ is

$$\hat{n}^2 = \sum \sum c_{jk} \hat{\tau}_{jk},$$

Based on results in Wilcox (2004), the null distribution of T is approximated with a Student's T distribution with $n-1$ degrees of freedom.

To elaborate on controlling the probability of at least one Type I error with Rom's method, and still focusing on Factor A, let $D = (J^2 - J)/2$ be the number of hypotheses to be tested and let $P_1, \ldots P_D$ be the corresponding p-values. Put the p-values in descending order yielding $P_{[1]} \geq P_{[2]} \geq \ldots P_{[D]}$.

Proceed as follows:

1. Set ℓ =1.

2. If $P_{[\ell]} \le d_\ell$, where d_ℓ is read from Table 1, stop and reject all *D* hypotheses; otherwise, go to step 3 (If $\ell > 10$, use $d_\ell = \alpha / \ell$).

3. Increment ℓ by 1. If $P_{[\ell]} \le d_{\ell,}$, stop and reject all hypotheses having a significance level less than or equal d_ℓ.

4. If $P_{[\ell]} > d_\ell$, repeat step 3.

5. Continue until a significant result is obtained or all *D* hypotheses have been tested.

A Simulation Study

Simulations were used to study the small-sample properties of the methods just described. Vectors of observations were generated from multivariate normal distributions having a common correlation, ρ. To study the effect of non-normality, observations were transformed to various g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If *Z* has a standard normal distribution, then

$$W = \begin{cases} \dfrac{\exp(gZ)-1}{g} \exp\left(hZ^2/2\right), & \text{if g > 0} \\ Z\exp(hZ^2/2) & \text{if g > 0} \end{cases}$$

has a g-and-h distribution where *g* and *h* are parameters that determine the first four moments. The four distributions used here were the standard normal (*g* = *h* =0.0), a symmetric heavy-tailed distribution (*h* = 0.5, *g* = 0.0), an asymmetric distribution with relatively light tails (*h* = 0.0, *g* = 0.5), and an asymmetric distribution with heavy tails (*g* = *h* = 0.5). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution considered. For *h* = .5, the third and fourth moments are not defined and so no values for the skewness and kurtosis are reported. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 1: Critical values, $d_{\ell,}$ for Rom's method.

ℓ	α = .05	α = .01
1	.05000	.01000
2	.02500	.00500
3	.01690	.00334
4	.01270	.00251
5	.01020	.00201
6	.00851	.00167
7	.00730	.00143
8	.00639	.00126
9	.00568	.00112
10	.00511	.00101

Table 2: Some properties of the g-and-h distribution.

g	h	(κ_1)	(κ_2)
0.0	0.0	0.00	3.0
0.0	0.5	0.00	—
0.5	0.0	1.81	8.9
0.5	0.5	—	—

Table 3: Estimated probability of a Type I error, $J = K = 2$, $n = 20$, $\alpha = .05$

			Method A		Method B		Method C	
g	h	ρ	Factor A	Inter	Factor A	Inter	Factor A	Inter
0.0	0.0	0.0	.074	.068	.046	.050	.051	.052
0.0	0.0	0.8	.072	.073	.032	.036	.048	.048
0.0	0.5	0.0	.046	.045	.048	.053	.025	.027
0.0	0.5	0.8	.049	.036	.047	.038	.026	.027
0.5	0.0	0.0	.045	.053	.045	.044	.045	.049
0.5	0.0	0.8	.044	.024	.047	.029	.043	.048
0.5	0.5	0.0	.030	.038	.030	.038	.021	.020
0.5	0.5	0.8	.019	.027	.032	.015	.023	.024

Table 4: Estimated Type I error rates using Methods A and C, $J = 2$, $K = 3$, $n = 20$, $\alpha = .05$

			Method A			Method C		
g	h	ρ	Factor A	Factor B	Inter	Factor A	Factor B	Inter
0.0	0.0	0.0	.047	.036	.043	.059	.044	.049
0.0	0.0	0.8	.062	.021	.023	.056	.057	.047
0.0	0.5	0.0	.034	.023	.026	.026	.018	.019
0.0	0.5	0.8	.038	.012	.015	.031	.023	.025
0.5	0.0	0.0	.040	.032	.039	.053	.040	.045
0.5	0.0	0.8	.055	.020	.016	.052	.047	.050
0.5	0.5	0.0	.027	.017	.023	.024	.015	.019
0.5	0.5	0.8	.035	.010	.010	.025	.024	.023

Simulations were run for the case $J = K = 2$ with $n = 20$. (Simulations also were run with $n = 100$ and 200 as a partial check on the software.) Table 3 shows the estimated probability of a Type I error when $\rho = 0$ or .8 when testing Factor A and the hypothesis of no interaction with method A. For brevity, results for Factor B are not shown because they are essentially the same as for Factor A, which should be the case. The estimates are based on 1,000 replications. (From Robey & Barcikowski, 1992, 1,000 replications is sufficient from a power point of view. More specifically, if we test the hypothesis that the actual Type I error rate is .05, and if we want power to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required).

As is evident, method A does a reasonable job of controlling the probability of a Type I error, the main difficulty being that when sampling from a very heavy-tailed distribution, the estimated probability of a Type I error can drop below .025. Switching to method B does not correct this problem. Generally, when using method B the estimated probability of a Type I error was approximately the same or smaller than the estimates shown in Table 3. For example, under normality with $\rho = .8$, the estimates corresponding to Factor A and the hypothesis of no interaction were .035 and .011, respectively. As for method C it performs well with the possible appeal that the estimate never drops below .02, unlike method B.

Table 4 reports results for methods A and C when $J = 2$ and $K = 3$. Both methods avoid Type I error probabilities well above the nominal level. Both methods have estimates that drop below .02, but in general method C seems a bit more satisfactory.

When $J = 3$ and $K = 5$, method A deteriorates even more when dealing with Factor B and interactions, with estimated Type I error probabilities typically below .01. (One exception is normality with $\rho = 0$; the estimates were .020 and .023.) All indications are that method C does better at providing actual Type I error probabilities close to the nominal level. For example, under normality with $\rho = .8$, method A has estimated Type I error probabilities equal to .044, .006 and .001 for Factors A, B and interactions, respectively. For method C, the estimates were .057, .042 and .068.

Conclusion

In summary, the bootstrap version of method A (method B) does not seem to have any practical value based on the criterion of controlling the probability of a Type I error. This is in contrast to the situations considered in Wilcox (2004) where pairwise multiple comparisons among J dependent groups were considered. A possible appeal of method B is that it uses the usual sample median when n is even rather than a single order statistic, but at the cost of risking actual Type I error probabilities well below the nominal level.

Methods A, B and C perform well in terms of avoiding Type I error probabilities well above the nominal level, but methods A and B become too conservative in certain situations where method C continues to perform reasonably well. It seems that applied researchers rarely have interest in an omnibus hypothesis only; the goal is to know which levels of the factor differ. Because the linear contrasts can be tested in a manner that controls FWE, all indications are that method C is the best method for routine use. Finally, S-PLUS and R functions are available from the author for applying method C. Please ask for the function mwwmcp.

References

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics, 37,* 577–580.

Davies, P. L., & Kovac, A. (2004). Densities, spectral densities and modality. *Annals of Statistics, 32,* 1093–1136.

Dawson, M., Schell, A., Rissling, A., & Wilcox, R. R. (2004). Evaluative learning and awareness of stimulus contingencies. Unpublished technical report, Dept of Psychology, University of Southern California.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes.* (p. 461–515). New York: Wiley.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283–288.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*, 663–666.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.

Staudte, R. G., & Sheather, S. J. (1990). *Robust Estimation and Testing.* New York: Wiley.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques.* San Diego, CA: Academic Press.

Wilcox, R. R. (2004). Pairwise comparisons of dependent groups based on medians. *Computational Statistics & Data Analysis*, submitted.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypot Testing*, 2nd Ed. San Diego, CA: Academic Press

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 11-34

1538 – 9472/05/$95.00

Regular Articles

Testing the Goodness of Fit of Multivariate Multiplicative-intercept Risk Models Based on Case-control Data

Biao Zhang
Department of Mathematics
The University of Toledo

The validity of the multivariate multiplicative-intercept risk model with $I+1$ categories based on case-control data is tested. After reparametrization, the assumed risk model is equivalent to an $(I+1)$-sample semiparametric model in which the I ratios of two unspecified density functions have known parametric forms. By identifying this $(I+1)$-sample semiparametric model, which is of intrinsic interest in general $(I+1)$-sample problems, with an $(I+1)$-sample semiparametric selection bias model, we propose a weighted Kolmogorov-Smirnov-type statistic to test the validity of the multivariate multiplicative-intercept risk model. Established are some asymptotic results associated with the proposed test statistic, also established is an optimal property for the maximum semiparametric likelihood estimator of the parameters in the $(I+1)$-sample semiparametric selection bias model. In addition, a bootstrap procedure along with some results on analysis of two real data sets is proposed.

Key words: Biased sampling problem, bootstrap, Kolmogorov-Smirnov two-sample statistic, logistic regression, mixture sampling, multivariate Gaussian process, semiparametric selection bias model, strong consistency, weak convergence

Introduction

Let Y be a multicategory response variable with $I+1$ categories and X be the associated $p\times 1$ covariate vector. When the possible values of the response variable Y are denoted by $y=0,1,\ldots,I$ and the first category (0) is the baseline category, Hsieh, Manski, and McFadden (1985) introduced the following multivariate multiplicative-intercept risk model:

$$\frac{P(Y=i\mid X=x)}{P(Y=0\mid X=x)}=\theta_i^* r_i(x;\beta_i),\quad i=1,\ldots,I,\ (1)$$

Biao Zhang is a Professor in the Department of Mathematics at the University of Toledo. His research interests include categorical data analysis and empirical likelihood. The author wishes to thank Xin Deng and Shuwen Wan for their help in the manuscript conversion process. Email: bzhang@utnet.utoledo.edu.

where $\theta_1^*,\ldots,\theta_I^*$ are positive scale parameters, $r_1,\ldots,r_I$ are, for fixed x, known functions from R^p to R^+, and $\beta_i=(\beta_i^1,\ldots,\beta_i^p)^\tau$ is a $p\times 1$ vector parameter for $i=1,\ldots I$. The class of multivariate multiplicative-intercept risk models includes the multivariate logistic regression models and the multivariate odds-linear models discussed by Weinberg and Sandler (1991) and Wacholder and Weinberg (1994). By generalizing earlier works of Anderson (1972, 1979), Farewell (1979), and Prentice and Pyke (1979) in the context of the logistic regression models, Weinberg and Wacholder (1993) and Scott and Wild (1997) showed that under model (1.1), a prospectively derived analysis, including parameter estimates and standard errors for $\beta_1,\ldots,\beta_I$, is asymptotically correct in case-control studies. In this article, testing the validity of model (1) based on case-control data as specified below is considered.

Let $X_{i1},\ldots,X_{in_i}$ be a random sample from $P(x\,|\,Y=i)$ for $i=0,1,\ldots,I$ and assume that $\{(X_{i1},\ldots,X_{in_i}):\ i=0,1,\ldots I\}$ are jointly independent. Let $\pi_i=P(Y=i)$ and $g_i(x)=f(x\,|\,Y=i)$ be the conditional density or frequency function of X given $Y=i$ for $i=0,1,\ldots I$. If $f(x)$ is the marginal distribution of X , then applying Bayes' rule yields

$$f(x\,|\,Y=i)=\frac{P(Y=i\,|\,X=x)}{\pi_i}\,f(x),\qquad i=0,1,\ldots I.$$

It is seen that

$$\frac{f(x\,|\,Y=i)}{f(x\,|\,Y=0)}=\frac{\pi_0}{\pi_i}\frac{P(Y=i\,|\,X=x)}{P(Y=0\,|\,X=x)}$$
$$=\frac{\pi_0}{\pi_i}\theta_i^{*}r_i(x;\beta_i),\qquad i=1,\ldots,I.$$

Consequently,

$$g_i(x)=f(x\,|\,Y=i)=\frac{\pi_0}{\pi_i}\theta_i^{*}r_i(x;\beta_i)\,f(x\,|\,Y=0)$$
$$=\exp[\theta_i+s_i(x;\beta_i)]g_0(x),\qquad i=1,\ldots,I,$$

where $\theta_i=\log\theta_i^{*}+\log(\pi_0/\pi_i)$ and

$s_i(x;\beta_i)=\log r_i(x;\beta_i)$ for $i=1,\ldots,I$. As a result, the following $(I+1)$ -sample semiparametric model is obtained:

$$X_{01},\ldots,X_{0n_0}\overset{i.i.d.}{\sim}g_0(x),$$
$$X_{i1},\ldots,X_{in_i}\overset{i.i.d.}{\sim}g_i(x)=\exp[\theta_i+s_i(x;\beta_i)]g_0(x),\qquad i=1,\ldots,I. \tag{2}$$

Throughout this article, let $\theta=(\theta_1,\ldots,\theta_I)^{\tau}$, $\beta=(\beta_1^{\tau},\ldots,\beta_I^{\tau})^{\tau}$, and $G_i(x)$ be the corresponding cumulative distribution function of $g_i(x)$ for $i=0,1,\ldots I$. Note that model (2) is equivalent to an $(I+1)$ -sample semiparametric model in which the i^{th} $(i=1,\ldots,I)$ ratio of a pair of unspecified density functions g_i and g_0 has a known parametric form, and thus is of intrinsic interest in general $(I+1)$ -sample problems. Model (2) is equivalent to model (1); it is an $(I+1)$ -sample semiparametric selection bias model with weight functions $w_0(x,\theta,\beta)=1$ and

$$w_i(x,\theta,\beta)=\exp[\theta_i+s(x;\beta_i)]$$

for $i=1,\ldots,I$ depending on the unknown parameters θ and β . The s -sample semiparametric selection bias model was proposed by Vardi (1985) and was further developed by Gilbert, Lele, and Vardi (1999). Vardi (1982, 1985), Gill, Vardi, and Wellner (1988), and Qin (1993) discussed estimating distribution functions in biased sampling models with known weight functions. Weinberg and Wacholder (1990) considered more flexible design and analysis of case-control studies with biased sampling. Qin and Zhang (1997) and Zhang (2002) considered goodness-of-fit tests for logistic regression models based on case-control data, whereas Zhang (2000) considered testing the validity of model (2) when $I=1$.

The focus in this article is to test the validity of model (1.2) for $I\ge 1$. Let $\{T_1,\ldots,T_n\}$ denote the pooled sample $\{X_{01},\ldots,X_{0n_0};X_{11},\ldots,X_{1n_1};\ldots;X_{I1},\ldots,X_{In_I}\}$ with $n=\sum_{i=0}^{I}n_i$. Furthermore, let

$$\widehat{G}_i(t)=n_i^{-1}\sum_{j=1}^{n_i}I_{[X_{ij}\le t]}$$

and

$$\widehat{G}_0(t)=n^{-1}\sum_{k=1}^{n}I_{[T_k\le t]}$$

be, respectively, the empirical distribution functions based on the sample $X_{i1},\ldots,X_{in_i}$ from the i^{th} $(i=0,1,\ldots,I)$ category and the pooled sample $T_1,\ldots,T_n$. In the special case of testing

the equality of G_0 and G_1 for which $I=1$ and $s_1(x;\beta_1)\equiv 0$ in model (2), as argued by (van der Vaart & Wellner, 1996, p. 361; Qin & Zhang, 1997), the Kolmogorov-Smirnov two-sample statistic is equivalent to a statistic based on the discrepancy between the empirical distribution function $\widehat{G}_0$ and the pooled empirical distribution function $\overline{G}_0$. This fact, along with the fact that $\widehat{G}_0$ and $\overline{G}_0$ are, respectively, the nonparametric maximum likelihood estimators of G_0 without and with the assumption of $G_0(t)=G_1(t)$, motivates us to employ a weighted average of the $I+1$ discrepancies between $\widehat{G}_i$ and $\widetilde{G}_i$ $(i=0,1,\ldots,I)$ to assess the validity of model (2), where $\widetilde{G}_i$ is the maximum semiparametric likelihood estimator of G_i under model (2) and is derived by employing the empirical likelihood method developed by Owen (1988, 1990). For a more complete survey of developments in empirical likelihood, see Hall and La Scala (1990) and Owen (1991).

This article is structured as follows: in the method section proposed is a test statistic by deriving the maximum semiparametric likelihood estimator of G_i under model (2). Some asymptotic results are then presented along with an optimal property for the maximum semiparametric likelihood estimator of (θ,β). This is followed by a bootstrap procedure which allows one to find P-values of the proposed test. Also reported are some results on analysis of two real data problems. Finally, proofs of the main theoretical results are offered.

Methodology

Based on the observed data in (2), the likelihood function can be written as

$$L(\theta,\beta,G_0)=\prod_{i=0}^{I}\prod_{j=1}^{n_i}\exp[\theta_i+s_i(X_{ij};\beta_i)]dG_0(X_{ij})$$
$$=\left(\prod_{k=1}^{n}p_k\right)\left[\exp\left(\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\theta_i+s_i(X_{ij};\beta_i)]\right)\right],$$

where $\theta_0=0$, $s_0(\bullet;\beta_0)\equiv 0$, and $p_k=dG_0(T_k)$, $k=1,\ldots,n$, are (nonnegative) jumps with total mass unity. Similar to the approach of Owen (1988, 1990) and Qin and Lawless (1994), it can be shown by using the method of Lagrange multipliers that for fixed (θ,β), the maximum value of L, subject to constraints $\sum_{k=1}^{n}p_k=1$, $p_k\geq 0$ and

$$\sum_{k=1}^{n}p_k\{\exp[\theta_i+s_i(T_k;\beta_i)]-1\}=0$$

for $i=1,\ldots,I$, is attained at

$$p_k=\frac{1}{n_0}\frac{1}{1+\sum_{i=1}^{I}\rho_i\exp[\theta_i+s_i(T_k;\beta_i)]},$$
$$k=1,\ldots,n,$$

where $\rho_i=n_i/n_0$ for $i=0,1,\ldots,I$. Therefore, the (profile) semiparametric log-likelihood function of (θ,β) is given by

$$\ell(\theta,\beta)=-n\log n_0$$
$$-\sum_{k=1}^{n}\log\left[1+\sum_{i=1}^{I}\rho_i\exp[\theta_i+s_i(T_k;\beta_i)]\right]$$
$$+\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\theta_i+s_i(X_{ij};\beta_i)].$$

Next, maximize ℓ over (θ,β). Let $(\tilde{\theta},\tilde{\beta})$ with $\tilde{\theta}=(\tilde{\theta}_1,\ldots\tilde{\theta}_I)^\tau$ and $\tilde{\beta}=(\tilde{\beta}_1^{\,\tau},\ldots\tilde{\beta}_I^{\,\tau})^\tau$ be the solution to the following system of score equations:

$$\frac{\partial \ell(\theta,\beta)}{\partial \theta_u} = n_u - \sum_{k=1}^{n} \frac{\rho_u \exp[\theta_u + s_u(T_k;\beta_u)]}{1+\sum_{i=1}^{I}\rho_i \exp[\theta_i + s_i(T_k;\beta_i)]}$$
$$= 0, \quad u = 1,\ldots,I,$$
$$\frac{\partial \ell(\theta,\beta)}{\partial \beta_u} = \sum_{j=1}^{n_u} d_u(X_{uj};\beta_u)$$
$$-\sum_{k=1}^{n} \frac{\rho_u \exp[\theta_u + s_u(T_k;\beta_u)]}{1+\sum_{m=1}^{I}\rho_m \exp[\theta_m + s_m(T_k;\beta_m)]} d_u(T_k;\beta_u)$$
$$= 0, \quad u = 1,\ldots,I, \qquad (3)$$

where $d_u(T_k;\beta_u) = \dfrac{\partial s_u(\theta,\beta_u)}{\partial \beta_u}$ for $u = 1,\ldots,I$. That produces the following,

$$\tilde{p}_k = \frac{1}{n_0}\frac{1}{1+\sum_{i=1}^{I}\rho_i \exp[\tilde{\theta}_i + s_i(T_k;\tilde{\beta}_i)]}, \quad k = 1,\ldots,n. \qquad (4)$$

On the basis of the $\tilde{p}_k$ in (4), it can be proposed to estimate $G_i(t)$, under model (2), by

$$\tilde{G}_i(t) = \sum_{k=1}^{n} \tilde{p}_k \exp[\tilde{\theta}_i + s_i(T_k;\tilde{\beta}_i)] I_{[T_k \le t]}$$
$$= \frac{1}{n_0}\sum_{k=1}^{n} \frac{\exp[\tilde{\theta}_i + s_i(T_k;\tilde{\beta}_i)]}{1+\sum_{m=1}^{I}\rho_m \exp[\tilde{\theta}_m + s_m(T_k;\tilde{\beta}_m)]} I_{[T_k \le t]}, \quad i = 0,\ldots,I, \qquad (5)$$

where $\tilde{\theta}_0 = 0$ and $s_0(\bullet;\tilde{\beta}_0) \equiv 0$. Throughout this article, $a \le b$ and $-\infty \le a \le \infty$ with $a = (a_1,\ldots,a_p)^\tau$ and $b = (b_1,\ldots,b_p)^\tau$ stand for, respectively, $a_i \le b_i$ and $-\infty \le a_i \le \infty$ for $i = 0,1,\ldots,p$. Note that $\tilde{G}_i$ is the maximum semiparametric likelihood estimator of G_i under model (2) for $i = 0,1,\ldots,I$. Let $\hat{G}_i(t) = n_i^{-1}\sum_{j=1}^{n_i} I_{[X_{ij} \le t]}$ be the empirical distribution function based on the sample $X_{i1},\ldots,X_{in_i}$ from the i th $(i = 0,1,\ldots,I)$ category. Moreover, let

$$\Delta_{ni}(t) = \sqrt{n}\left(\tilde{G}_i(t) - \hat{G}_i(t)\right), \qquad \Delta_{ni} = \sup_{-\infty \le t \le \infty} |\Delta_{ni}(t)|, \quad i = 0,1,\ldots,I.$$

Then, Δ_{ni} is the discrepancy between the two estimators $\tilde{G}_i(t)$ and $\hat{G}_i(t)$, and thus measures the departure from the assumption of the multivariate multiplicative-intercept risk model (1) within the i th $(i = 1,\ldots,I)$ pair of category i and the baseline category (0). Since $\sum_{i=0}^{I}\rho_i\Delta_{ni}(t) = \sqrt{n}\sum_{i=0}^{I}\rho_i[\tilde{G}_i(t) - \hat{G}_i(t)] = 0$, there exists a motivation to employ the weighted average of the Δ_{ni} defined by

$$\Delta_n = \frac{1}{I+1}\sum_{i=0}^{I}\rho_i\Delta_{ni} \qquad (6)$$

to assess the validity of model (2). Clearly, the proposed test statistic Δ_n measures the global departure from the assumption of the multivariate multiplicative-intercept risk model (1). Because the same value of Δ_n occurs no matter which category is the baseline category, there is a symmetry among the $I+1$ category designations for such a global test. Thus, the choice of the baseline category in model (1) is arbitrary for testing the validity of model (1) or model (2) based on Δ_n. Note that the test statistic Δ_n reduces to that of Zhang (2000) when $I = 1$ in model (1) since $\Delta_n = 2^{-1}(\Delta_{n0} + \rho_1\Delta_{n1}) = \Delta_{n0}$ for $I = 1$.

Remark 1: The test statistic Δ_n can also be applied to mixture sampling data in which a sample of $n = \sum_{i=0}^{I} n_i$ members is randomly selected from the whole population with $n_0, n_1,\ldots,n_I$ being random (Day & Kerridge, 1967). Let (X_k, Y_k), $k = 1,\ldots,n$, be a random sample from the joint distribution of (X,Y), then the likelihood has the form of

$$L = \prod_{k=1}^{n} P(Y_k \mid X_k) f(X_k)$$
$$= \prod_{i=0}^{I} \prod_{j=1}^{n_i} [\pi_i f(X_{ij} \mid Y = i)],$$

where $\pi_i = P(Y = i)$ for $i = 1, \ldots, I$. The first expression is a prospective decomposition and the second one is a retrospective decomposition.

Remark 2: In light of Anderson (1972, 1979), the case-control data may be treated as the prospective data to compute the maximum likelihood estimate of (θ^*, β) under model (1), where $\theta^* = (\theta_1^*, \ldots, \theta_I^*)^\tau$. Suppose that the sample data in model (2) are collected prospectively, then the (prospective) likelihood function is, by (1),

$$L(\theta^*, \beta) = \prod_{i=0}^{I} \prod_{j=1}^{n_i} P(Y = i \mid X = X_{ij}) = \prod_{i=0}^{I} \prod_{j=1}^{n_i} \left[\frac{\theta_i^* r_i(X_{ij}; \beta_i)}{1 + \sum_{m=1}^{I} \theta_m^* r_m(X_{ij}; \beta_m)} \right].$$

The log-likelihood function is

$$\ell(\theta^*, \beta) = \sum_{i=0}^{I} \sum_{j=1}^{n_j} [\log \theta_i^* + s_i(X_{ij}; \beta_i)]$$
$$- \sum_{k=1}^{n} \log \left[1 + \sum_{m=1}^{I} \theta_m^* \exp[s_m(T_k; \beta_m)] \right].$$

The system of score equations is given by

$$\frac{\partial \ell(\theta^*, \beta)}{\partial \theta_u^*}$$
$$= \frac{1}{\theta_u^*} \left[n_u - \sum_{k=1}^{n} \frac{\theta_u^* \exp[s_u(T_k; \beta_u)]}{1 + \sum_{m=1}^{I} \theta_m^* \exp[s_m(T_k; \beta_m)]} \right]$$
$$= 0, \quad u = 1, \ldots, I,$$

$$\frac{\partial \ell(\theta^*, \beta)}{\partial \beta_u} = \sum_{j=1}^{n_u} d_u(X_{uj}; \beta_u)$$
$$- \sum_{k=1}^{n} \frac{\theta_u^* \exp[s_u(T_k; \beta_u)]}{1 + \sum_{m=1}^{I} \theta_m^* \exp[s_m(T_k; \beta_m)]} d_u(T_k; \beta_u) = 0,$$
$$u = 1, \ldots, I. \qquad (7)$$

Let $(\widehat{\theta^*}, \widehat{\beta})$ with $\widehat{\theta^*} = (\widehat{\theta_1^*}, \ldots, \widehat{\theta_I^*})^\tau$ and $\widehat{\beta} = (\widehat{\beta_1}^\tau, \ldots, \widehat{\beta_I}^\tau)^\tau$ denote the solution to the system of score equations in (7). Then comparing (7) with (3) implies that $\widetilde{\theta_u} = \log \widehat{\theta_u^*} + \log(n_0 / n_u)$ and $\widetilde{\beta_u} = \widehat{\beta_u}$ for $u = 1, \ldots, I$. Thus, the maximum likelihood estimates of are identical under the retrospective sampling scheme and the prospective sampling scheme. In addition, the two estimated asymptotic variance-covariance matrices for $\widehat{\beta}$ and $\widetilde{\beta}$ based on the observed information matrices coincide. See also Remarks 3 and 4 below.

Asymptotic results

In this section, the asymptotic properties of the proposed estimator $\widetilde{G_i}(t)$ $(i = 0, 1, \ldots, I)$ in (5) and the proposed test statistic Δ_n in (6) are studied. To this end, let $(\theta_{(0)}, \beta_{(0)})$ be the true value of (θ, β) under model (2) with

$$\theta_{(0)} = (\theta_{10}, \ldots, \theta_{p0})^\tau$$

and

$$\beta_{(0)} = (\beta_{10}{}^\tau, \ldots, \beta_{p0}{}^\tau)^\tau.$$

Throughout this article, it is assumed that $\rho_i = n_i / n_0$ $(i = 0, 1, \ldots, I)$ is positive and finite and remains fixed as $n = \sum_{i=0}^{I} n_i \to \infty$. Write $\rho = \sum_{i=0}^{I} \rho_i$ and

$$d_i(t; \beta_i) = \frac{\partial s_i(t; \beta_i)}{\partial \beta_i},$$

$$D_i(t; \beta_i) = \frac{\partial d_i(t; \beta_i)}{\partial \beta_i^\tau} = \frac{\partial^2 s_i(t; \beta_i)}{\partial \beta_i \partial \beta_i^\tau}$$

$$i = 1, \ldots, I,$$

$$s_{11}^{uv} = -\frac{1}{1+\rho}\times$$
$$\int \frac{\rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right]}{1+\sum_{i=1}^{I}\rho_i \exp[\theta_{i0}+s_i(y;\beta_{i0})]} dG_0(y),$$
$$u \neq v = 0,1,\ldots,I,$$
$$s_{11}^{uu} = -\sum_{v=0,v\neq u}^{I} s_{11}^{uv}, \quad u = 0,1,\ldots,I,$$
$$S_{11} = (s_{11}^{uv})_{u,v=1,\ldots,I},$$
$$s_{21}^{uv} = -\frac{1}{1+\rho}\times$$
$$\int \frac{\rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right]}{1+\sum_{i=1}^{I}\rho_i \exp[\theta_{i0}+s_i(y;\beta_{i0})]} d_u(y,\beta_u) dG_0(y),$$
$$u \neq v = 0,1,\ldots,I$$
$$s_{21}^{uu} = -\sum_{v=0,v\neq u}^{I} s_{21}^{uv},$$
$$u = 0,1,\ldots,I, \quad S_{21} = (s_{21}^{uv})_{u,v=1,\ldots,I},$$
$$s_{22}^{uv} = -\frac{1}{1+\rho}\times$$
$$\int \frac{\rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right]}{1+\sum_{i=1}^{I}\rho_i \exp[\theta_{i0}+s_i(y;\beta_{i0})]} \times$$
$$d_u(y,\beta_u) d_v^{\tau}(y,\beta_v)]dG_0(y), \quad u \neq v = 0,1,\ldots,I,$$
$$s_{22}^{uu} = -\sum_{v=0,v\neq u}^{I} s_{22}^{uv}, \quad u = 0,1,\ldots,I,$$
$$S_{22} = (s_{22}^{uv})_{u,v=1,\ldots,I}, \qquad S = \begin{pmatrix} S_{11} & S_{21}^{\tau} \\ S_{21} & S_{22} \end{pmatrix}$$
$$\Sigma = S^{-1} - (1+\rho)\begin{pmatrix} D+J & 0 \\ 0 & 0 \end{pmatrix},$$
$$B_{uv}(t) =$$
$$\int_{-\infty}^{t} \frac{\rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right]}{1+\sum_{m=1}^{I}\rho_m \exp[\theta_{m0}+s_m(y;\beta_{m0})]} dG_0(y),$$
$$u,v = 0,1,\ldots,I,$$
$$a_{uv}(t;\theta,\beta) = \rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\times$$
$$\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right], \quad u \neq v = 0,1,\ldots,I,$$
$$a_{uu}(t) = -\sum_{v=0,v\neq u}^{I} a_{uv}(t;\theta,\beta), \quad u = 0,1,\ldots,I,$$
$$C_{1h}(t;\theta,\beta) = (a_{h1}(t;\theta,\beta),\ldots,a_{hI}(t;\theta,\beta))^{\tau},$$
$$h = 0,1,\ldots,I,$$
$$b_{uv}(t;\theta,\beta) = \rho_u \exp\left[\theta_{u0}+s_u(y;\beta_{uo})\right]\times$$
$$\rho_v \exp\left[\theta_{v0}+s_v(y;\beta_{vo})\right] d_u(t;\beta_u),$$
$$u \neq v = 0,1,\ldots,I,$$
$$b_{uu}(t) = -\sum_{v=0,v\neq u}^{I} b_{uv}(t;\theta,\beta), \quad u = 0,1,\ldots,I,$$
$$C_{2h}(t;\theta,\beta) = (b_{h1}(t;\theta,\beta),\ldots,b_{hI}(t;\theta,\beta))^{\tau},$$
$$h = 0,1,\ldots,I,$$
$$A_{kh}(t) =$$
$$\int_{-\infty}^{t} \frac{C_{kh}(y;\theta_{(0)},\beta_{(0)})}{1+\sum_{m=1}^{I}\rho_m \exp[\theta_{m0}+s_m(y;\beta_{m0})]} dG_0(y),$$
$$k = 1,2, \quad h = 0,1,\ldots,I, \tag{8}$$

where J is an $I\times I$ matrix of 1 elements and $D = \mathrm{Diag}(\rho_1^{-1},\ldots,\rho_I^{-1})$ is the $I\times I$ diagonal matrix having elements $\{\rho_1^{-1},\ldots,\rho_I^{-1}\}$ on the main diagonal. In order to formulate the results, the following assumptions are stated.

(A1) There exists a neighborhood Θ_0 of the true parameter point $\beta_{(0)}$ such that for all t the function $r_i(t;\beta_i)$ $(i = 1,\ldots,I)$ admits all third derivatives $\dfrac{\partial^3 r_i(t;\beta_i)}{\partial\beta_i^k \partial\beta_i^l \partial\beta_i^m}$ for all $\beta\in\Theta_0$

(A2) For $i = 1,\ldots,I$, there exists a function Q_1 such that $\left|\dfrac{\partial s_i(t;\beta_i)}{\partial\beta_i^k}\right| \le Q_1(t)$ for all $\beta\in\Theta_0$ and $k = 1,\ldots,p$, where

$$q_{1j} = \int Q_1^j(y)\{1+\rho_i \exp[\theta_{i0}+s_i(y;\beta_{io})]\} dG_0(y) < \infty,$$
$$j = 1,2,3.$$

(A3) For $i = 1,\ldots,I$, there exists a function Q_2

such that $\left|\frac{\partial^2 s_i(t;\beta_i)}{\partial\beta_i^k\partial\beta_i^l}\right| \le Q_2(t)$ for all $\beta\in\Theta_0$ and $k,l=1,\ldots,p$, where

$$q_{2j} = \int Q_2^j(y)\{1+\rho_i\exp[\theta_{i0}+s_i(y;\beta_{io})]\}dG_0(y)<\infty,$$
$$j=1,2.$$

(A4) For $i=1,\ldots,I$, there exists a function Q_3 such that $\left|\frac{\partial^3 s_i(t;\beta_i)}{\partial\beta_i^k\partial\beta_i^l\partial\beta_i^m}\right| \le Q_3(t)$ for all $\beta\in\Theta_0$ and $k,l,m=1,\ldots,p$, where

$$q_3 = \int Q_3(y)\{1+\rho_i\exp[\theta_{i0}+s_i(y;\beta_{io})]\}dG_0(y)<\infty$$

First, study the asymptotic behavior of the maximum semiparametric likelihood estimate $(\tilde\theta,\tilde\beta)$ defined in (3). Theorem 8 concerns the strong consistency and the asymptotic distribution of $(\tilde\theta,\tilde\beta)$

Theorem 1: Suppose that model (2) and Assumptions (A1)−(A4) hold. Suppose further that S is positive definite. (a) As $n\to\infty$, with probability 1 there exists a sequence $(\tilde\theta,\tilde\beta)$ of roots of the system of score equations (2.1) such that $(\tilde\theta,\tilde\beta)$ is strongly consistent for estimating $(\theta_{(0)},\beta_{(0)})$, i.e., $(\tilde\theta,\tilde\beta)\xrightarrow{a.s.}(\theta_{(0)},\beta_{(0)})$.

(b) As $n\to\infty$, it may be written

$$\begin{pmatrix}\tilde\theta-\theta_{(0)}\\ \tilde\beta-\beta_{(0)}\end{pmatrix} = \frac{1}{n}S^{-1}\begin{pmatrix}\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix} + o_p(n^{-1/2}), \quad (9)$$

where $\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta} = \frac{\partial\ell(\theta,\beta)}{\partial\theta}\Big|_{(\theta,\beta)=(\theta_{(0)},\beta_{(0)})}$ and $\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta} = \frac{\partial\ell(\theta,\beta)}{\partial\beta}\Big|_{(\theta,\beta)=(\theta_{(0)},\beta_{(0)})}$. As a result,

$$\sqrt{n}\begin{pmatrix}\tilde\theta-\theta_{(0)}\\ \tilde\beta-\beta_{(0)}\end{pmatrix} \xrightarrow{d} N_{(p+1)I}(0,\Sigma). \quad (10)$$

Remark 3: A consistent estimate of the covariance matrix Σ is given by

$$\tilde\Sigma = \tilde S^{-1} - (1+\rho)\begin{pmatrix}D+J & 0\\ 0 & 0\end{pmatrix}$$

where $\tilde S$ is obtained from S with $(\theta_{(0)},\beta_{(0)})$ replaced by $(\tilde\theta,\tilde\beta)$ and G_0 replaced by $\widetilde{G_0}$.

Remark 4: Because S^{-1} is the prospectively derived asymptotic variance-covariance matrix of $(\widehat{\theta^*},\hat\beta)$ on the basis of the prospective likelihood function given in Remark 2, it is seen from the expression for the asymptotic variance-covariance matrix Σ of $(\tilde\theta,\tilde\beta)$ that the asymptotic variance-covariance matrices for $\hat\beta$ and $\tilde\beta$ coincide under the retrospective sampling scheme and the prospective sampling scheme. Consequently, a prospectively derived analysis under model (1.1) on parameter estimates and standard errors for β is asymptotically correct in case-control studies. These results match those of Weinberg and Wacholder (1993) and Scott and Wild (1997).

The two-step profile maximization procedure, by which the maximum semiparametric likelihood estimator $(\tilde\theta,\tilde\beta,\widetilde{G_0})$ is derived, relies on first maximizing the nonparametric part G_0 with (θ,β) fixed and then maximizing $\ell(\theta,\beta)$ with respect to (θ,β). The estimator $(\tilde\theta,\tilde\beta,\widetilde{G_0})$ can also be derived by employing the following "method of moments". Motivated by the work of Gill,

Vardi, and Wellner (1988), let $F = \sum_{i=0}^{I} \frac{n_i}{n} G_i$ be the "average distribution function", then by (2)

$$G_i(t) = \frac{n}{n_0} \int \frac{\exp[\theta_i + s_i(y;\beta_i)]}{\sum_{i=0}^{I} \rho_i \exp[\theta_i + s_i(y;\beta_i)]} I_{[y \le t]} dF(y), \quad i = 0,1,\ldots,I.$$

Let $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[T_i \le t]}$ be the empirical distribution function of the pooled sample $\{T_1,\ldots,T_n\}$. Then G_i can be estimated for fixed (θ,β) by

$$\overline{G}_i(t) = \frac{n}{n_0} \int \frac{\exp[\theta_i + s_i(y;\beta_i)]}{\sum_{i=0}^{I} \rho_i \exp[\theta_i + s_i(y;\beta_i)]} I_{[y \le t]} dF_n(y) = \frac{1}{n_0}\sum_{k=1}^{n} \frac{\exp[\theta_i + s_i(T_k;\beta_i)]}{\sum_{i=0}^{I} \rho_i \exp[\theta_i + s_i(T_k;\beta_i)]} I_{[T_k \le t]}$$

for $i = 0,1,\ldots,I$. Let $\widehat{G}_i(t) = n_i^{-1}\sum_{j=1}^{n_i} I_{[X_{ij} \le t]}$ be the empirical distribution function based on the sample $X_{i1},\ldots,X_{in}$ from the i th response category. Let $\psi_i(t;\theta,\beta)$ be a real function from R^p to R^{p+1} for $i = 1,\ldots,I$ and let $\psi(t;\theta,\beta) = (\psi_1^{\tau}(t;\theta,\beta),\ldots,\psi_I^{\tau}(t;\theta,\beta))^{\tau}$. Then, for a particular choice of $\psi(t;\theta,\beta)$, (θ,β) can be estimated by matching the expectation of $n_i\psi_i(t;\theta,\beta)$ under $\overline{G}_i$ with that under $\widehat{G}_i$ for $i = 1,\ldots,I$:

$$E_{\overline{G}_i}[n_i\psi_i(T;\theta,\beta)] = \int n_i\psi_i(t;\theta,\beta)\, d\overline{G}_i(t) = \int n_i\psi_i(t;\theta,\beta)\, d\widehat{G}_i(t) = E_{\widehat{G}_i}[n_i\psi_i(T;\theta,\beta)]$$

for $i = 1,\ldots,I$. In other words, (θ,β) can be estimated by seeking a root to the following system of equations:

$$L_i(\theta,\beta) = \sum_{k=1}^{n} \frac{\rho_i \exp[\theta_i + s_i(T_k;\beta_i)]}{\sum_{m=0}^{I} \rho_m \exp[\theta_m + s_m(T_k;\beta_m)]}\ \psi_i(T_k,\theta,\beta) - \sum_{j=1}^{n_i} \psi_i(X_{ij},\theta,\beta) = 0 \quad i = 1,\ldots,I. \tag{11}$$

It is easy to see that the above system of equations reduces to the system of score equations in (3) if $\psi_i(t;\theta,\beta) = (1, d_i^{\tau}(t;\beta_i))^{\tau}$ is taken for $i = 1,\ldots,I$. Let $(\overline{\theta},\overline{\beta})$ with $\overline{\theta} = (\overline{\theta}_1,\ldots,\overline{\theta}_I)^{\tau}$ and $\overline{\beta} = (\overline{\beta}_1,\ldots,\overline{\beta}_I)^{\tau}$ be a solution to the system of equations in (11). Note that $(\overline{\theta},\overline{\beta})$ depends on the choice of $\psi_i(t;\theta,\beta)$ for $i = 1,\ldots,I$. The following theorem demonstrates that the choice of $\psi_i(t;\theta,\beta) = (1, d_i^{\tau}(t;\beta_i))^{\tau}$ for $i = 1,\ldots,I$ is optimal in the sense that the difference between the asymptotic variance-covariance matrices of $(\overline{\theta},\overline{\beta})$ and $(\tilde{\theta},\tilde{\beta})$ is positive semidefinite for any set of measurable functions $\{\psi_i(t;\theta,\beta):\ i = 1,\ldots,I\}$. Qin (1998) established this optimal property when $I = 1$.

Theorem 2: Under the conditions of Theorem 1, we have

$$\sqrt{n}\begin{pmatrix} \overline{\theta} - \theta_{(0)} \\ \overline{\beta} - \beta_{(0)} \end{pmatrix} \xrightarrow{d} N_{(p+1)I}(0, \Sigma_{\psi}),$$

where $\Sigma_{\psi} = V^{-1}B_{\psi}(V^{\tau})^{-1}$ with V and B_{ψ} defined in (18) of the proof section. Moreover, the maximum semiparametric likelihood estimator $(\tilde{\theta},\tilde{\beta})$ is optimal in the sense that $\Sigma_{\psi} - \Sigma$ is positive semidefinite for any set of measurable functions $\{\psi_i(t;\theta,\beta): i = 1,\ldots,I\}$.

In the following case, $p = 1$ is considered, although the results can be naturally generalized to the case of $p > 1$. The weak convergence of $\sqrt{n}(\tilde{G}_0 - \widehat{G}_0,\ldots,\tilde{G}_I - \widehat{G}_I)^{\tau}$ is

now established to a multivariate Gaussian process by representing $\widetilde{G}_i - \widehat{G}_i \ (i = 0,1,\ldots,I)$ as the mean of a sequence of independent and identically distributed stochastic processes with a remainder term of order $o_p(n^{-1/2})$.

Theorem 3: Suppose that model (2) and Assumptions (A1)−(A4) hold. Suppose further that S is positive definite. For $i = 0,1,\ldots,I$, one can write

$$\widetilde{G}_i(t) - \widehat{G}_i(t) = H_{1i}(t) - \widehat{G}_i(t) - H_{2i}(t) + R_{in}(t), \tag{12}$$

where

$$H_{1i}(t) = \frac{1}{n_0}\sum_{k=1}^{n} \frac{\exp[\theta_{i0} + s_i(T_k;\beta_{i0})]}{1 + \sum_{m=1}^{I} \rho_m \exp[\theta_{m0} + s_m(T_k;\beta_{m0})]} I_{[T_k \le t]},$$

$$H_{2i}(t) = \frac{1}{n\rho_i}(A_{1i}^{\tau}(t), A_{2i}^{\tau}(t))S^{-1}\begin{pmatrix} \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial \theta} \\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial \beta} \end{pmatrix}, \tag{13}$$

and the remainder term $R_{in}(t)$ satisfies

$$\sup_{-\infty \le t \le \infty} |R_{in}(t)| = o_p(n^{-1/2}). \tag{14}$$

As a result,

$$\sqrt{n}\begin{pmatrix} \widetilde{G}_0 - \widehat{G}_0 \\ \widetilde{G}_1 - \widehat{G}_1 \\ \vdots \\ \widetilde{G}_I - \widehat{G}_I \end{pmatrix} \xrightarrow{D} \begin{pmatrix} W_0 \\ W_1 \\ \vdots \\ W_I \end{pmatrix}$$

$$in \quad D^{I+1}[-\infty,\infty] \tag{15}$$

where $D^{I+1}[-\infty,\infty]$ is the product space defined by $D[-\infty,\infty]\times\cdots\times D[-\infty,\infty]$ and $(W_0, W_1, \ldots, W_I)^{\tau}$ is a multivariate Gaussian process with continuous sample path and satisfies, for $-\infty \le s \le t \le \infty$,

$$EW_i(t) = 0, \qquad i = 0,1,\ldots,I,$$

$$EW_i(s)W_i(t) = \frac{1+\rho}{\rho_i^2}[G_i(s) - B_{ii}(s)]$$

$$-\frac{1}{\rho_i^2}(A_{1i}^{\tau}(s), A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix} A_{1i}(t) \\ A_{2i}(t) \end{pmatrix}, \qquad i = 0,1,\ldots,I,$$

$$EW_i(s)W_j(t) = -\frac{1+\rho}{\rho_i\rho_j}B_{ij}(s) - \frac{1}{\rho_i\rho_j}(A_{1i}^{\tau}(s), A_{2i}^{\tau}(s))S^{-1}$$

$$\begin{pmatrix} A_{1j}(t) \\ A_{2j}(t) \end{pmatrix}, \qquad i \ne j = 0,1,\ldots,I. \tag{16}$$

Theorem 3 forms the basis for testing the validity of model (2) on the basis of the test statistic Δ_n in (6). Let w_α denote the α-quantile of the distribution of $\frac{1}{I+1}\sum_{i=0}^{I}\rho_i\{\sup_{-\infty\le t\le\infty}|W_i(t)|\}$, i.e., w_α satisfies

$$P(\tfrac{1}{I+1}\textstyle\sum_{i=0}^{I}\rho_i\{\sup_{-\infty\le t\le\infty}|W_i(t)|\} \le w_\alpha) = \alpha.$$

According to Theorem 3 and the continuous Mapping Theorem (Billingsley, 1968, p. 30):

$$\lim_{n\to\infty} P(\Delta_n \ge w_{1-\alpha})$$

$$= \lim_{n\to\infty} P\left(\frac{1}{I+1}\sum_{i=0}^{I}\rho_i\{\sup_{-\infty\le t\le\infty}\sqrt{n}\,|\widetilde{G}_i(t) - \widehat{G}_i(t)|\} \ge w_{1-\alpha}\right)$$

$$= P\left(\frac{1}{I+1}\sum_{i=0}^{I}\rho_i\left\{\sup_{-\infty\le t\le\infty}|W_i(t)|\right\} \ge w_{1-\alpha}\right) = \alpha$$

Thus, the proposed goodness of fit test procedure has the following decision rule: reject model (2) at level α if $\Delta_n > w_{1-\alpha}$. In order for this proposed test procedure to be useful in practice, the distribution of $\frac{1}{I+1}\sum_{i=0}^{I}\rho_i\{\sup_{-\infty\le t\le\infty}|W_i(t)|\}$ must be found and the $(1-\alpha)$-quantile $w_{1-\alpha}$ calculated.

Unfortunately, no analytic expressions appear to be available for the distribution function of $\frac{1}{I+1}\sum_{i=0}^{I}\rho_i\{\sup_{-\infty\le t\le\infty}|W_i(t)|\}$ and the quantile function thereof. A way out is to employ a bootstrap procedure as described in the next section.

A Bootstrap Procedure

In this section is presented a bootstrap procedure which can be employed to approximate the quantile $w_{1-\alpha}$ defined at the end of the last section. If model (1) is valid, since $\theta^* = (\theta_1^*,\ldots,\theta_I^*)^\tau$ is not estimable in general on the basis of the case-control data $T_1,\ldots,T_n$, only generated data, respectively, from $\widetilde{G}_0,\widetilde{G}_1,\ldots,\widetilde{G}_I$, where $\widetilde{G}_i$ $(i=0,1,\ldots,I)$ is given by (5). Specifically, let $X_{i1}^*,\ldots,X_{in_i}^*$ be a random sample from $\widetilde{G}_i$ for $i=0,1,\ldots,I$ and assume that $\{(X_{i1}^*,\ldots,X_{in_i}^*):\ i=0,1,\ldots,I\}$ are jointly independent. Let $\{T_1^*,\ldots,T_n^*\}$ denote the combined bootstrap sample $\{X_{01}^*,\ldots,X_{0n_0}^*;X_{11}^*,\ldots,X_{1n_1}^*;\cdots;X_{I1}^*,\ldots,X_{In_I}^*\}$ and $(\widetilde{\theta}^*,\widetilde{\beta}^*)$ with $\widetilde{\theta}^* = (\widetilde{\theta}_1^*,\ldots,\widetilde{\theta}_I^*)^\tau$ and $\widetilde{\beta}^* = (\widetilde{\beta}_1^{*\tau},\ldots,\widetilde{\beta}_I^{*\tau})^\tau$ be the solution to the system of score equations in (3) with the T_k^* in place of the T_k. Moreover, similar to (4)−(6), let $\widehat{G}_i^*(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} I_{[X_{ij}^*\le t]}$ for $i=0,1,\ldots,I$ and

$$\widetilde{p}_k^* = \frac{1}{n_0}\frac{1}{1+\sum_{i=1}^{I}\rho_i\exp[\widetilde{\theta}_i^* + s_i(T_k^*;\widetilde{\beta}_i^*)]},\quad k=1,\ldots,n.$$

$$\widetilde{G}_i^*(t) = \sum_{k=1}^{n}\widetilde{p}_k^*\exp[\widetilde{\theta}_i^* + s_i(T_k^*;\widetilde{\beta}_i^*)]I_{[T_k^*\le t]}$$

$$= \frac{1}{n_0}\sum_{k=1}^{n}\frac{\exp[\widetilde{\theta}_i^* + s_i(T_k^*;\widetilde{\beta}_i^*)]I_{[T_k^*\le t]}}{1+\sum_{m=1}^{I}\rho_m\exp[\widetilde{\theta}_m^* + s_m(T_k^*;\widetilde{\beta}_m^*)]},\quad i=0,\ldots,I,$$

where $\widetilde{\theta}_0^* = 0$ and $s_0(\bullet;\widetilde{\beta}_0^*)\equiv 0$. Then the corresponding bootstrap version of the test statistic Δ_n in (6) is given by

$$\Delta_n^* = \frac{1}{I+1}\sum_{i=0}^{I}\rho_i\Delta_{ni}^*,$$

where $\Delta_{ni}^* = \sup_{-\infty\le t\le\infty}|\Delta_{ni}^*(t)|$ with $\Delta_{ni}^*(t) = \sqrt{n}(\widetilde{G}_i^*(t) - \widehat{G}_i^*(t))$ for $i=0,1,\ldots,I$. To see the validity of the proposed bootstrap procedure, the proofs of Theorems 1 and 3 can be mimicked with slight modification to show the following theorem. The details are omitted here.

Theorem 4: Suppose that model (2) and Assumptions (A1)−(A4) hold. Suppose further that S is positive definite and

$$\int_{-\infty}^{-\infty} Q_1^2(y)Q_2(y)\{1+\rho_i\exp[\theta_{i0}+s_i(y;\beta_{i0})]\}dG_0(y) < \infty$$

for $i=1,\ldots,I$.

(a) Along almost all sample sequences $T_1,T_2,\ldots$, given $(T_1,\ldots,T_n)$, as $n\to\infty$, we have

$$\sqrt{n}\begin{pmatrix}\widetilde{\theta}^* - \widetilde{\theta}\\ \widetilde{\beta}^* - \widetilde{\beta}\end{pmatrix} \xrightarrow{d} N_{(p+1)I}(0,\Sigma).$$

(b) Along almost all sample sequences $T_1,T_2,\ldots$, given $(T_1,\ldots,T_n)$, as $n\to\infty$, we have

$$\sqrt{n}\begin{pmatrix}\widetilde{G}_0^* - \widehat{G}_0^*\\ \widetilde{G}_1^* - \widehat{G}_1^*\\ \vdots\\ \widetilde{G}_I^* - \widehat{G}_I^*\end{pmatrix} \xrightarrow{D} \begin{pmatrix}W_0\\ W_1\\ \vdots\\ W_I\end{pmatrix} \quad \text{in} \quad D^{I+1}[-\infty,\infty],$$

where $(W_0,W_1,\ldots,W_I)^\tau$ is the multivariate Gaussian process defined in Theorem 3.

Theorem 3 and part (b) of Theorem 4 indicate that the limit process of $\sqrt{n}(\widetilde{G}_0^* - \widehat{G}_0^*,\ldots,\widetilde{G}_I^* - \widehat{G}_I^*)^\tau$ agrees with that

of $\sqrt{n}(\widetilde{G}_0 - \widehat{G}_0, \ldots, \widetilde{G}_I - \widehat{G}_I)^\tau$. It follows from the Continuous Mapping Theorem that $\Delta_n^* = \frac{1}{I+1}\sum_{i=0}^{I} \rho_i \Delta_{ni}^*$ has the same limiting behavior as does $\Delta_n = \frac{1}{I+1}\sum_{i=0}^{I} \rho_i \Delta_{ni}$. Thus, the quantiles of the distribution of Δ_n can be approximated by those of Δ_n^*. For $\alpha \in (0,1)$, let $w_{1-\alpha}^n = \inf\{t;\ P^*(\Delta_n^* \le t) \ge 1-\alpha\}$, where P^* stands for the bootstrap probability under $\widetilde{G}_i$ $(i = 0,1,\ldots,I)$. Then there is the following bootstrap decision rule: reject model (2) at level α if $\Delta_n > w_{1-\alpha}^n$.

Two real data sets are next considered. Note that the multivariate logistic regression model is a special case of the multivariate multiplicative-intercept risk model (1) with $\theta_i^* = \exp(\alpha_i^*)$ and $r_i(x;\beta_i) = \exp(\beta_i^\tau x)$ for $i = 1,\ldots,I$. In this case, we have $\theta_i = \alpha_i^* + \log(\frac{\pi_0}{\pi_1})$ and $s_i(x;\beta_i) = \beta_i^\tau x$ in model (2) for $i = 1,\ldots,I$.

Example 1: Agresti (1990) analyzed, by employing the continuation-ratio logit model, the relationship between the concentration level of an industrial solvent and the outcome for pregnant mice in a developmental toxicity study. The complete dataset is listed on page 320 in his book. Let X denote "concentration level (in mg/kg per day)" and Y represent "pregnancy outcome", in which $Y = 0,1,$ and 2 stand for three possible outcomes: Normal, Malformation, and Non-live. Here this data set is analyzed on the basis of the multivariate logistic regression model. Because the sample data (X_i, Y_i) , $i = 1,\ldots,1435$, can be thought as being drawn independently and identically from the joint distribution of (X,Y), Remark 1 implies that the test statistic Δ_n in (6) can be used to test the validity of the multivariate logistic regression model. Under model (2),

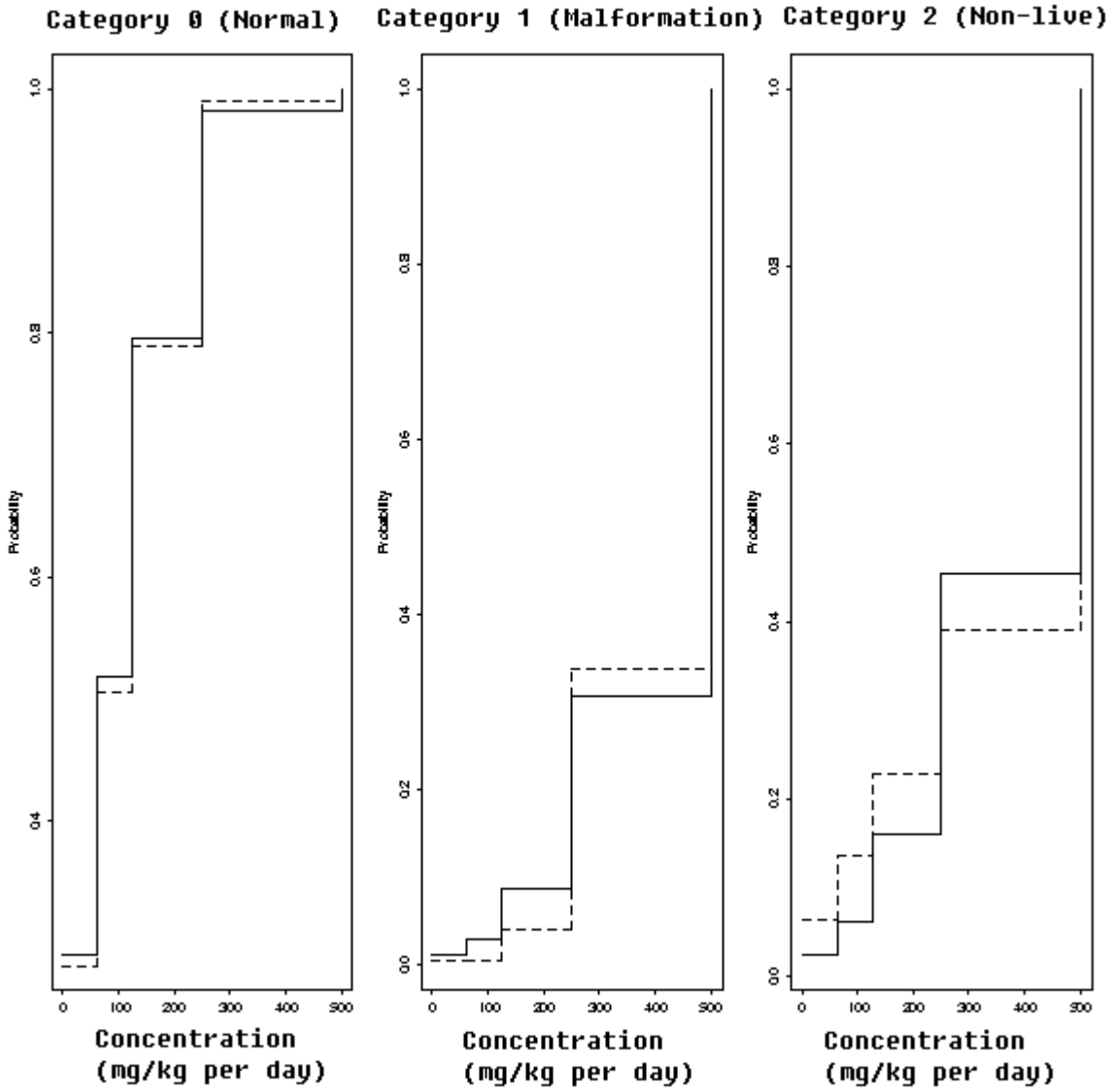
$$(\widetilde{\theta}_1, \widetilde{\beta}_1, \widetilde{\theta}_2, \widetilde{\beta}_2) = (-3.33834,\ 0.01401,\ -2.52553,\ 0.01191)$$

and $\Delta_n = 0.49439$ with the observed P -value equal to 0 based on 1000 bootstrap replications of Δ_n^* . Because $n_0 = 1000$, $n_1 = 199$, and $n_2 = 236$, $\alpha_1^* = \log\theta_1^*$ and $\alpha_2^* = \log\theta_2^*$ can be estimated by

$\widetilde{\alpha}_1^* = -3.33834 + \log(199/1000) = -4.95279$ and $\widetilde{\alpha}_2^* = -2.52553 + \log(236/1000) = -3.96945$, respectively.

Figure 1 shows the curves of $\widetilde{G}_0$ and $\widehat{G}_0$ (left panel), the curves of $\widetilde{G}_1$ and $\widehat{G}_1$ (middle panel), and the curves of $\widetilde{G}_2$ and $\widehat{G}_2$ (right panel) based on this data set. The middle and right panels indicate strong evidence of the lack of fit of the multivariate logistic regression model to these data within the categories for Malformation and Non-live.

Figure 1. Example 1: Developmental toxicity study with pregnant mice. Left panel: estimated cumulative distribution functions $\tilde{G}_0$ (solid curve) and $\hat{G}_0$ (dashed curve). Middle panel: estimated cumulative distribution functions $\tilde{G}_1$ (solid curve) and $\hat{G}_1$ (dashed curve). Right panel: estimated cumulative distribution functions $\tilde{G}_2$ (solid curve) and $\hat{G}_2$ (dashed curve).



Example 2: Table 9.12 in Agresti (1990, p. 339) contains data for the 63 alligators caught in Lake George. Here the relationship between the alligator length and the primary food choice of alligators is analyzed by employing the multivariate logistic regression model. Let X denote "length of alligator (in meters)" and Y represent "primary food choice" in which $Y = 0, 1,$ and 2 stand for three categories: Other, Fish, and Invertebrate. Since the sample data (X_i, Y_i), $i = 1, \ldots, 63$, can be thought as being drawn independently and identically from the joint distribution of (X, Y), Remark 1 implies that the test statistic Δ_n in (6) can be used to test the validity of the multivariate logistic regression model.

For the male data, we find $(\widetilde{\theta}_1, \widetilde{\beta}_1, \widetilde{\theta}_2, \widetilde{\beta}_2)$ = (0.41781, -0.17678, 4.83809, -2.60093) and Δ_n = 1.33460 with the observed P-value identical to 0.389 based on 1000 bootstrap replications of Δ_n^*. For the female data, we find $(\widetilde{\theta}_1, \widetilde{\beta}_1, \widetilde{\theta}_2, \widetilde{\beta}_2)$ = ($-$ 5.58723, 2.57174, 2.70962, -1.50304) and Δ_n = 1.63346 with the observed P-value equal to 0.249 based on 1000 bootstrap replications of Δ_n^*. For the combined male and female data, $(\widetilde{\theta}_1, \widetilde{\beta}_1, \widetilde{\theta}_2, \widetilde{\beta}_2)$ = (-0.19542, 0.08481, 4.48780, -2.38837) and Δ_n = 1.73676 is found with the observed P-value identical to 0.225 based on 1000 bootstrap replications of Δ_n^*, indicating that we can ignore the gender effect on primary food choice. Because n_0 = 10, n_1 = 33, and n_2 = 20, $\alpha_1^* = \log \theta_1^*$ and $\alpha_2^* = \log \theta_2^*$ can be estimated by $\widetilde{\alpha}_1^*$ = $-$ 0.19542 + log(33/10) = 0.99850 and $\widetilde{\alpha}_2^*$ = 4.48780 + log(20/10) = 5.18094, respectively.

Figures 2-4 display the curves of $\widetilde{G}_0$ and $\widehat{G}_0$ (left panel), the curves of $\widetilde{G}_1$ and $\widehat{G}_1$ (middle panel), and the curves of $\widetilde{G}_2$ and $\widehat{G}_2$ (right panel) based, respectively, on the male, female, and combined data set. For the combined data, the curve of $\widetilde{G}_1(\widetilde{G}_2)$ bears a resemblance to that of $\widehat{G}_1(\widehat{G}_2)$, whereas the dissimilarity between the curves of $\widetilde{G}_0$ and $\widehat{G}_0$ indicates some evidence of lack of fit of the multivariate logistic regression model to these data within the baseline category for Other.

Proofs

First presented are four lemmas, which will be used in the proof of the main results. The proofs of Lemmas 1, 2, and 3 are lengthy yet straightforward and are therefore omitted here. Throughout this section, the norm of a $m_1 \times m_2$ matrix $A = (a_{ij})_{m_1 \times m_2}$ is defined by

$$\| A \| = \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} a_{ij}^{\ 2} \right)^{1/2} \quad \text{for} \quad m_1, m_2 \geq 1.$$

Furthermore, in addition to the notation in (8) we introduce some further notation. Write

$$Q_{i11} = (s_{11}^{1i}, \cdots, s_{11}^{Ii})^{\tau}, \quad Q_{i21} = ((s_{21}^{1i})^{\tau}, \cdots, (s_{21}^{Ii})^{\tau})^{\tau},$$

$$Q_i = \begin{pmatrix} Q_{i11} \\ Q_{i21} \end{pmatrix}, \quad i = 0,\ 1,\ \cdots, I,$$

$$S_{n11} = -\frac{1}{n} \frac{\partial^2 \ell(\theta_{(0)}, \beta_{(0)})}{\partial \theta \, \partial \theta^{\tau}},$$

$$S_{n21} = -\frac{1}{n} \frac{\partial^2 \ell(\theta_{(0)}, \beta_{(0)})}{\partial \beta \, \partial \theta^{\tau}},$$

$$S_{n22} = -\frac{1}{n} \frac{\partial^2 \ell(\theta_{(0)}, \beta_{(0)})}{\partial \beta \, \partial \beta^{\tau}}, \quad S_n = \begin{pmatrix} S_{n11} & S_{n21}^{\tau} \\ S_{n12} & S_{n22} \end{pmatrix}$$

$$H_{0i}(t) = \frac{1}{n_i} \sum_{k=1}^{n} \frac{C_{1i}(T_k; \theta_{(0)}, \beta_{(0)}) I_{[T_k \leq t]}}{\{1 + \sum_{m=1}^{I} \rho_m \exp[\theta_{m0} + s_m(T_k; \beta_{m0})]\}^2},$$

$$i = 0,\ 1,\ \cdots, I,$$

Figure 2. Example 2: Primary food choice for 39 male Florida alligators. Left panel: estimated cumulative distribution functions $\widetilde{G}_0$ (solid curve) and $\hat{G}_0$ (dashed curve). Middle panel: estimated cumulative distribution functions $\widetilde{G}_1$ (solid curve) and $\hat{G}_1$ (dashed curve). Right panel: estimated cumulative distribution functions $\widetilde{G}_2$ (solid curve) and $\hat{G}_2$ (dashed curve).

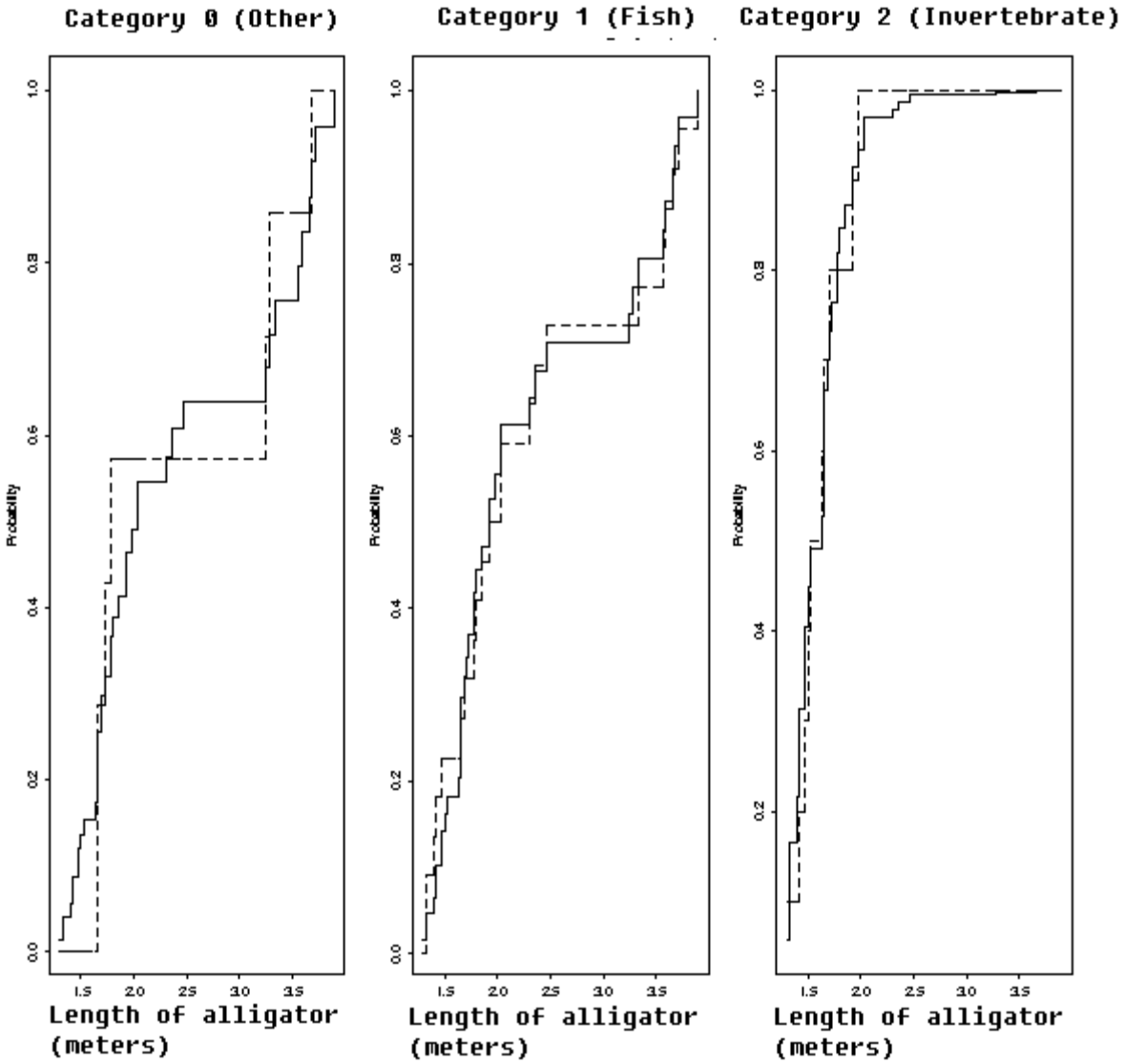


Figure 3. Example 2: Primary food choice for 24 female Florida alligators. Left panel: estimated cumulative distribution functions $\tilde{G}_0$ (solid curve) and $\hat{G}_0$ (dashed curve). Middle panel: estimated cumulative distribution functions $\tilde{G}_1$ (solid curve) and $\hat{G}_1$ (dashed curve). Right panel: estimated cumulative distribution functions $\tilde{G}_2$ (solid curve) and $\hat{G}_2$ (dashed curve).

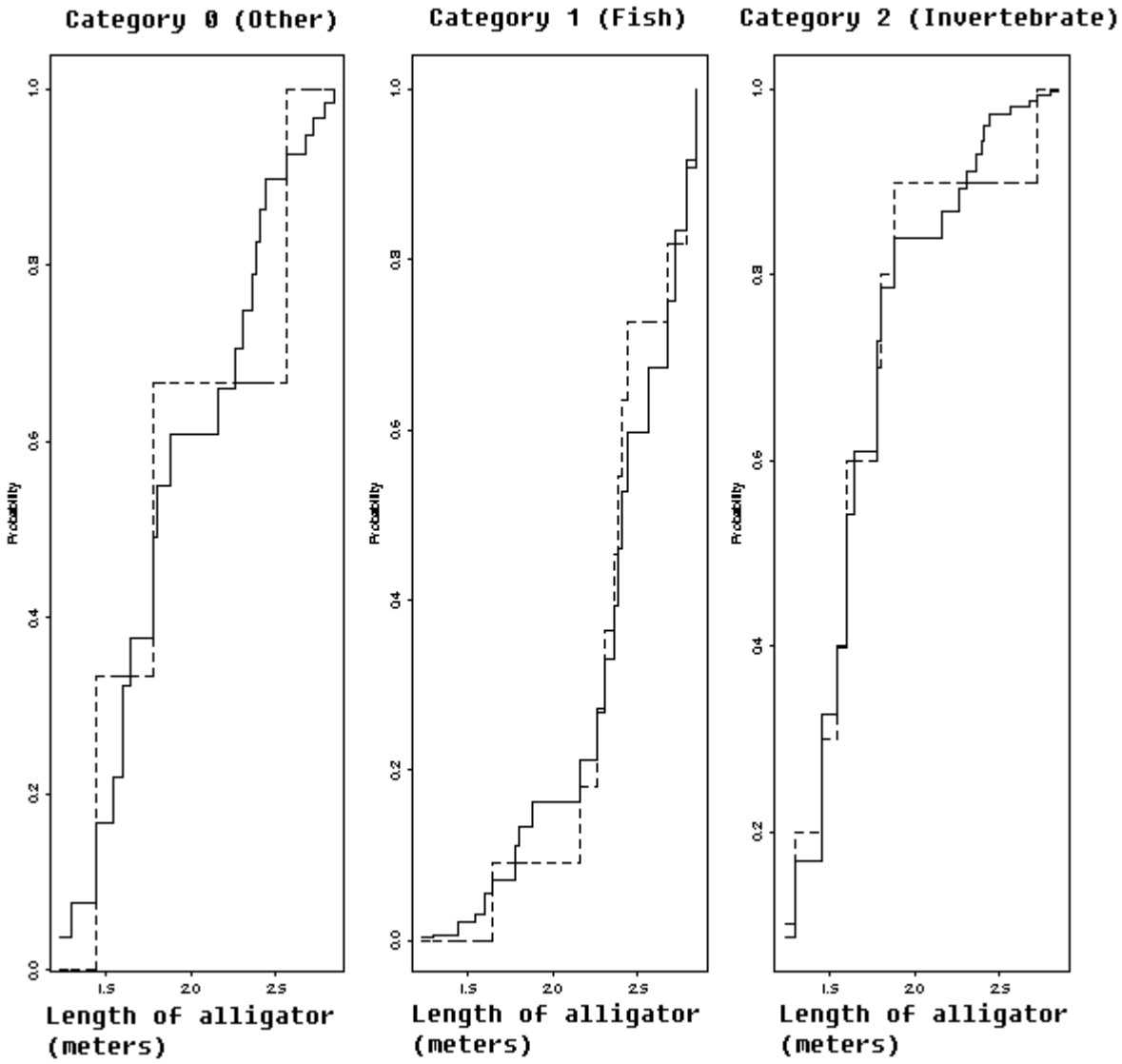
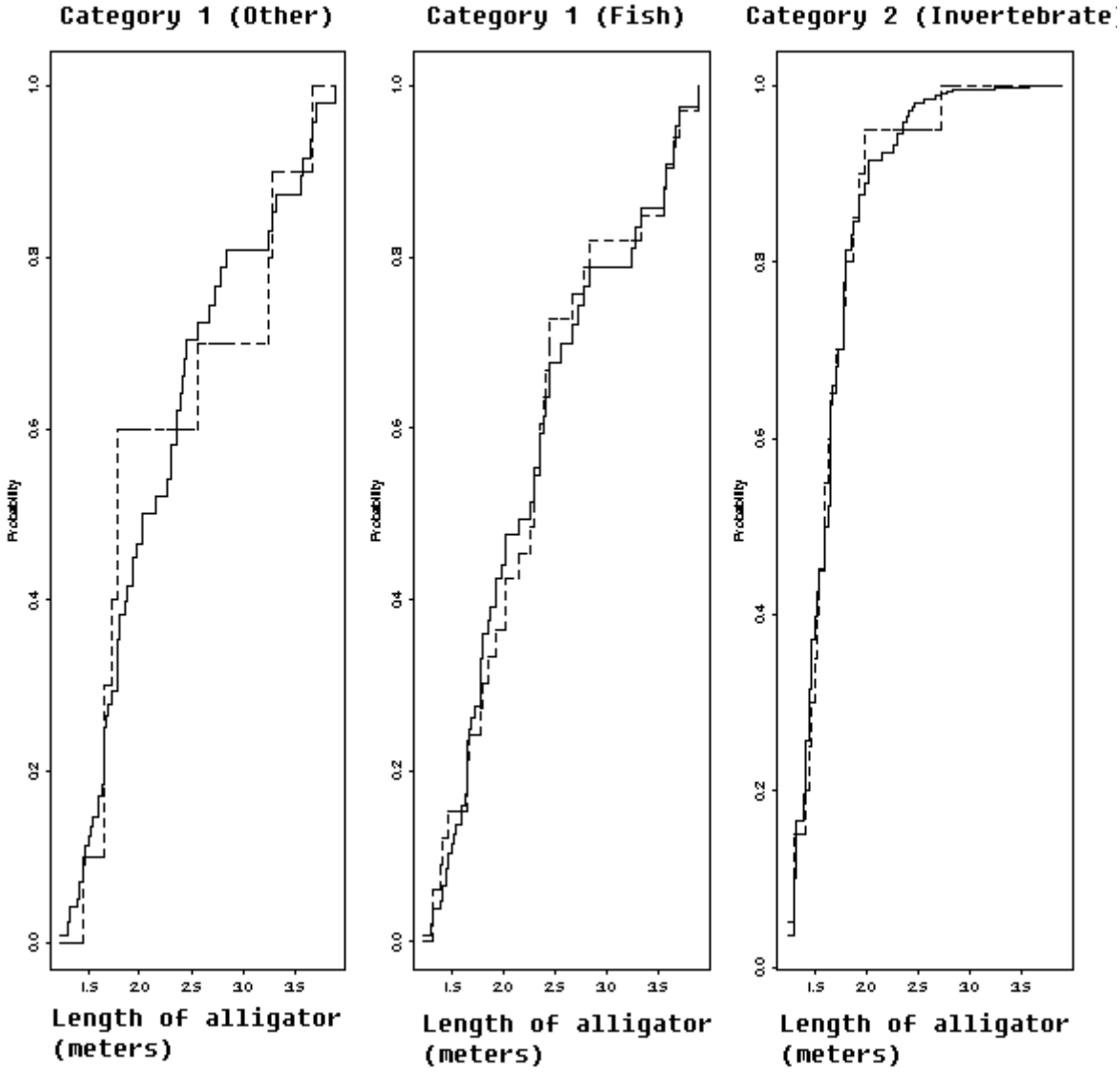


Figure 4. Example 2: Primary food choice for 63 male and female Florida alligators. Left panel: estimated cumulative distribution functions $\tilde{\tilde{G}}_0$ (solid curve) and $\hat{G}_0$ (dashed curve). Middle panel: estimated cumulative distribution functions $\tilde{\tilde{G}}_1$ (solid curve) and $\hat{G}_1$ (dashed curve). Right panel: estimated cumulative distribution functions $\tilde{\tilde{G}}_2$ (solid curve) and $\hat{G}_2$ (dashed curve).



$$H_{3i}(t)=\frac{1}{n_i}\sum_{k=1}^{n}\frac{C_{2i}(T_k;\theta_{(0)},\beta_{(0)})I_{[T_k\le t]}}{\{1+\sum_{m=1}^{I}\rho_m\exp[\theta_{m0}+s_m(T_k;\beta_{m0})]\}^2},$$
$$i=0,\ 1,\ \cdots,I.$$

Lemma 1: Suppose that model (2) holds and S is positive definite. Let J be an $I\times I$ matrix of 1 elements and let $D=\mathrm{Diag}(\rho_1^{-1},\cdots,\rho_I^{-1})$ denote the $I\times I$ diagonal matrix having elements $\{\rho_1^{-1},\cdots,\rho_I^{-1}\}$ on the main diagonal, then

$$B\equiv\frac{1}{n}\mathrm{Var}\begin{pmatrix}\dfrac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}=S-\sum_{i=0}^{I}\frac{1+\rho}{\rho_i}Q_iQ_i^{\tau},$$
$$S^{-1}BS^{-1}=S^{-1}-(1+\rho)\begin{pmatrix}D+J & 0\\ 0 & 0\end{pmatrix}=\Sigma.$$

Lemma 2: Suppose that model (2) holds and S is positive definite. For $-\infty\le s\le t\le\infty$,

$$\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)])$$
$$=\frac{1+\rho}{\rho_i^2}[G_i(s)-B_{ii}(s)]$$
$$-\frac{1+\rho}{\rho_i^2}\sum_{k=0,k\ne i}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{ik}(t)$$
$$-\frac{1+\rho}{\rho_i^3}[G_i(s)-B_{ii}(s)][G_i(t)-B_{ii}(t)],$$
$$i=0,\ 1,\ \cdots,I,$$

$$\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\sqrt{n}[H_{1j}(t)-\hat{G}_j(t)])$$
$$=-\frac{1+\rho}{\rho_i\rho_j}B_{ij}(s)-\frac{1+\rho}{\rho_i\rho_j}\sum_{k=0}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{jk}(t)$$
$$+\frac{1+\rho}{\rho_i\rho_j^2}B_{ij}(s)G_j(t)+\frac{1+\rho}{\rho_i^2\rho_j}G_i(s)B_{ij}(t),$$
$$i\ne j=0,\ 1,\ \cdots,I.$$

Lemma 3: Suppose that model (2) holds and S is positive definite. For $-\infty\le s\le t\le\infty$, we have

$$\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\ \sqrt{n}H_{2i}(t))$$
$$=\mathrm{Cov}(\sqrt{n}H_{2i}(s),\ \sqrt{n}H_{2i}(t))$$
$$=\frac{1}{\rho_i^2}(A_{1i}^{\tau}(t),A_{2i}^{\tau}(t))S^{-1}\begin{pmatrix}A_{1i}(s)\\ A_{2i}(s)\end{pmatrix}$$
$$-\frac{1+\rho}{\rho_i^2}\sum_{k=0,k\ne i}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{ik}(t)$$
$$-\frac{1+\rho}{\rho_i^3}[G_i(s)-B_{ii}(s)][G_i(t)-B_{ii}(t)],$$
$$i=0,\ 1,\ \cdots,I,$$

$$\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\sqrt{n}H_{2j}(t))$$
$$=\mathrm{Cov}(\sqrt{n}H_{2i}(s),\sqrt{n}H_{2j}(t))$$
$$=\frac{1}{\rho_i\rho_j}(A_{1i}^{\tau}(s),A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix}A_{1j}(t)\\ A_{2j}(t)\end{pmatrix}$$
$$-\frac{1+\rho}{\rho_i\rho_j}\sum_{k=0}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{jk}(t)$$
$$+\frac{1+\rho}{\rho_i\rho_j^2}B_{ij}(s)G_j(t)+\frac{1+\rho}{\rho_i^2\rho_j}G_i(s)B_{ij}(t),$$
$$i\ne j=0,\ 1,\ \cdots,I.$$

Lemma 4: Suppose that model (2) and Assumption (A2) hold. If S is positive definite and G_0 is continuous, then the stochastic process $\{\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)-H_{2i}(t)],\ -\infty\le t\le\infty\}$ is tight in $D[-\infty,\infty]$ for $i=0,1,\cdots,I$, where $H_{1i}(t)$ and $H_{2i}(t)$ are defined in (13).

Proof: Because $n=\frac{1+\rho}{\rho_i}n_i$ for $i=0,1,\cdots,I$, it can be shown after some algebra that

$$\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)-H_{2i}(t)]$$
$$=\frac{1}{\rho_i}\sum_{k=0,k\neq i}^{I}\sqrt{\frac{1+\rho}{\rho_k}}\sqrt{n_k}U_{ik}(t)$$
$$-\frac{1}{\rho_i}\sqrt{\frac{1+\rho}{\rho_i}}\sqrt{n_i}U_{ii}(t)-\sqrt{n}H_{2i}(t), \qquad (17)$$

where

$$U_{ii}(t)=\frac{1}{n_i}$$
$$\sum_{j=1}^{n_i}\frac{\sum_{m=0,m\neq i}^{I}\rho_m\exp[\theta_{m0}+s_m(X_{ij};\beta_{m0})]}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_{m0}+s_m(X_{ij};\beta_{m0})]}\rho_i I_{[X_{ij}\le t]}$$
$$-[G_i(t)-B_{ii}(t)],$$

$$U_{ik}(t)=\frac{1}{n_k}$$
$$\sum_{j=1}^{n_k}\frac{\rho_i\exp[\theta_{i0}+s_i(X_{kj};\beta_{i0})]\rho_k I_{[X_{kj}\le t]}}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_{m0}+s_m(X_{kj};\beta_{m0})]}$$
$$-B_{ik}(t), \quad k\neq i=0,\ 1,\ \cdots,I.$$

Let $\Im=\{I_{(-\infty,t]}: t\in \mathrm{R}\}$ be the collection of all indicator functions of cells $(-\infty,t]$ in R. According to the classical empirical process theory, $\Im$ is a $P_{X_{k1}}$ -Donsker class for $k=0,1,\cdots,I,$ where $P_{X_{k1}}=P\circ X_{k1}^{-1}$ is the law of X_{k1} for $k=0,\ 1,\ \cdots,I.$ For each $i=0,1,\cdots,I,$ let us define $I+1$ fixed functions $f_{i0}, f_{i1},\cdots, f_{iI}$ by

$$f_{ii}(y)=\frac{\rho_i\sum_{m=0,m\neq i}^{I}\rho_m\exp[\theta_{m0}+s_m(y;\beta_{m0})]}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_{m0}+s_m(y;\beta_{m0})]},$$

$$f_{ik}(y)=\frac{\rho_k\rho_i\exp[\theta_{i0}+s_i(y;\beta_{i0})]}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_{m0}+s_m(y;\beta_{m0})]},$$
$$k\neq i=0,\ 1,\ \cdots,I.$$

Then it is seen that $f_{i0}, f_{i1},\cdots, f_{iI}$ are uniformly bounded functions. According to Example 2.10.10 of van der Vaart and Wellner (1996, p. 192), it can be concluded that $\Im\cdot f_{ik}$ is a $P_{X_{k1}}$ -Donsker class for $k=0,\ 1,\ \cdots,I.$

Let $P_{n_k}=\frac{1}{n_k}\sum_{j=1}^{n_k}\delta_{X_{kj}}$ be the empirical measure of $X_{k1},\cdots,X_{kn_k}$ for $k=0,1,\cdots,I,$ where δ_x is the measure with mass one at x. Then, it can be shown that

$$\sqrt{n_k}(P_{n_k}-P_{X_{k1}})(I_{(-\infty,t]}f_{ik})$$
$$=\sqrt{n_k}U_{ik}(t), \qquad i,k=0,1,\cdots,I.$$

As a result, there exist $I+1$ zero-mean Gaussian processes $V_{i0},V_{i1},\cdots,V_{iI}$ such that

$$\sqrt{n_k}U_{ik}\xrightarrow{D}V_{ik} \quad \text{on } D[-\infty,\infty],$$
$$i,k=0,1,\cdots,I$$

Thus, the stochastic process $\{\sqrt{n_k}U_{ik}(t),\ -\infty\le t\le\infty\}$ is tight on $D[-\infty,\infty]$ for $i,k=0,1,\cdots,I$. Moreover, it can be shown by using the tightness axiom (Sen & Singer, 1993, p. 330) that the stochastic process $\{\sqrt{n}H_{2i}(t),\ -\infty\le t\le\infty\}$ is tight on $D[-\infty,\infty]$ for $i=0,1,\cdots,I$. These results, along with (17), imply that the stochastic process

$\{\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)-H_{2i}(t)],\ -\infty\le t\le\infty\}$

is tight in $D[-\infty,\infty]$ for $i=0,1,\cdots,I$. The proof is complete.

Proof of Theorem 1: Fro part (a), let $B_\varepsilon=\{(\theta,\beta):\|\theta-\theta_{(0)}\|^2+\|\beta-\beta_{(0)}\|^2\le\varepsilon^2\}$ be the ball with center at the true parameter point $(\theta_{(0)},\beta_{(0)})$ and radius ε for some $\varepsilon>0$. For small ε, it can be shown that we can expand $n^{-1}\ell(\theta,\beta)$ on the surface of B_ε about $(\theta_{(0)},\beta_{(0)})$ to find

$$\frac{1}{n}\ell(\theta,\beta)-\frac{1}{n}\ell(\theta_{(0)},\beta_{(0)})=W_{n1}+W_{n2}+W_{n3},$$

where

$$W_{n1}=(\theta^\tau-\theta_{(0)}^\tau,\beta^\tau-\beta_{(0)}^\tau)\frac{1}{n}\begin{pmatrix}\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix},$$

$$W_{n2}=-\frac{1}{2}(\theta^\tau-\theta_{(0)}^\tau,\beta^\tau-\beta_{(0)}^\tau)S_n\begin{pmatrix}\theta-\theta_{(0)}\\ \beta-\beta_{(0)}\end{pmatrix},$$

and W_{n3} satisfies $|W_{n3}| \le c_3\varepsilon^3$ for some constant $c_3>0$ and sufficiently large n with probability 1. Because $\frac{1}{n}\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta} \xrightarrow{a.s.} 0$ and $\frac{1}{n}\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta} \xrightarrow{a.s.} 0$ by the strong law of large numbers, it follows that for any given $\varepsilon>0$, with probability 1 $|W_{n1}| \le 2\varepsilon^3$ for sufficiently large n. Furthermore, because $S_n \xrightarrow{a.s.} S$ again by the strong law of large numbers, it follows that with probability 1, $\|S_n-S\| < 2\varepsilon$ for sufficiently large n. Because S is positive definite, on the surface of B_ε there is,

$$-\frac{1}{2}(\theta^\tau-\theta_{(0)}^\tau,\beta^\tau-\beta_{(0)}^\tau)S\begin{pmatrix}\theta-\theta_{(0)}\\ \beta-\beta_{(0)}\end{pmatrix}$$
$$\le-\frac{\varepsilon^2}{2}\inf_{x\neq0}\frac{x^\tau Sx}{x^\tau x}\le-\frac{\varepsilon^2}{2}\lambda_1,$$

where $\lambda_1>0$ is the smallest eigenvalue of S. As a result, $W_{n2}<-c_2\varepsilon^2$ for sufficiently large n with probability 1 with $c_2=\frac{\lambda_1}{2}-\varepsilon>0$ for sufficiently small $\varepsilon>0$. Consequently, if $\varepsilon<\frac{c_2}{2+c_3}$, then on the surface of B_ε,

$$\frac{1}{n}\ell(\theta,\beta)-\frac{1}{n}\ell(\theta_{(0)},\beta_{(0)})$$
$$\le|W_{n1}|+|W_{n2}|+|W_{n3}|\le 2\varepsilon^3-c_2\varepsilon^2+c_3\varepsilon^3<0$$

for sufficiently large n with probability 1. It has been shown that for any sufficiently small $\varepsilon>0$ and sufficiently large n, with probability 1, $\ell(\theta,\beta)<\ell(\theta_{(0)},\beta_{(0)})$ at all points (θ,β) on the surface of B_ε, and hence that $\ell(\theta,\beta)$ has a local maximum in the interior of B_ε. Because at a local maximum the score equations (3) must be satisfied it follows that for any sufficiently small $\varepsilon>0$ and sufficiently large n, with probability 1, the system of score equations (3) has a solution $(\tilde\theta,\tilde\beta)$ within B_ε. Because $\varepsilon>0$ is arbitrary, $(\tilde\theta,\tilde\beta)$ is strongly consistent for estimating $(\theta_{(0)},\beta_{(0)})$, i.e., $(\tilde\theta,\tilde\beta) \xrightarrow{a.s.} (\theta_{(0)},\beta_{(0)})$.

For part (b), since $(\tilde\theta,\tilde\beta)$ is strongly consistent by part (a), expanding $\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}$ and $\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}$ at $(\theta_{(0)},\beta_{(0)})$ gives

$$0=\frac{\partial\ell(\tilde\theta,\tilde\beta)}{\partial\theta}$$
$$=\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}+\frac{\partial\ell^2(\theta_{(0)},\beta_{(0)})}{\partial\theta\,\partial\theta^\tau}(\tilde\theta-\theta_{(0)})$$
$$+\frac{\partial\ell^2(\theta_{(0)},\beta_{(0)})}{\partial\theta\,\partial\beta^\tau}(\tilde\beta-\beta_{(0)})+o_p(\delta_n),$$
$$0=\frac{\partial\ell(\tilde\theta,\tilde\beta)}{\partial\beta}$$
$$=\frac{\partial\ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}+\frac{\partial\ell^2(\theta_{(0)},\beta_{(0)})}{\partial\beta\,\partial\theta^\tau}(\tilde\theta-\theta_{(0)})$$
$$+\frac{\partial\ell^2(\theta_{(0)},\beta_{(0)})}{\partial\beta\,\partial\beta^\tau}(\tilde\beta-\beta_{(0)})+o_p(\delta_n),$$

where

$$\delta_n = \|\tilde\theta-\theta_{(0)}\| + \|\tilde\beta-\beta_{(0)}\| = o_p(1).$$

Thus,

$$nS_n\begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix}=\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}+o_p(\delta_n).$$

Because $S_n = S + o_p(1)$ by the weak law of large numbers and $\frac{1}{\sqrt{n}}\begin{pmatrix}\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}=O_p(1)$ by the central limit theorem, it follows that

$$\begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix}=\frac{1}{n}S^{-1}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}$$

$$+o_p(n^{-1/2})\frac{1}{\sqrt{n}}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}+o_p(n^{-1}\delta_n)$$

$$=\frac{1}{n}S^{-1}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}+o_p(n^{-1/2}),$$

thus establishing (9). To prove (10), it suffices to show that

$$\frac{1}{\sqrt{n}}S^{-1}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}\xrightarrow{d} N_{(p+1)I}\left(0,\Sigma\right).$$

Because each term in $\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}$ and $\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}$ has mean 0, it follows from the multivariate central limit theorem that

$$\frac{1}{\sqrt{n}}B^{-1/2}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}\xrightarrow{d} N_{(p+1)I}\left(0,I_{(p+1)I}\right),$$

where B is defined in Lemma 1. By Slutsky's Theorem and Lemma 1,

$$\frac{1}{\sqrt{n}}S^{-1}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}$$

$$=S^{-1}B^{1/2}\frac{1}{\sqrt{n}}B^{-1/2}\begin{pmatrix}\dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \dfrac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}$$

$$\xrightarrow{d} S^{-1}B^{1/2}N_{(p+1)I}\left(0,I_{(p+1)I}\right)\overset{d}{=}N_{(p+1)I}\left(0,\Sigma\right).$$

The proof is complete.

Proof of Theorem 2: Let

$$v_{11}^{ij}=-\frac{1}{1+\rho}\times$$

$$\int\frac{\rho_i\exp[\theta_{i0}+s_i(t;\beta_{i0})]\rho_j\exp[\theta_{j0}+s_j(t;\beta_{j0})]}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_m+s_m(t;\beta_{m0})]}\times$$

$$\psi_i(t;\theta_{(0)},\beta_{(0)})dG_0(t),\quad i\neq j=0,1,\cdots,I,$$

$$v_{11}^{ii}=-\sum_{j=0,\,j\neq i}^{I}v_{11}^{ij},$$

$$i=0,1,\cdots,I,\qquad V_{11}=(v_{11}^{ij})_{i,\,j=1,\cdots,I},$$

$$v_{12}^{ij}=-\frac{1}{1+\rho}\times$$

$$\int\frac{\rho_i\exp[\theta_{i0}+s_i(t;\beta_{i0})]\rho_j\exp[\theta_{j0}+s_j(t;\beta_{j0})]}{1+\sum_{m=1}^{I}\rho_m\exp[\theta_m+s_m(t;\beta_{m0})]}\times$$

$$\psi_i(t;\theta_{(0)},\beta_{(0)})d_j^{\tau}(t;\beta_{(j0)})dG_0(t),\quad i\neq j=0,1,\cdots,I,$$

$$v_{12}^{ii} = -\sum_{j=0,\, j\neq i}^{I} v_{12}^{ij}, \quad i = 0,1,\cdots,I,$$

$$V_{12} = (v_{12}^{ij})_{i,\, j=1,\cdots,I}, \quad V = (V_{11}, V_{12}),$$

$$L(\theta,\beta) = (L_1^{\tau}(\theta,\beta),\cdots,L_I^{\tau}(\theta,\beta))^{\tau},$$

$$B_{\psi} = \mathrm{Var}\left[\frac{1}{\sqrt{n}} L(\theta_{(0)},\beta_{(0)})\right]. \tag{18}$$

The proof of the asymptotic normality of $\begin{pmatrix}\bar{\theta}-\theta_{(0)}\\ \bar{\beta}-\beta_{(0)}\end{pmatrix}$ is similar to that of $\begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix}$ in Theorem 1 by noting that

$$\begin{pmatrix}\bar{\theta}-\theta_{(0)}\\ \bar{\beta}-\beta_{(0)}\end{pmatrix} = \frac{1}{n} V^{-1} L(\theta_{(0)},\beta_{(0)}) + o_p(n^{-1/2}).$$

To prove the second part of Theorem 2, notice that $\begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix}$ and $\begin{pmatrix}\bar{\theta}-\theta_{(0)}\\ \bar{\beta}-\beta_{(0)}\end{pmatrix} - \begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix}$ are asymptotically independent because it can be shown after very extensive algebra that

$$\mathrm{Cov}\left[S^{-1}\begin{pmatrix}\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}, V^{-1}L(\theta_{(0)},\beta_{(0)}) - S^{-1}\begin{pmatrix}\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix}\right] = 0.$$

Consequently, there is

$$0 \le \mathrm{Var}\left[\frac{1}{\sqrt{n}} S^{-1}\begin{pmatrix}\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix} - \frac{1}{\sqrt{n}} V^{-1}L(\theta_{(0)},\beta_{(0)})\right] = \Sigma_{\psi} - \Sigma.$$

This completes the proof of Theorem 2.

Proof of Theorem 3: Since $\mathrm{E}H_{0i}(t) = \rho_i^{-1} A_{1i}(t)$ and $\mathrm{E}H_{3i}(t) = \rho_i^{-1} A_{2i}(t)$ for $i = 0,1,\cdots,I$ and $(\tilde{\theta},\tilde{\beta})$ is strongly consistent, applying a first-order Taylor expansion and Theorem 1 gives, uniformly in t,

$$\begin{aligned}
\tilde{G}_i(t) &= \frac{1}{n_i}\sum_{k=1}^{n} \frac{\rho_i \exp[\tilde{\theta}_i + s_i(T_k;\tilde{\beta}_i)] I_{[T_k\le t]}}{1+\sum_{m=1}^{I}\rho_m \exp[\tilde{\theta}_m + s_m(T_k;\tilde{\beta}_m)]} \\
&= H_{1i}(t) - H_{0i}^{\tau}(t)(\tilde{\theta}-\theta_{(0)}) - H_{3i}^{\tau}(t)(\tilde{\beta}-\beta_{(0)}) + o_p(\delta_n) \\
&= H_{1i}(t) - (\mathrm{E}H_{0i}^{\tau}(t),\ \mathrm{E}H_{3i}^{\tau}(t))\begin{pmatrix}(\tilde{\theta}-\theta_{(0)})\\ (\tilde{\beta}-\beta_{(0)})\end{pmatrix} \\
&\quad - r_{in}(t) + o_p(\delta_n) \\
&= H_{1i}(t) - \frac{1}{n\rho_i}(A_{1i}^{\tau}(t),\ A_{2i}^{\tau}(t)) S^{-1}\begin{pmatrix}\frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\theta}\\ \frac{\partial \ell(\theta_{(0)},\beta_{(0)})}{\partial\beta}\end{pmatrix} \\
&\quad + o_p(n^{-1/2}) - r_{in}(t) + o_p(\delta_n) \\
&= H_{1i}(t) - H_{2i}(t) + R_{in}(t), \qquad i = 0,1,\cdots,I,
\end{aligned} \tag{19}$$

where $\delta_n = \|\tilde{\theta}-\theta_{(0)}\| + \|\tilde{\beta}-\beta_{(0)}\|$ and for $i = 0,1,\cdots,I$,

$$r_{in}(t) = (H_{0i}^{\tau}(t) - \mathrm{E}H_{0i}^{\tau}(t),\ H_{3i}^{\tau}(t) - \mathrm{E}H_{3i}^{\tau}(t))\begin{pmatrix}\tilde{\theta}-\theta_{(0)}\\ \tilde{\beta}-\beta_{(0)}\end{pmatrix},$$

$$R_{in}(t) = o_p(n^{-1/2}) - r_{in}(t) + o_p(\delta_n).$$

It follows from part (b) of Theorem 3 that $\delta_n = O_p(n^{-1/2})$. Furthermore, it can be shown that $\sup_{-\infty\le t\le\infty} |r_{in}(t)| = o_p(n^{-1/2})$. As a result, $\sup_{-\infty\le t\le\infty} |R_{in}(t)| = o_p(n^{-1/2})$, which along with (19) establishes (12) and (14). To prove (15), according to (12) and (14), it suffices to show that

$$\sqrt{n}\begin{pmatrix}H_{10}-\hat{G}_0-H_{20}\\ H_{11}-\hat{G}_1-H_{21}\\ \vdots\\ H_{1I}-\hat{G}_I-H_{2I}\end{pmatrix} \xrightarrow{D} \begin{pmatrix}W_0\\ W_1\\ \vdots\\ W_I\end{pmatrix} \text{ in } D^{I+1}[-\infty,\infty]. \tag{20}$$

Under the assumption that the underlying distribution function G_0 is continuous (20) is proven. According to (16) and Lemmas 2 and 3, we have for $-\infty \le s \le t \le \infty$,

$$\mathrm{E}\{\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)-H_{2i}(t)]\}=0=\mathrm{E}W_i(t),\quad i=0,1,\cdots,I,$$

$$\begin{aligned}&\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)]-\sqrt{n}H_{2i}(s),\ \sqrt{n}[H_{1i}(t)-\hat{G}_i(t)]-\sqrt{n}H_{2i}(t))\\&=\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\sqrt{n}[H_{1i}(t)-\hat{G}_i(t)])\\&\quad-\mathrm{Cov}(\sqrt{n}H_{2i}(s),\ \sqrt{n}H_{2i}(t))\\&=\frac{1+\rho}{\rho_i^2}[G_i(s)-B_{ii}(s)]\\&\quad-\frac{1+\rho}{\rho_i^2}\sum_{k=0,k\neq i}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{ik}(t)\\&\quad-\frac{1+\rho}{\rho_i^3}[G_i(s)-B_{ii}(s)][G_i(t)-B_{ii}(t)]\\&\qquad-\frac{1}{\rho_i^2}(A_{1i}^{\tau}(s),A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix}A_{1i}(t)\\A_{2i}(t)\end{pmatrix}\\&\qquad+\frac{1+\rho}{\rho_i^2}\sum_{k=0,k\neq i}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{ik}(t)\\&\qquad+\frac{1+\rho}{\rho_i^3}[G_i(s)-B_{ii}(s)][G_i(t)-B_{ii}(t)]\\&=\frac{1+\rho}{\rho_i^2}[G_i(s)-B_{ii}(s)]\\&\quad-\frac{1}{\rho_i^2}(A_{1i}^{\tau}(s),A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix}A_{1i}(t)\\A_{2i}(t)\end{pmatrix}=\mathrm{E}W_i(s)W_i(t),\\&\quad i=0,1,\cdots,I,\end{aligned}$$

$$\begin{aligned}&\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)]-\sqrt{n}H_{2i}(s),\ \sqrt{n}[H_{1j}(t)-\hat{G}_j(t)]-\sqrt{n}H_{2j}(t))\\&=\mathrm{Cov}(\sqrt{n}[H_{1i}(s)-\hat{G}_i(s)],\sqrt{n}[H_{1j}(t)-\hat{G}_j(t)])\\&\quad-\mathrm{Cov}(\sqrt{n}H_{2i}(s),\ \sqrt{n}H_{2j}(t))\\&=-\frac{1+\rho}{\rho_i\rho_j}B_{ij}(s)-\frac{1+\rho}{\rho_i\rho_j}\sum_{k=0}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{jk}(t)\\&\quad+\frac{1+\rho}{\rho_i\rho_j^2}B_{ij}(s)G_j(t)+\frac{1+\rho}{\rho_i^2\rho_j}G_i(s)B_{ij}(t)\\&\qquad-\frac{1}{\rho_i\rho_j}(A_{1i}^{\tau}(s),A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix}A_{1j}(t)\\A_{2j}(t)\end{pmatrix}\\&\qquad+\frac{1+\rho}{\rho_i\rho_j}\sum_{k=0}^{I}\frac{1}{\rho_k}B_{ik}(s)B_{jk}(t)\\&\quad-\frac{1+\rho}{\rho_i\rho_j^2}B_{ij}(s)G_j(t)-\frac{1+\rho}{\rho_i^2\rho_j}G_i(s)B_{ij}(t)\\&=-\frac{1+\rho}{\rho_i\rho_j}B_{ij}(s)-\frac{1}{\rho_i\rho_j}(A_{1i}^{\tau}(s),A_{2i}^{\tau}(s))S^{-1}\begin{pmatrix}A_{1j}(t)\\A_{2j}(t)\end{pmatrix}\\&=\mathrm{E}W_i(s)W_i(t),\qquad i\neq j=0,\ 1,\ \cdots,I.\end{aligned}$$

It then follows from the multivariate central limit theorem for sample means and the Cramer-Wold device that the finite-dimensional distributions of

$$\sqrt{n}(H_{10}-\hat{G}_0-H_{20},\cdots,H_{1I}-\hat{G}_I-H_{2I})^{\tau}$$

converge weakly to those of $(W_0,\cdots,W_I)^{\tau}$. Thus, in order to prove (20), it is enough to show that the process

$$\{\sqrt{n}(H_{10}(t)-\hat{G}_0(t)-H_{20}(t),\cdots,\ H_{1I}(t)-\hat{G}_I(t)-H_{2I}(t))^{\tau},\quad -\infty\le t\le\infty\}$$

is tight in $D^{I+1}[-\infty,\infty]$. But this has been established by Lemma 4 for continuous G_0. Thus, (20) has been proven when G_0 is continuous.

Suppose now that G_0 is an arbitrary distribution function over $[-\infty,\infty]$. Define the inverse of G_0, or quantile function associated with G_0, by $G_0^{-1}(x)=\inf\{t: G_0(t)\ge x\}$, $x\in(0,1)$. Let $\xi_{i1},\cdots,\xi_{in_i}$ be independent random variables having the same density function $h_i(x)=\exp[\theta_i+s_i(G_0^{-1}(x);\beta_i)]$ on $(0,1)$ for $i=0,1,\cdots,I$ and assume that

$\{(\xi_{i1},\cdots,\xi_{in_i}): i=0,1,\cdots,I\}$ are jointly independent. Thus, we have the following $(I+1)$-sample semiparametric model analogous to (2):

$$\xi_{01},\cdots,\xi_{0n_0} \overset{i.i.d.}{\sim} h_0(x)=I_{(0,1)}(x),$$

$$\xi_{i1},\cdots,\xi_{in_i} \overset{i.i.d.}{\sim} h_i(x)=\exp[\theta_i+s_i(G_0^{-1}(x);\beta_i)]h_0(x),$$
$$i=0,1,\cdots,I. \qquad (21)$$

Then, it is easy to see that $(X_{i1},\cdots,X_{in_i})$ and $(G_0^{-1}(\xi_{i1}),\cdots,G_0^{-1}(\xi_{in_i}))$ have the same distribution, i.e., $(X_{i1},\cdots,X_{in_i}) \overset{d}{=} (G_0^{-1}(\xi_{i1}),\cdots,G_0^{-1}(\xi_{in_i}))$ for $i=0,1,\cdots,I$. Let $\{\psi_1,\cdots,\psi_n\}$ denote the pooled random variables $\{\xi_{01},\cdots,\xi_{0n_0};\xi_{11},\cdots,\xi_{1n_1};\cdots;\xi_{I1},\cdots,\xi_{In_I}\}$, then $(T_1,\cdots,T_n) \overset{d}{=} (G_0^{-1}(\psi_1),\cdots,G_0^{-1}(\psi_n))$. For $u\in(0,1)$ and $m=0,1,\cdots,I$, let $\tilde{H}_{1m}(u),\tilde{H}_{2m}(u)$, and $\hat{\tilde{G}}_m(u)$ be the corresponding counterparts of $\tilde{H}_{1m}(t)$, $\tilde{H}_{2m}(t)$, and $\hat{\tilde{G}}_m(t)$ under model (21). Now define $\phi: D^{I+1}[-\infty,\infty]\to D^{I+1}[-\infty,\infty]$ by $(\phi K)(t)=K(G_0(t))$, then it can be shown that

$$\sqrt{n}(\tilde{H}_{10}[G_0(t)]-\hat{\tilde{G}}_0[G_0(t)]-\tilde{H}_{20}[G_0(t)],\cdots,$$
$$\tilde{H}_{1I}[G_0(t)]-\hat{\tilde{G}}_I[G_0(t)]-\tilde{H}_{2I}[G_0(t)])^\tau$$
$$\overset{d}{=}\sqrt{n}(H_{10}(t)-\hat{G}_0(t)-H_{20}(t),\cdots,H_{1I}(t)-\hat{G}_I(t)]$$
$$-H_{2I}(t))^\tau \quad \text{and}$$
$$\sqrt{n}(\tilde{H}_{10}-\hat{\tilde{G}}_0-\tilde{H}_{20},\cdots,\tilde{H}_{1I}-\hat{\tilde{G}}_I-\tilde{H}_{2I})^\tau$$
$$\overset{D}{\to}(\tilde{W}_0,\cdots\tilde{W}_I)^\tau$$

in $D^{I+1}[0,1]$, where $(\tilde{W}_0,\cdots\tilde{W}_I)^\tau$ is a multivariate Gaussian process satisfying $\phi[(\tilde{W}_0,\cdots\tilde{W}_I)^\tau] \overset{d}{=} (W_0,\cdots W_I)^\tau$. If K_n converges to K in the Skorohod topology and $K\in C^{I+1}[-\infty,\infty]$, then the convergence is uniform, so that ϕK_n converges to ϕK uniformly and hence in the Skorohod topology. As a result, Theorem 5.1 of Billingsley (1968, page 30) implies that

$$\sqrt{n}(H_{10}-\hat{G}_0-H_{20},\cdots,H_{1I}-\hat{G}_I-H_{2I})^\tau$$
$$\overset{d}{=}\phi[\sqrt{n}(\tilde{H}_{10}-\hat{\tilde{G}}_0-\tilde{H}_{20},\cdots,\tilde{H}_{1I}-\hat{\tilde{G}}_I-\tilde{H}_{2I})^\tau]$$
$$\overset{D}{\to}\phi[(\tilde{W}_0,\cdots\tilde{W}_I)^\tau] \overset{d}{=} (W_0,\cdots W_I)^\tau.$$

Therefore, (20) holds for general G_0, and this completes the proof of Theorem 3.

References

Agresti, A. (1990). *Categorical data analysis.* NY: John Wiley & Sons.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika, 59,* 19-35.

Anderson, J. A. (1979). Robust inference using logistic models. *International Statistical Institute Bulletin,* 48, 35-53.

Billingsley, P. (1968). *Convergence of probability measures.* NY: John Wiley & Sons.

Day, N. E., & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics, 23,* 313-323.

Farewell, V. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika, 66,* 27-32.

Gilbert, P., Lele, S., & Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika, 86,* 27-43.

Gill, R. D., Vardi, Y., & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics, 16,* 1069-1112.

Hall, P., & La Scala, B. (1990). Methodology and algorithms of empirical likelihood, *International Statistical Review,58,* 109-127.

Hsieh, D. A., Manski, C. F., & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, 80, 651-662.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A. B. (1990). Empirical likelihood confidence regions. *Annals of Statistics*, 18, 90-120.

Owen, A. B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19, 1725-1747.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, 21, 1182-96.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* 85, 619-30.

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300-325.

Qin, J., & Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, *84*, 609-618.

Scott, A. J., & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, *84*, 57-71.

Sen, P. K., & Singer, J. M. (1993). *Large sample methods in statistics: an introduction with applications*. NY: Chapman & Hall.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes with applications to statistics*. NY: Springer.

Vardi, Y. (1982). Nonparametric estimation in presence of length bias. *Annals of Statistics*, 10, 616-620.

Vardi, Y. (1985). Empirical distribution in selection bias models, *Annals of Statistics*, *13*, 178-203.

Wacholder, S., & Weinberg, C. R. (1994). Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics*, *50*, 350-357.

Weinberg, C. R., & Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics*, *46*, 963-975.

Weinberg, C. R., & Sandler, D. P. (1991). Randomized recruitment in case-control studies. *American Journal of Epidemiology*, *134*, 421-433.

Weinberg, C. R., & Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika*, 80, 461-465.

Zhang, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica*, *10*, 839-865.

Zhang, B. (2002). Assessing goodness-of-fit of generalized logit models based on case-control data. *Journal of Multivariate Analysis*, *82*, 17-38.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 35-42

1538 – 9472/05/$95.00

Two Sides Of The Same Coin: Bootstrapping The Restricted Vs. Unrestricted Model

Panagiotis Mantalos
Department of Statistics
Lund University, Sweden

The properties of the bootstrap test for restrictions are studied in two versions: 1) bootstrapping under the null hypothesis, restricted, and 2) bootstrapping under the alternative hypothesis, unrestricted. This article demonstrates the equivalence of these two methods, and illustrates the small sample properties of the Wald test for testing Granger-Causality in a stable stationary VAR system by Monte Carlo methods. The analysis regarding the size of the test reveals that, as expected, both bootstrap tests have actual sizes that lie close to the nominal size. Regarding the power of the test, the Wald and bootstrap tests share the same power as the use of the Size-Power Curves on a correct size-adjusted basis.

Key words: Bootstrap, Granger-Causality, VAR system, Wald test

Introduction

When studying the small sample properties of a test procedure by comparing different tests, two aspects are of importance:

a) to find the test that has actual size closest to the nominal size, and given that (a) holds, and
b) to find the test that has the greatest power.

In most cases, however, the distributions of the test statistic used are known only asymptotically and, unfortunately, unless the sample size is very large, the tests may not have the correct size. Inferential comparisons and judgements based on them might be misleading. Gregory and Veall (1985) can be consulted for an illustrative example.

One of the ways to deal with this situation is to use the bootstrap. The use of this procedure is increasing with the advent of personal computers.

Panagiotis Mantalos, Department of Statistics Lund University, Box 743 SE-22007 Lund, Sweden. E-mail: Panagiotis.Mantalos@stat.lu.se

However, the issue of the bootstrap test, even it is applied, is not trivial. One of the problems is that one needs to decide how to resample the data, and whether to resample under the null hypothesis or under the alternative hypothesis.

By bootstrapping under the null hypothesis, an approximation is made of the distribution of the test statistic, thereby generating more robust critical values for our test statistic. Alternately, by bootstrapping under the alternative hypothesis, an approximation is made of the distribution of the parameter, and is subsequently used to make inferences.

In either case, it does not matter whether the nature of the theoretical distribution of the parameter estimator or the theoretical distribution of the test statistic is known. What matters is that the bootstrap technique approximates those distributions.

In this article, the bootstrap test procedure shows that

a) by bootstrapping under the null hypothesis (that is, bootstrapping the restricted model), and
b) by bootstrapping under the alternative hypothesis (that is, bootstrapping the unrestricted model)

will lead to the same results.

The properties of the two different methods will be illustrated and investigated using Monte Carlo methods. The Residual Bootstrap, (RB), will be used to study the properties of the test procedure when the errors are identically and independently distributed. To provide an example that is easy to be extended to a more general hypothesis, it is convenient to use the Wald test for restrictions for testing Granger-causality in a stable stationary VAR system.

The Model

Consider the general linear model

$$y = \mathbf{X}\beta + \delta \qquad (1)$$

where y is an $(n \times 1)$ vector, $\mathbf{X}$ is an $(n \times K)$ matrix and b is a $(K \times 1)$ vector. It is assumed that δ is an n-dimensional normal vector $\mathbb{N}(0, \Omega)$.

Consider testing q independent linear restrictions:

$$H_o : \mathrm{R}\beta = \mathrm{r} \quad \text{vs.} \quad H_1 : \mathrm{R}\beta \neq \mathrm{r}, \qquad (2)$$

where q and r are fixed (q x 1) vectors and R is a fixed $(q \times K)$ matrix with rank q. It is possible to base a test of H_0 on the Wald criterion

$$T_s = (R\hat{\beta} - r)' \left[Var(R\hat{\beta}) \right]^{-1} (R\hat{\beta} - r). \qquad (3)$$

Bootstrap critical values

The bootstrap technique improves the critical values, so that the true size of the test approaches its nominal value. The principle of bootstrap critical values is to draw a number of Bootstrap samples from the model under the null hypothesis, calculate the Bootstrap test statistic T_s^*, and compare it with the observed test statistic.

The bootstrap procedure for calculation of the critical values is given by the following steps:

a) Estimate the test statistic as in (3)

b) Use the adjusted OLS residuals $(\hat{\delta}_i - \overline{\delta})$ i = 1,...,T to draw i.i.d $\delta_1^*, \ldots, \delta_T^*$ data. Define

$$y^* = X\hat{\beta}_0 + \delta^*. \qquad (4)$$

c) Then, calculate the test statistic T_s^* as in (3), i.e., by applying the Wald test procedure to the (4) model. Repeat this step N_b times and take the $(1-\alpha)^{th}$ quintile of the bootstrap distribution of T_s^* to obtain the α - level Bootstrap critical values ($c_{t\alpha}^*$). Reject H_o if $T_s \geq c_{t\alpha}^*$.

Among articles that advocated this approach are Horowitz (1994) and Mantalos and Shukur (1998), whereas Davidson and MacKinnon (1999) and Mantalos (1998) advocated the estimate of the P-value. A bootstrap estimate of the P-value for testing is $P^*\{T_s^* \geq T_s\}$.

Bootstrap-hypothesis testing

One of the important considerations for generating the y_t^* leading to the bootstrap critical values is whether to impose the null hypothesis on the model from which is generated the y_t^*. However, some authors, including Jeong and Chung (2001), argued for bootstrapping under the alternative hypothesis. 'Let the data speak' is their principle in apply the bootstrap. The bootstrap procedure to resample the data from the unrestricted model consists of the following steps:

a) Estimate the test statistic as in (3)

b) Use the adjusted OLS residuals $(\hat{\delta}_i - \overline{\delta})$ i = 1,...,T. to draw i.i.d $\delta_1^*, \ldots, \delta_T^*$ data. Define $y^* = X\hat{\beta} + \delta^*$, noting that $\hat{\beta}$ is the unconstrained LS estimator of β. That is, the unrestricted model is used to simulate the y^*.

c) Calculate

$$T_s^* = \frac{\left(R(\hat{\beta}^* - \hat{\beta}) \right)^2}{Var(R\hat{\beta}^*)}. \qquad (5)$$

By repeating this step N_b times the $(1-\alpha)^{th}$ quintile can be used of the bootstrap distribution of the (5) as the α - level Bootstrap critical values ($c^*_{t\alpha}$). Reject H_o if $T_s \geq c^*_{t\alpha}$. The bootstrap estimate of the P-value is $P^*\{T^*_s \geq T_s\}$.

Since Efron's (1979) introduction of the bootstrap as a computer-based method for evaluating the accuracy of a statistic, there have been significant theoretical refinements of the technique. Horowitz (1994) and Hall and Horowitz (1996) discussed the method and showed that bootstrap tests are more reliable than asymptotic tests, and can be used to improve finite-sample performance. They provided a heuristic explanation of why the bootstrap provides asymptotic refinements to the critical values of test statistics. See Hall (1992) for a wider discussion on bootstrap refinements based on the Edgeworth expansion.

Davidson and MacKinnon (1999) provided an explanation of why the bootstrap provides asymptotic refinements to the p- values of a test. The same authors conclude that by using the bootstrap critical values or bootstrap test, the size distortion of a bootstrap test is at least of order $T^{-1/2}$ smaller than that of the corresponding asymptotic test.

Two sides of the same coin

Consider the general linear model

$$y = \mathbf{X}\beta + \delta \quad (1)$$

and suppose that the interest is in testing the q independent linear restrictions

$$H_o: \mathrm{R}\beta = \mathrm{r} \quad \text{vs.} \quad H_1: \mathrm{R}\beta \neq \mathrm{r}. \quad (2)$$

Let the LS unconstrained estimator of β be denoted by $\hat{\beta}$ and the equality-constrained estimator be denoted by $\hat{\beta}_0$. The bootstrap GDPs are:

a) Restricted: $y^*_R = \mathbf{X}\hat{\beta}_0 + \delta^*$. (6)

b) Unrestricted: $y^*_u = \mathbf{X}\hat{\beta} + \delta^*$. (7)

Let $\hat{\beta}^*$ be the LS estimator of the *b* coefficient in the model relating y^*_R to **X,** and $\hat{\hat{\beta}}^*$ be the LS estimator of the *b* in the y^*_u on **X** model. Thus,

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y^*_R = \hat{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta^* \quad (8)$$

and

$$\hat{\hat{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y^*_u = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta^*. \quad (9)$$

From (8) and (9)

$$\hat{\beta}^* - \hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta^* \quad (10)$$

and

$$\hat{\hat{\beta}}^* - \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta^*. \quad (11)$$

Because the right-hand components of the (10) and (11) are equal,

$$\left(\hat{\beta}^* - \hat{\beta}_0\right) = \left(\hat{\hat{\beta}}^* - \hat{\beta}\right). \quad (12)$$

It is not difficult to see from (12) that the same results from the both methods are expected: there are two sides to the same coin. These results will be illustrated by a Monte Carlo experiment.

Wald test for restrictions in a VAR model

Consider a data-generation process (DGP) that consists of the k-dimensional multiple time series generated by the VAR(p) process

$$y_t = A_1 y_{t-1} + \ldots + \mathrm{A_p} y_{t-p} + \varepsilon_t, \quad (13)$$

where $\varepsilon_t = \left(\varepsilon_{1t}, \ldots, \varepsilon_{kt}\right)'$ is a zero mean independent white noise process with non-singular covariance matrix Σ_ε and, for j = 1, ... ,

k, $\mathrm{E}\left|\varepsilon_{jt}\right|^{2+\tau} < \infty$ for some $\tau > 0$. The order p of the process is assumed to be known. Define

$$\mathbf{Y}: = \left(y_1, \cdots, y_T\right) \quad \left(\text{k x T}\right) \text{ matrix,}$$

$$\mathbf{B}: = \left(v, A_1, \cdots, A_p\right) \quad \text{(k x (kp+ 1)) matrix,}$$

$$\mathbf{Z}_t: = \begin{bmatrix} 1 \\ y_t \\ \vdots \\ y_{t-p+1} \end{bmatrix} \quad \left(\text{(kp +1) x 1}\right) \text{ matrix,}$$

$$\mathbf{Z}: = \left(\mathbf{Z_0}, \cdots, \mathbf{Z_{T-1}}\right) \quad \left(\text{(kp+1) x T}\right) \text{ matrix}$$

and

$$\delta: = \left(\varepsilon_1, \cdots, \varepsilon_T\right) \quad \left(\text{k x T}\right) \text{ matrix.}$$

By using this notation, for t = 1, …, T, the VAR (p) model including a constant term (v) can be written compactly as

$$\mathbf{Y} = \mathbf{BZ} + \delta. \tag{14}$$

Then, the LS estimator of the $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \mathbf{YZ}'\left(\mathbf{ZZ}'\right)^{-1}. \tag{15}$$

Let $\alpha_p = vec\left[A_1, \cdots, A_p\right]$ be the vector of the true parameters, and $\hat{\alpha}_p = \mathrm{vec}\left[\hat{\mathrm{A}}_1, \cdots, \hat{\mathrm{A}}_\mathrm{p}\right]$ be the vector the LS estimators of the parameters, where vec[.] denotes the vectorization operator that stacks the columns of the argument matrix. Then,

$$T^{1/2}\left(\hat{\alpha}_p - \alpha_p\right) \Rightarrow N\left(0, \Sigma_p\right), \tag{16}$$

where $\Rightarrow$ denotes weak convergence in distribution and the $[k^2(p)$ x $k^2(p)]$ covariance matrix Σ_p is non-singular.

Now, suppose that in testing q independent linear restrictions is of interest

$$H_o: \mathrm{R}\alpha_p = \mathrm{s} \quad \text{vs.} \quad H_1: \mathrm{R}\alpha_p \neq \mathrm{s}, \tag{17}$$

where q and s are fixed (q x 1) vectors and R is a fixed [q x $k^2 p$] matrix with rank q.

We can base a test of H_0 on the Wald criterion

$$T_{s\_wald} = (R\hat{\alpha}_p - s)'\left[Var(R\hat{\alpha}_p - s)\right]^{-1}(R\hat{\alpha}_p - s). \tag{18}$$

Let

$$\hat{\Sigma}_\delta = \frac{1}{T - kp - 1}\left(\mathbf{YY'} - \mathbf{YZ'(ZZ')^{-1}ZY'}\right) \tag{19}$$

be the estimate of the residual covariance matrix.

Then, the diagonal elements of $\left(ZZ'\right)^{-1} \otimes \hat{\Sigma}_\delta$ form the variance vector of the LS estimated parameters. Substitute (19) into (18) in order to have

$$T_{s\_wald} = (R\hat{\alpha}_p - s)'\left[R\left(\left(ZZ'\right)^{-1} \otimes \hat{\Sigma}_\delta\right)R'\right]^{-1}(R\hat{\alpha}_p - s). \tag{20}$$

The null hypothesis of no Granger-causality may be expressed in terms of the coefficients of VAR process as

$$H_o: \mathrm{R}\alpha_p = 0 \quad \text{vs.} \quad H_1: \mathrm{R}\alpha_p \neq 0. \tag{21}$$

Then, (20) can be written as

$$T_{s\_wald} = (R\hat{\alpha}_p)'\left[R\left(\left(ZZ'\right)^{-1} \otimes \hat{\Sigma}_\delta\right)R'\right]^{-1}(R\hat{\alpha}_p) \tag{22}$$

and the bootstrap variations as

$$T^*_{s\_wald} = (R\hat{\alpha}^*_p)'\left[R\left(\left(Z^*Z^{*'}\right)^{-1} \otimes \hat{\Sigma}^*_\delta\right)R'\right]^{-1}(R\hat{\alpha}^*_p) \tag{23}$$

for the restricted form and

$$T^{*}_{s\_wald} = (R\hat{\alpha}^{*}_{p} - R\hat{\alpha}_{p})' \left[R\left(\left(Z^{*}Z^{*'} \right)^{-1} \otimes \hat{\Sigma}^{*}_{\delta} \right) R' \right]^{-1} (R\hat{\alpha}^{*}_{p} - R\hat{\alpha}_{p}) \quad (24)$$

for the unrestricted form.

Methodology

Monte Carlo experiment

This section illustrates various generalizations of the Granger-causality tests in VAR systems with stationary variables, using Monte Carlo methods. The estimated size is calculated by observing how many times the null is rejected in repeated samples under conditions where the null is true.

The following VAR(1) process is generated:

$$y_t = \begin{bmatrix} 0.5 & 0.3 \\ T^{-1/2}\gamma & 0.5 \end{bmatrix} y_{t-1} + \varepsilon_t, \quad (25)$$

where $\varepsilon_t \sim N(0, I_2)$, $y_t = \left(y_{1t}.y_{2t} \right)'$. If $\gamma = 0$, y_{1t} is Granger-noncausal for y_{2t} and if $\gamma \neq 0$, y_{1t} causes y_{2t} . Therefore, $\gamma = 0$ is used to study the size of the tests.

The order p of the process is assumed to be known. Because this assumption might be too optimistic, a VAR(2) is fitted: $y_t = v + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$.

For each time series, 20 pre-sample values were generated with zero initial conditions, taking net sample sizes of T = 25 and 50. The Bootstrap test statistic (T^{*}_{s}) is calculated. As for N_b, which is the size of the bootstrap sample used to estimate bootstrap critical values and the P-value, $N_b = 399$ is used. Note that there are no initial bootstrap observations in bootstrap procedure.

Next presented are the results of the Monte Carlo experiment concerning the sizes of the various versions of the tests statistics using the VAR(2) model. Graphical methods are used that were developed and illustrated by Davidson and MacKinnon (1998) because they are easy to interpret. The P-value plot is used to study the size, and the Size-Power curves is used to study the power of the tests. The graphs, the P-value plots and Size-Power curves are based on the empirical distribution function, the EDF of the P-values, denoted as $\hat{F}\left(x_j \right)$.

For the P-value plots, if the distribution used to compute the p_s terms is correct, each of the p_s terms should be distributed uniformly on (0,1). Therefore the resulting graph should be close to the 45^{o} line.

Furthermore, to judge the reasonableness of the results, a 95% confidence interval is used for the nominal size (π_0) as: $\pi_0 \pm 2\sqrt{\frac{\pi_0(1-\pi_0)}{N}}$, where N is the number of Monte Carlo replications. Results that lie between these bounds will be considered satisfactory. For example, if the nominal size is 5%, define a result as reasonable if the estimated size lies between 3.6% and 6.4%. The P-value plots also make it possible and easy to distinguish between tests that systematically over-reject or under-reject, and those that reject the null hypothesis about the right proportion of the time.

Figure 1 shows the truncated P-value plots for the actual size of the bootstrap and the Wald tests, using 25 and 50 observations. Looking at these curves, it is not difficult to make the inference that both the bootstrap tests perform adequately, as they lie inside the confidence bounds. However, using the asymptotic critical values, the Wald test shows a tendency to over-reject the null hypothesis.

The superiority of the bootstrap test over the Wald test, concerning the size of the tests, is considerable, and more noticeable in small samples of size 25. The power of the Wald and bootstrap tests by using sample sizes of 25 and 50 observations was examined. The power function is estimated by calculating the rejection frequencies in 1000 replications using the value $\gamma = 2$.

The Size-Power Curves are used to compare the estimated power functions of the alternative test statistics. This proved to be quite

adequate, because those tests that gave reasonable results regarding size usually differed very little regarding power.

The same processes are followed for the size investigation to evaluate the EDFs denoted by $\hat{F}^{\oplus}\left(x_j\right)$, by using the same sequence of random numbers used to estimate the size of the tests. Size-Power Curves are used to plot the estimated power functions against the nominal size. The estimated power functions are plotted against the true size, that is, plotting $\hat{F}^{\oplus}\left(x_j\right)$ against $\hat{F}\left(x_j\right)$, which produces the Size-Power Curves on a correct size-adjusted basis.

Figure 2 shows the results of using the Size-Power Curves. The Wald test has higher power than the restricted and unrestricted bootstrap tests. A sample effect can also be seen. The larger the sample, the larger is the power of the tests. As the sample size increases, the power difference decreases.

However, the most interesting result is that both the restricted and unrestricted bootstrap tests share the same power. This result confirms the view that these two bootstrap methods are two sides of the same coin.

When using the Size-Power Curves on a correct size-adjusted basis, however, the situation is different concerning the power of the Wald and the bootstrap tests. Now the Wald, restricted and unrestricted bootstrap tests share the same power, as seen in Figure 3.

Conclusion

The purpose of this study was to provide advice on whether to resample under the null hypothesis or under the alternative hypothesis. In summary:

a) the restricted bootstrap test was used, in which the distribution of the test statistic was approximated, generating more robust critical values for our test statistic, and
b) the unrestricted bootstrap test, where the distribution of the parameter (coefficient) was approximated.

In both cases it does not matter whether or not the nature of the theoretical distribution of the parameter estimator or the theoretical distribution of the test statistic is known. What matters is that the bootstrap technique well approximates those distributions. Moreover, this article demonstrated the equivalence of these two methods.

The conclusion to this investigation for the Granger-causality test is that both bootstrap tests have an actual size that lies close to the nominal size. Given that the both unrestricted and restricted models have the same power, it makes sense to choose the bootstrap ahead of the classical tests, especially in small samples.

Figure 1. P-values Plots Estimated Size of the Wald and Bootstrap Tests.

Figure 1a: 25 observations

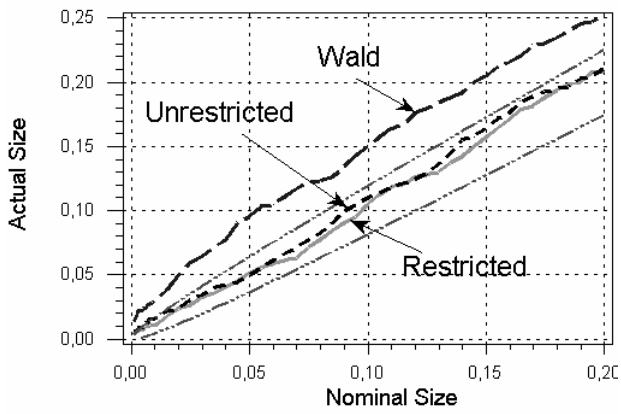
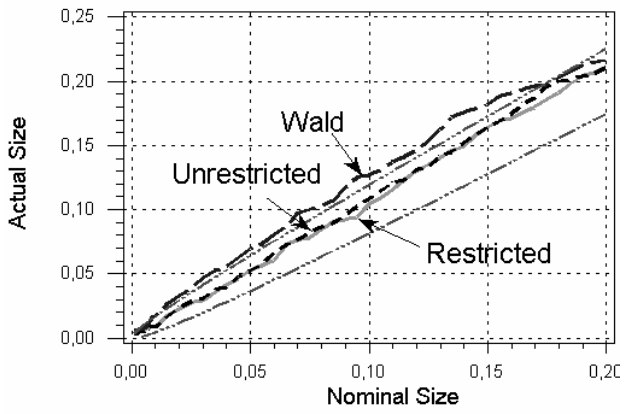


Figure 1b: 50 observations



Dash 3Dot lines: 95% Confidence interval

Figure 2. Estimated Power of the Wald and Bootstrap Tests.

Figure 2a: 25 observations

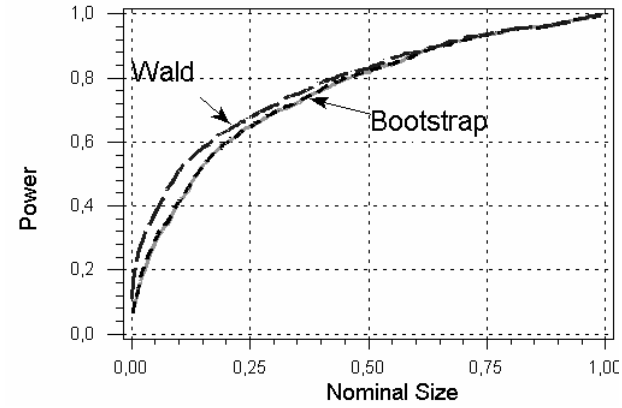


Figure 2b: 50 observations

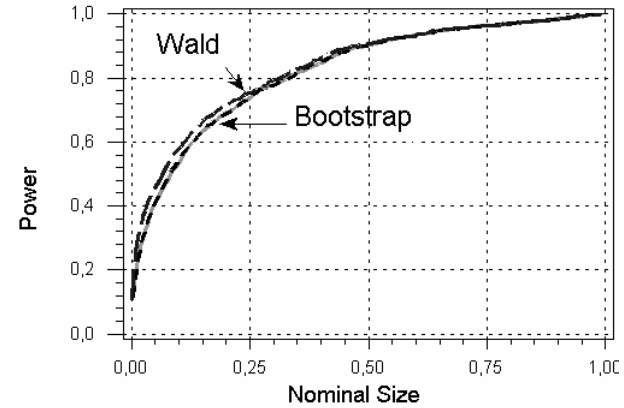


Figure3. Size-adjusted Power of the Wald and Bootstrap Tests.

Figure 3a: 25 observations

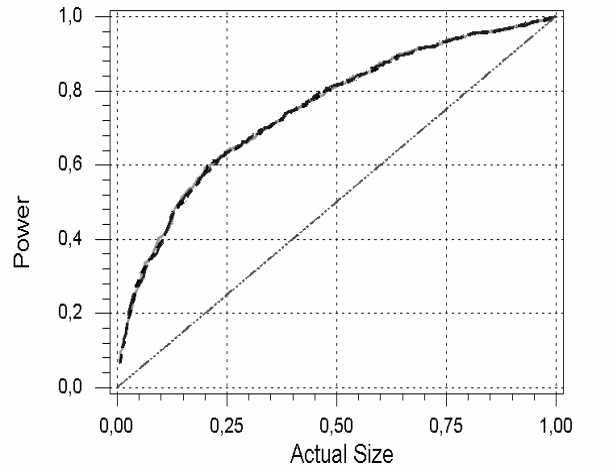
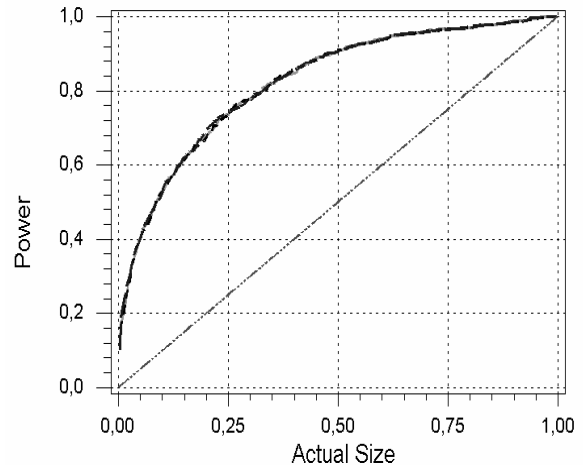


Figure 3b: 50 observations



References

Davidson, R., & MacKinnon, J. G. (1998). The size distortion of bootstrap tests. *Econometric Theory, 15*, 361-376.

Davidson, R., & MacKinnon, J. G. (1996). The Power of bootstrap tests. Discussion paper, *Queen's University, Kingston*, Ontario.

Davidson, R., & MacKinnon, J. G. (1998). Graphical methods for investigating the size and power of test statistics. *The Manchester School*, *66*, 1-26.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, 1-26.

Gregory, A. W., & Veall, M. R. (1985). Formulating Wald tests of nonlinear restrictions. *Econometrica 53*, 1465-1468.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York: Springer-Verlag.

Hall, P., & Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized - method - of - moments estimators. *Econometrica, 64*, 891–916.

Horowitz, J. L. (1994). Bootstrap-based critical values for the information matrix test. *Journal of Econometrics, 61*, 395-411.

Jeong, J., & Chung, S. (2001). Bootstrap test for autocorrelation. *Computational Statistics and Data Analysis*, *38*(1) 49-69.

MacKinnon, J. G. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature, 30*, 102-146.

Maddala, G. S. (1992). *Introduction to Econometrics,* (2ed,). New York: Maxwell Macmillan.

Mantalos**,** P. (1998). A graphical investigation of the size and power of the Granger-causality tests in Integrated-Cointegrated VAR Systems. *Studies in Nonlinear Dynamics and Econometrics* 4.1, 17-33

Mantalos, P., & Shukur. G. (1998). Size and power of the error correction model (ECM) cointegration test - A Bootstrap approach. *Oxford Bulletin of Economics and Statistics, 60*, 249-255.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 43-52

1538 – 9472/05/$95.00

Coverage Properties Of Optimized Confidence Intervals For Proportions

John P. Wendell Sharon P. Cox
College of Business Administration
University of Hawai`i at Mānoa

Wardell (1997) provided a method for constructing confidence intervals on a proportion that modifies the Clopper-Pearson (1934) interval by allowing for the upper and lower binomial tail probabilities to be set in a way that minimizes the interval width. This article investigates the coverage properties of these optimized intervals. It is found that the optimized intervals fail to provide coverage at or above the nominal rate over some portions of the binomial parameter space but may be useful as an approximate method.

Key words: Attribute, Bernoulli, dichotomous, exact, sampling

Introduction

A common task in statistics is to form a confidence interval on the binomial proportion *p*. The binomial probability distribution function is defined as

$$\Pr[Y = y \mid p, n] = \mathrm{b}(p, n, y) = \binom{n}{y} p^{y}\left(1 - p^{n-y}\right),$$

where the proportion of elements with a specified characteristic in the population is *p*, the sample size is *n*, and *y* is the outcome of the random variable *Y* representing the number of elements with a specified characteristic in the sample.

The coverage probability for a given value of *p* is

$$C_{n,CL^*}(p) = \sum_{i=0}^{n} \mathrm{I}(i, p)\mathrm{b}(p, n, i),$$

where $C_{n,CL^*}(p)$ is the coverage probability for a particular method with a nominal confidence level CL^* for samples of size *n* taken from a population with binomial parameter *p* and $\mathrm{I}(i, p)$ is 1 if the interval contains *p* when $y = i$ and 0 otherwise. The actual confidence level of a method for a given CL^* and *n* $\left(CL_{n,CL^*}\right)$ is the infimum over *p* of $C_{n,CL^*}(p)$. Exact confidence interval methods (Blyth & Still, 1983) have the property that $CL_{n,CL^*} \geq CL^*$ for all *n*, and CL^*.

The most commonly used exact method is due to Clopper and Pearson (1934) and is based on inverting binomial tests of $H_0 : p = p_0$. The upper bound of the Clopper-Pearson interval (*U*) is the solution in p_0 to the equation

$$\sum_{i=y}^{n} \mathrm{b}(\mathrm{p}_0, n, i) = \alpha_U,$$

except that when $y = n$, $U = 1$. The lower bound, *L*, is the solution in p_0 to the equation

$$\sum_{i=0}^{y} \mathrm{b}(\mathrm{p}_0, n, i) = \alpha_L,$$

except that when $y = 0$, $L = 0$. The nominal confidence level $CL^* = 1 - \alpha$ where

John P. Wendell is Professor, College of Business Administration, University of Hawai`i at Mānoa. E-mail: cbaajwe@hawaii.edu. Sharon Cox is Assistant Professor, College of Business Administration, University of Hawai`i at Mānoa.

$\alpha = \alpha_U + \alpha_L$. Because the Clopper-Pearson bounds are determined by inverting hypothesis tests, both α_U and α_L are set *a priori* and remain fixed regardless of the value of *y*. In practice, the values of α_U and α_L are often set to $\alpha_U = \alpha_L = \alpha/2$.

Wardell (1997) modified the Clopper-Pearson bounds by replacing the condition that α_U and α_L are fixed with the condition that only α is fixed. This allows α to be partitioned differently between α_U and α_L for each sample outcome *y*. Wardell (1997) provided an algorithm for accomplishing this partitioning in such a way that the confidence interval width is minimized for each *y*. Intervals calculated in this way are referred to here as optimized intervals. Wardell (1997) was concerned with determining the optimized intervals and not the coverage properties of the method. The purpose of this article is to investigate the coverage properties.

Coverage Properties of Optimized Intervals

Figure 1 plots $C_{n,.95}(p)$ against *p* for sample sizes of 5, 10, 20, and 50. The discontinuity evident in the Figure 1 plots is due to the abrupt change in the coverage probability when *p* is at *U* or *L* for any of the $n+1$ confidence intervals. Berger and Coutant (2001) demonstrated that the optimized interval method is an approximate and not an exact method by showing that $CL_{5,.95} = .9375 < .95$. Figure 1 confirms the Berger and Coutant result and extends it to sample sizes of 10, 20, and 50.

Agresti and Coull (1998) argued that some approximate methods have advantages over exact methods that make them preferable in many applications. In particular, they recommended two approximate methods for use by practitioners: the score method and adjusted Wald method. The interval bounds for the score method are

$$\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\left[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n \right]/n} \right) / \left(1 + z^2_{\alpha/2}/n\right),$$

where $\hat{p} = y/n$ and z_c is the $1-c$ quantile of the standard normal distribution. The adjusted Wald method interval bounds are

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/(n+4)},$$

where

$$\tilde{p} = (y+2)/(n+4).$$

One measure of the usefulness of an approximate method is the average coverage probability over the parameter space when *p* has a uniform distribution. This measure is used by Agresti and Coull (1998). Ideally, the average coverage probability should equal the nominal coverage probability. Figure 2 is a plot of the average coverage probabilities for the optimized interval, adjusted Wald and score methods for sample sizes of 1 to 100 and nominal confidence levels of .80, .90, 95, and .99.

Both the adjusted Wald and the score method perform better on this measure than the optimized interval method in the sense the average coverage probability is closer to the nominal across all of the nominal confidence levels and sample sizes. However, the optimized interval method has the desirable property that the average coverage probability never falls below the nominal for any of the points plotted. The score method is below the nominal for the entire range of sample sizes at the nominal confidence level of .99 and the same is true for the adjusted Wald method at the nominal confidence level of .80.

Figure 1. Coverage Probabilities of Optimized Intervals Across Binomial Parameter p. The disjointed lines plot the actual coverage probabilities of the optimized interval method across the entire range of values of p at a nominal confidence level of .95 for sample sizes of 5, 10, 20, and 50. The discontinuities occur at the boundary points of the n + 1 confidence intervals. The horizontal dotted line is at the nominal confidence level of .95. For all four sample sizes the actual coverage probability falls below the nominal for some values of p, demonstrating that the optimized bounds method is not an exact method.

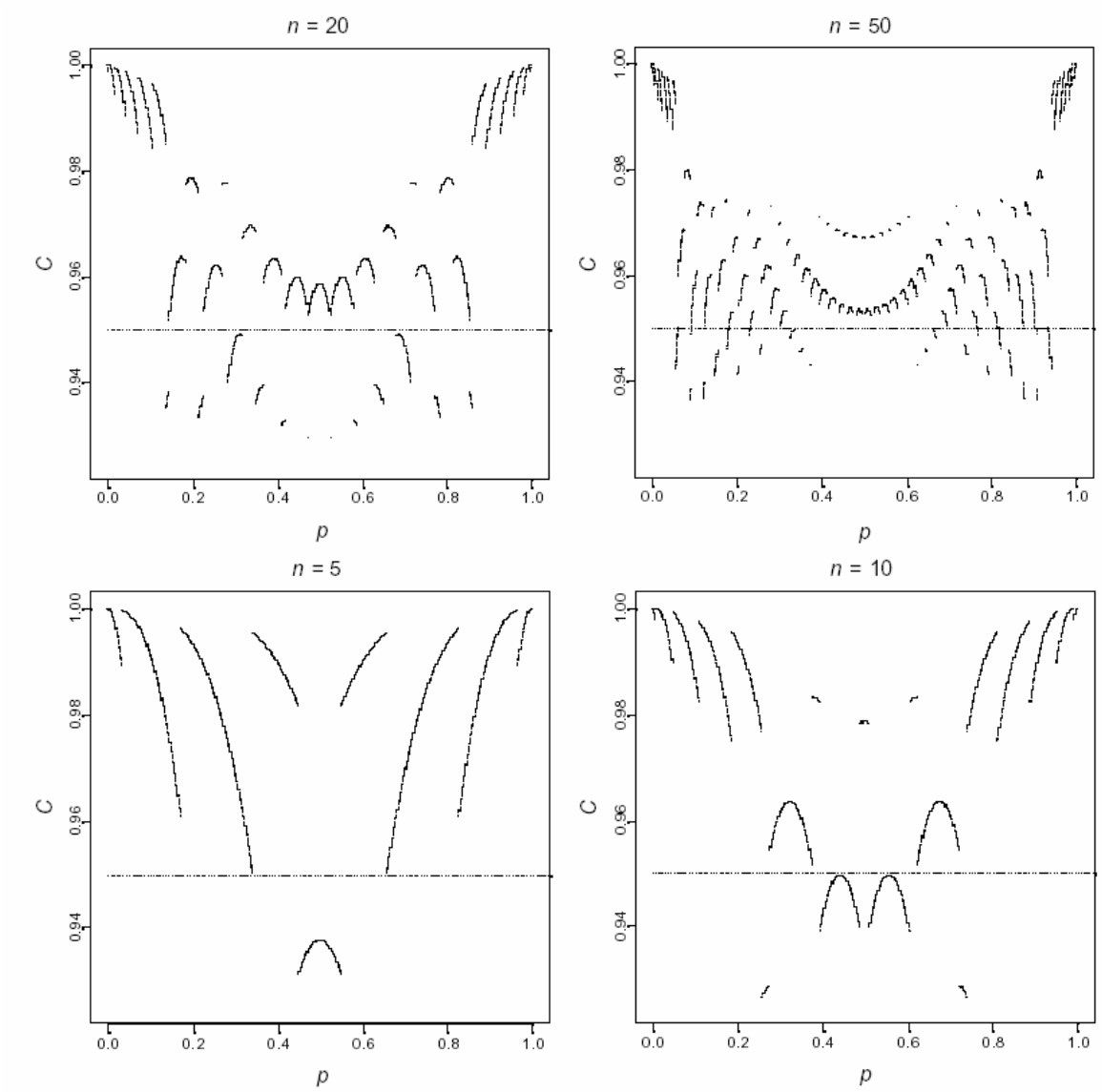
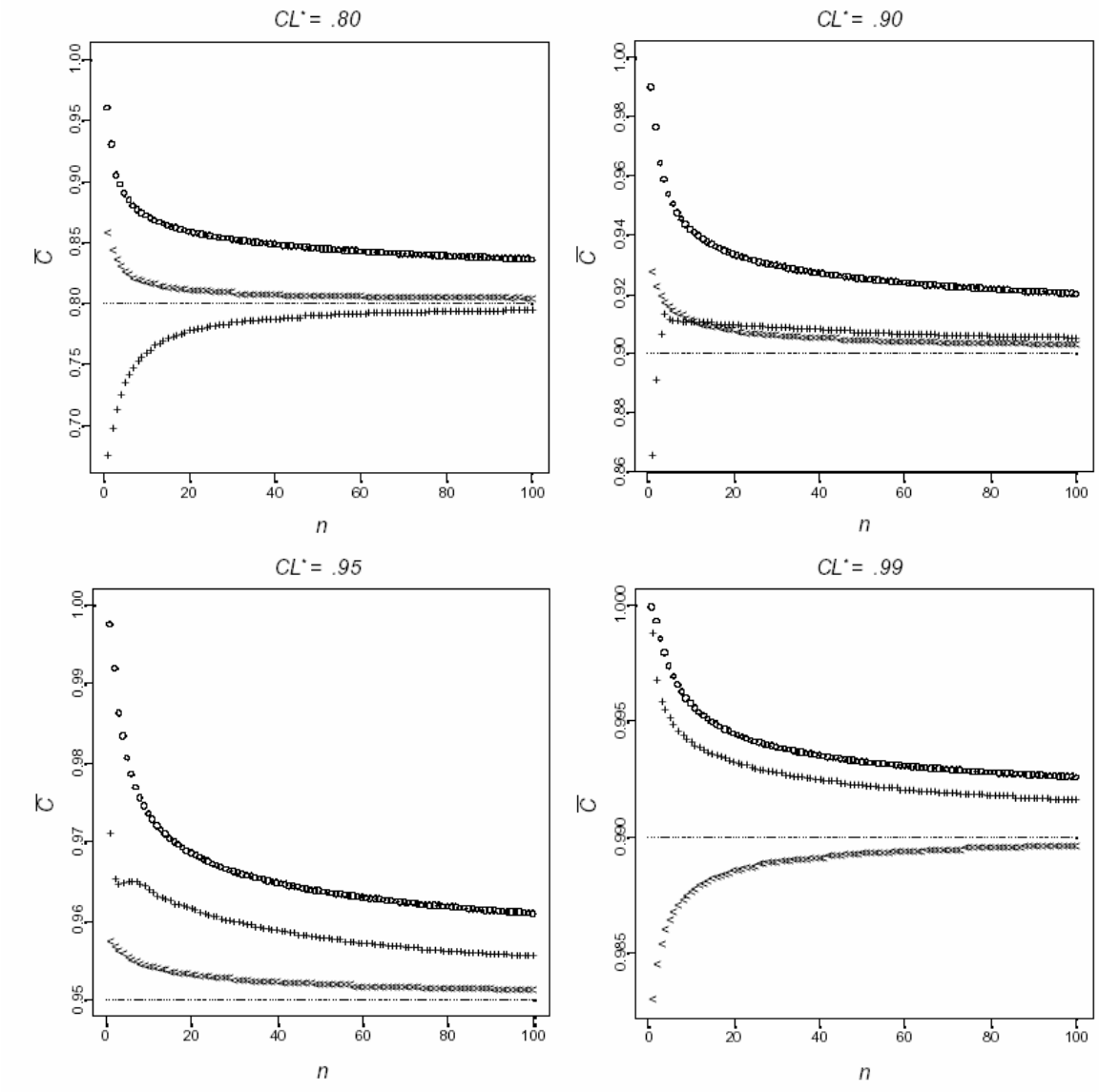


Figure 2. Average Coverage Probabilities of Three Approximate Methods. The scatter is of the average coverage probabilities of three approximate methods when p is uniformly distributed for sample sizes of from 1 to 100 with nominal confidence levels of .80, .90, .95, and .99. The optimized interval method is indicated by a "o", the adjusted Wald method by a "+", and the score method by a "<". The horizontal dotted line is at the nominal confidence level. The optimized interval method's average coverage probability tends to be further away from the nominal than the other two methods for all four nominal confidence levels and is always higher than the nominal. The average coverage probabilities of the other two methods tend to be closer to, and sometimes below, the nominal level.



A second measure used by Agresti and Coull (1998) is $\sqrt{\int_0^1 \left(C_{n,CL^*}(p) - CL^*\right)^2 dp}$, the uniform-weighted root mean squared error of the average coverage probabilities about the nominal confidence level. Ideally, this mean squared error would equal zero. Figure 3 plots the root mean squared error for the three methods over the same range of sample sizes and nominal confidence levels as Figure 2. The relative performance of the three methods for this metric varies according to the nominal confidence level. Each method has at least one nominal confidence level where the root mean squared error is furthest from zero for most of the sample sizes. The score method is worst at nominal confidence level of .99, the adjusted Wald at .80, and the optimized interval method at both .90 and .95.

Agresti and Coull (1998) also advocated comparing one method directly to another by measuring the proportion of the parameter space where the coverage probability is closer to the nominal for one method than the other. Figure 4 plots this metric for both the score method and the adjusted Wald method versus the optimized interval method for the same sample sizes and nominal confidence levels as Figures 2 and 3.

The results are mixed. At the .99 nominal confidence level the coverage of the adjusted Wald method is closer to the nominal in less than 50% of the range of *p* for all sample sizes, whereas the score method is closer for more than 50% of the range of *p* for all sample sizes above 40. At the other three nominal confidence levels both the adjusted Wald and score methods are usually closer to the nominal than the optimized interval method in more 50% of the range of *p* when sample sizes are greater than 20 and less than 50% for smaller sample sizes. Neither method is closer than the optimized interval method to the nominal confidence level in more than 65% of the range of *p* for any of the pairs of sample sizes and nominal confidence levels.

Another metric of interest is the proportion of the range of *p* where the coverage probability is less than the nominal. For exact methods, this proportion is zero by definition. For approximate methods, a small proportion of the range of *p* with coverage probabilities less than the nominal level is preferred. Figure 5 plots this metric over the same sample sizes and nominal confidence levels as Figures 2 to 4. The optimized interval method is closer to zero than the other methods for almost all of the sample sizes and nominal confidence levels. The adjusted Wald is the next best, with the score method performing the worst on this metric.

The approximate methods all have the property that $CL_{n,CL^*} < CL^*$ for most values of CL^* and *n*, so it is of interest how far below the nominal confidence level the actual confidence level is. The actual coverage probability of the optimized interval method can never fall below the nominal minus α, that is $CL_{n,CL^*} \geq CL^* - \alpha$ for every *n* and CL^*. This follows from the restriction that $\alpha_U + \alpha_L = \alpha$ which requires that α_U and $\alpha_L \leq \alpha$ for all *y*. As a result, the $CL^* = 1-\alpha$ level optimized intervals must be contained within the Clopper-Pearson $CL^* = 1-2\alpha$ level intervals. Because the Clopper-Pearson method is an exact method, it follows directly that $CL_{n,CL^*} \geq CL^* - \alpha$ for all *n* and CL^*. The score and the adjusted Wald method have no such restriction on CL_{n,CL^*}.

Figure 6 plots the actual coverage probability of the optimized interval method against sample sizes ranging from 1 to 100 for nominal confidence levels of .80, .90, .95, and .99. Figure 6 shows that the optimized method is always below the nominal except for very small sample sizes. It is often within a distance of $\alpha/2$ of the nominal confidence level, particularly for sample sizes over 20. The performance of the adjusted Wald method for this metric is very similar to the optimized interval method for sample sizes over 10 at the .95 and .99 confidence level. At the .80 and .90 confidence level the adjusted Wald performs very badly, with coverage probabilities of zero for all of the sample sizes when the nominal level is .80. The score method is the opposite, with actual confidence levels substantially below the nominal at the .95 and .99 nominal levels and closer at the .90 and .80 levels.

Figure 3. Root Mean Square Error of Three Approximate Methods. The scatter is of the uniform-weighted root mean squared error of the average coverage probabilities of three approximate methods when p is uniformly distributed for sample sizes of from 1 to 100 with nominal confidence levels of .80, .90, .95, and .99. The optimized interval method is indicated by a "o", the adjusted Wald method by a "+", and the score method by a "<". The relative performance of the three methods for this metric varies according to the nominal confidence level. Each method has at least one nominal confidence level where the root mean squared error is furthest from zero for most of the sample sizes.

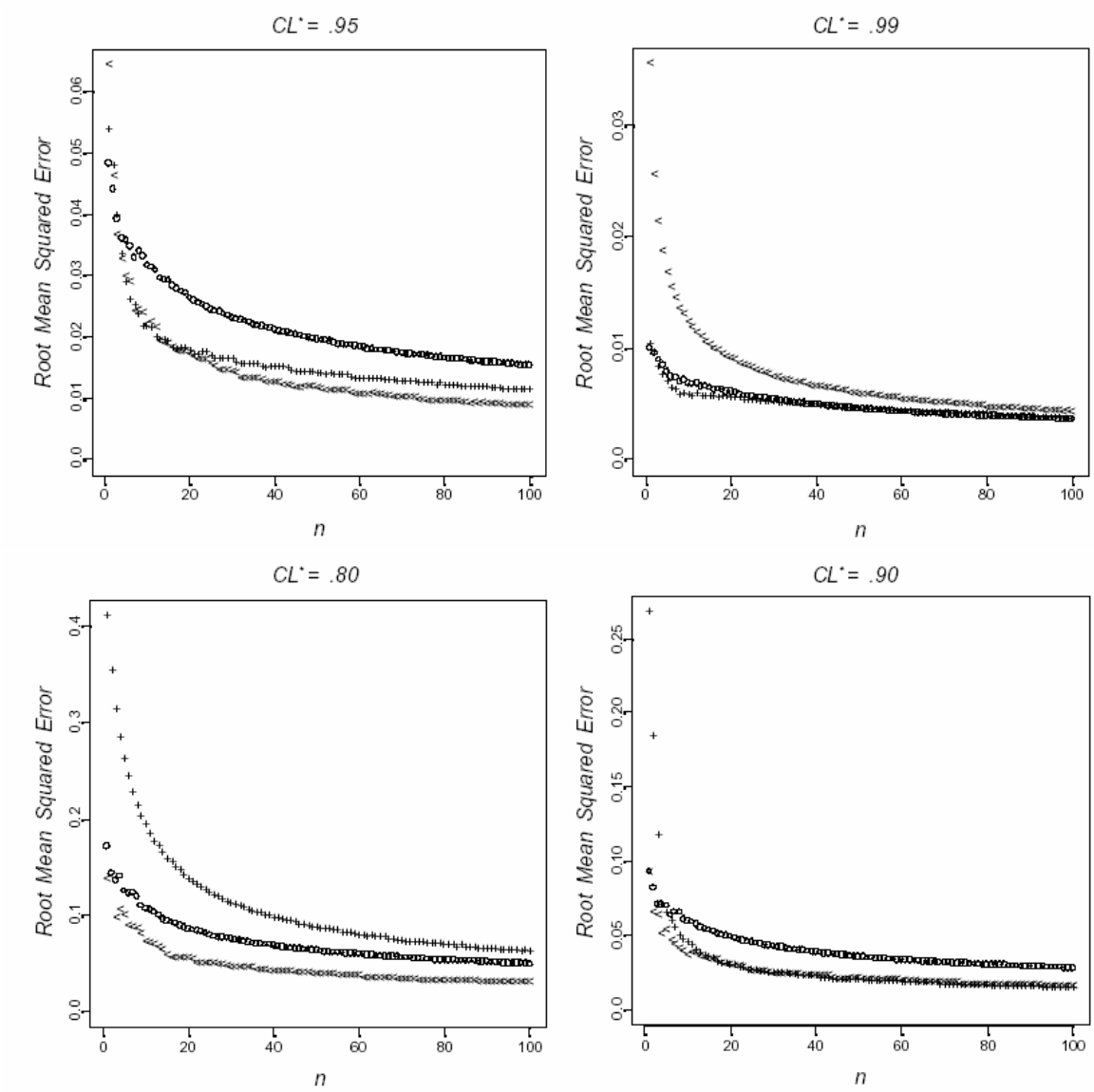


Figure 4. Proportion of Values of p Where Coverage is Closer to Nominal. The scatter is of the proportion of the uniformly distributed values of p for which the adjusted Wald or score method has actual coverage probability closer to the nominal coverage probability than the optimized method for sample sizes of from 1 to 100 with nominal confidence levels of .80, .90, .95, and .99. The adjusted Wald method is indicated by a "o" and the score method by a "+". The horizontal dotted line is at 50%. At the .80, .90, and .95 nominal confidence levels both the adjusted Wald and Score method tend to have coverage probabilities closer to the nominal for more than half the range of p sample sizes over 20 and this is also true for the score method at a nominal confidence level of .99. For the adjusted Wald at nominal confidence level of .99, and for both methods with sample sizes less than 20, the coverage probability is closer to the nominal than the optimized method for less than half the range of for p.

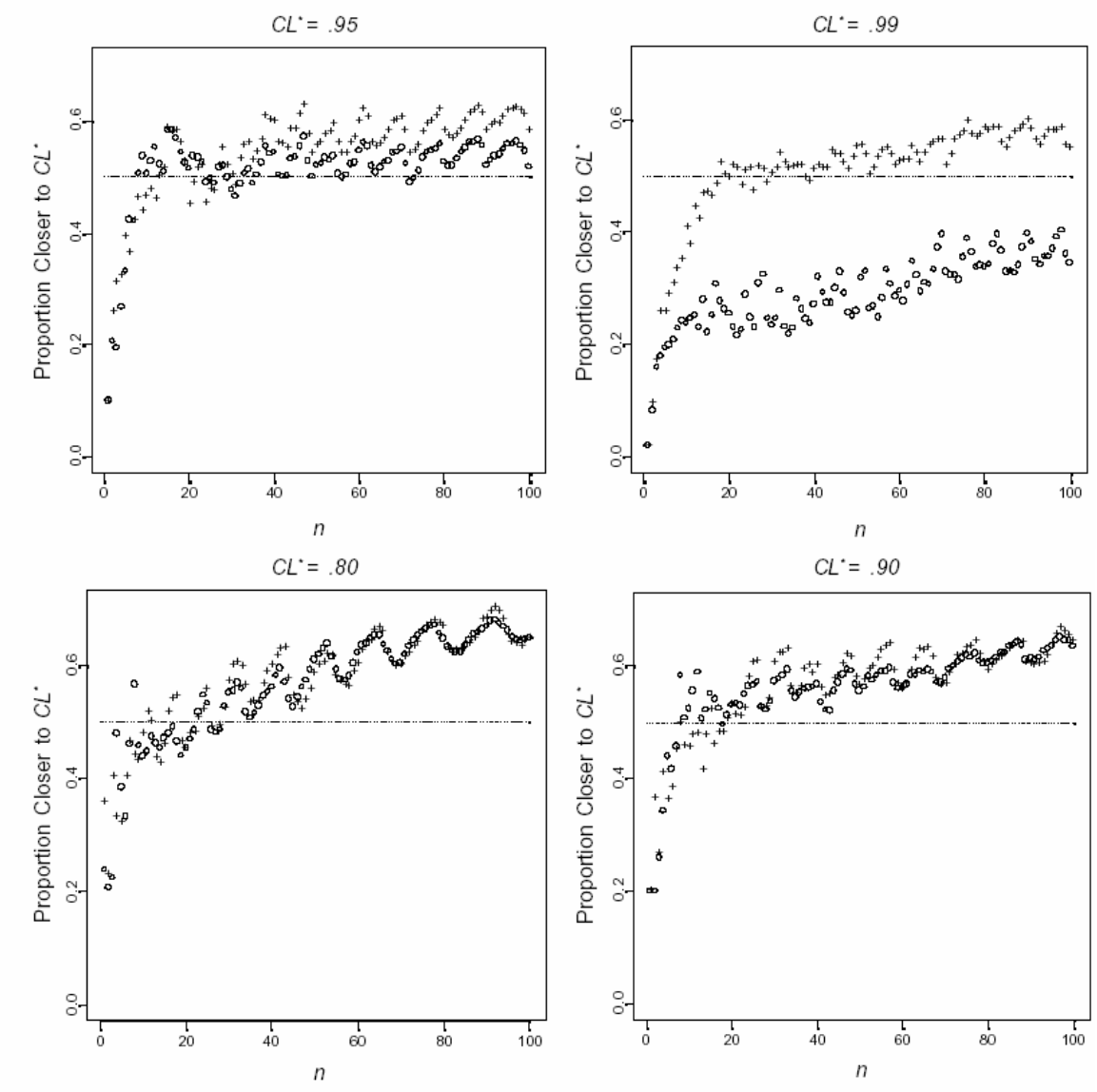


Figure 5. Proportion of p Where Coverage is Less Than the Nominal. The scatter is of the proportion of the uniformly distributed values of p for which a coverage method has actual coverage probability less than the nominal coverage probability for sample sizes of from 1 to 100 with nominal confidence levels of .80, .90, .95, and .99. The optimized interval method is indicated by a "o", the adjusted Wald method by a "+", and the score method by a "<". In general, the optimized interval method has a smaller proportion of the range of p where the actual coverage probability is less than the nominal than the other two methods and this proportion tends to decrease as the sample size increases while it increases for the adjusted Wald and stays at approximately the same level for the score method.

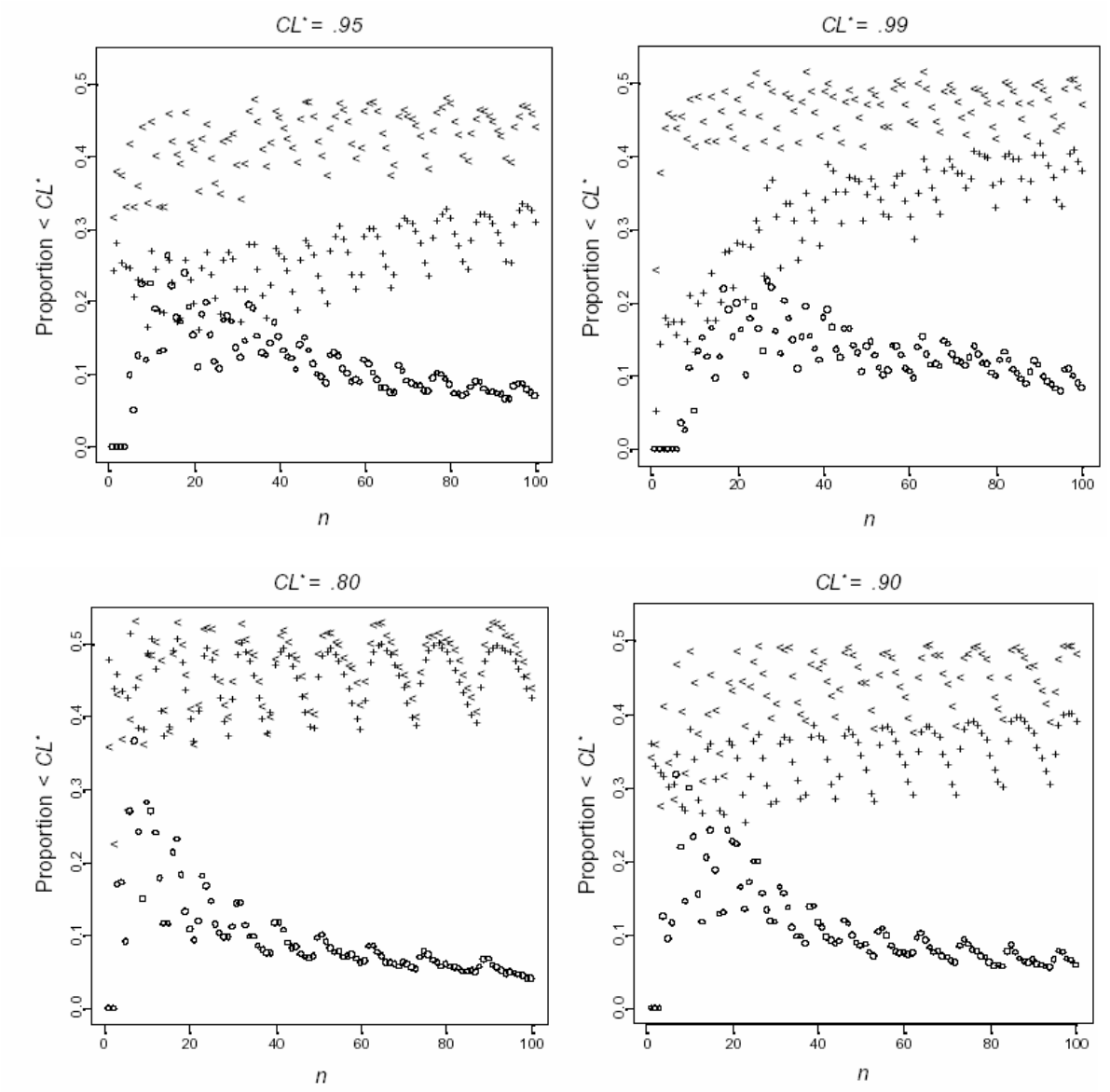
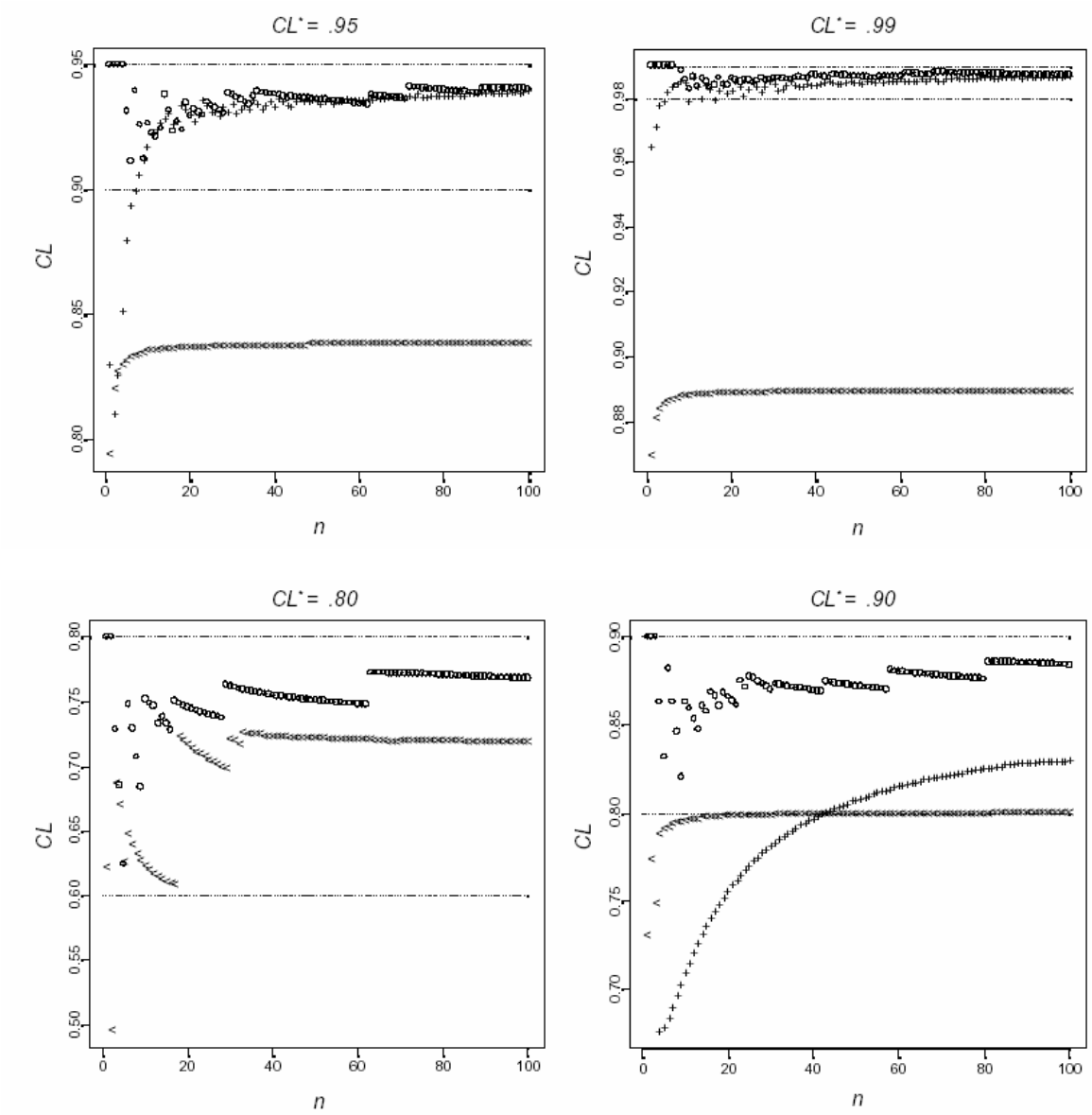


Figure 6. Actual Confidence Levels. The scatter is of the actual confidence levels for three approximate methods for sample sizes of from 1 to 100 with nominal confidence levels of .80, .90, .95, and .99. The optimized interval method is indicated by a "o", the adjusted Wald method by a "+", and the score method by a "<". No actual confidence levels for any sample size are shown for the adjusted Wald method at a nominal confidence level of .80 or for sample sizes less than four at a nominal confidence level of .90. The actual confidence level is zero at all of those points. The upper horizontal dotted line is at the nominal confidence level and the lower dotted line is at the nominal confidence level minus *a*. The actual confidence level for the optimized bound method is always less than nominal level except for very small sample sizes, but it is never less than the nominal level minus *a*. The actual confidence level of the other two methods can be substantially less than the nominal.



Conclusion

The optimized interval method is not an exact method. It should not be used in applications where it is essential that the actual coverage probability be at or above the nominal confidence level across the entire parameter space. For applications where an exact method is not required the optimized method is worth consideration.

Figures 2 – 6 demonstrate that none of the three approximate methods considered in this paper is clearly superior for all of the metrics across all of the sample sizes and nominal confidence levels considered. The investigator needs to determine which metrics are most important and then consult Figures 2 – 6 to determine which method performs best for those metrics at the sample size and nominal confidence level that will be used. If the distance of the actual confidence level from the nominal confidence level and the proportion of the parameter space where coverage falls below the nominal are important considerations then the optimized bound method will often be a good choice.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions, *The American Statistician, 52*, 119-126.

Berger, R. L., & Coutant, B. W. (2001). Comment on small-sample interval estimation of bernoulli and poission parameters, by D. Wardell. *The American Statistician, 55*, 85.

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association, 78*, 108-116.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404-413.

Wardell, D. G. (1997). Small-Sample interval estimation of bernoulli and poisson parameters. *The American Statistician, 51*, 321-325.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 53-62

1538 – 9472/05/$95.00

Inferences About Regression Interactions Via A Robust Smoother With An Application To Cannabis Problems

Rand R. Wilcox Mitchell Earleywine
Department of Psychology
University of Southern California, Los Angeles

A flexible approach to testing the hypothesis of no regression interaction is to test the hypothesis that a generalized additive model provides a good fit to the data, where the components are some type of robust smoother. A practical concern, however, is that there are no published results on how well this approach controls the probability of a Type I error. Simulation results, reported here, indicate that an appropriate choice for the span of the smoother is required so that the actual probability of a Type I error is reasonably close to the nominal level. The technique is illustrated with data dealing with cannabis problems where the usual regression model for interactions provides a poor fit to the data.

Key words: Robust smoothers, curvature, interactions

Introduction

A combination of extant regression methods provides a very flexible and robust approach to detecting and modeling regression interactions. In particular, both curvature and nonnormality are allowed. The main goal in this paper is to report results on the small-sample properties of this approach when a particular robust smoother is used to approximate the regression surface. The main result is that in order to control the probability of a Type I error, an appropriate choice for the span must be used which is a function of the sample size. However, before addressing this issue, we provide a motivating example for considering smoothers when investigating interactions.

A well-known approach to detecting and modeling regression interactions is to assume that for a sample of *n* vectors of observations, (Y_i, X_{i1}, X_{i2}),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i, \quad (1)$$

$i=1,...,n$, where ε is independent of X_{i1} and X_{i2}, $E(\varepsilon) = 0$. The hypothesis of no interaction corresponds to

$$H_0 : \beta_3 = 0.$$

This approach appears to have been first suggested by Saunders (1956). A practical issue is whether this approach is flexible enough to detect and to model an interaction if one exists. We consider data collected by the second author to illustrate that at least in some situations, a more flexible model is required. The data deal with cannabis problems among adult males. Responses from n=296 males were obtained where the two regressors were the participants' use of cannabis (X_1) and consumption of alcohol (X_2). The dependent measure (Y) reflected cannabis dependence as measured by the number of DSM-IV symptoms reported. An issue of interest was determining whether the amount of alcohol consumed alters the association between Y and the amount of cannabis used, and there is the issue of

Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology at the University of Southern California, Los Angeles. M. Earleywine is an Associate Professor at the University of Southern California, Los Angeles.

understanding how the association changes if an interaction exists.

Using a method derived by Stute, González-Manteiga and Presedo-Quindimil (1998), it is possible to test the hypothesis that the model given by equation (1) provides a good fit to the data. If, for example,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}^2 + \varepsilon_i,$$

then there is an interaction, but the family of regression equations given by (1) is inappropriate. The Stute et al. method can be applied using the S-PLUS or R function lintest in Wilcox (2003). Estimating the unknown parameters via least squares, this hypothesis is rejected at the .05 level. A criticism is that when testing the hypothesis that (1) is an appropriate model for the data, and when using the ordinary least squares estimator when estimating the unknown parameters, the probability of a Type I error might not be controlled (Wilcox, 2003).

Replacing the least squares estimator with various robust estimators corrects this problem. Here, using the robust M-estimator derived by Coakley and Hettmansperger (1993), or using a generalization of the Theil-Sen estimator to multiple predictors (see Wilcox, 2005), again the hypothesis is rejected. Moreover, the R (or S-PLUS) function pmodchk in Wilcox (2005) provides a graphical check of how well the model given by (1) fits the data when a least squares estimate of the parameters is used, versus a more flexible fit based on what is called a running interval smoother, and a poor fit based on (1) is indicated. Robust variations give similar results. So, at least in this case, an alternative and more flexible approach to testing the hypothesis of no interaction seems necessary.

To provide more motivation for a more flexible approach when modeling interactions, note that equation (1) implies a nonlinear association between *Y* versus X_1 and X_2. A concern, however, is that a nonlinear association does not necessarily imply an interaction. If, for example, X_1, X_2 and ε are independent and have standard normal distributions, and if $Y = X_1 + X_2^2 + \varepsilon$, the probability of rejecting H_0: $\beta_3 = 0$ is .18 when testing at the .05 level with a sample size of twenty. Of course, in this case, standard diagnostics can be used to detect the curvature, but experience with smoothers suggest that dealing with curvature is not always straightforward.

Suppose instead $Y = X_1 + X_1^2 + |X_2| + \varepsilon$, so there is no interaction even though there is a nonlinear association. Then with a sample size of fifty, and when testing at the .05 level, the probability of rejecting H_0: $\beta_3 = 0$ is .30. In contrast, using the more flexible method described here, the probability of rejecting the hypothesis of no interaction is .042.

If we ignore the result that (1) is an inadequate model for the cannabis data and simply test H_0: $\beta_3 = 0$ (using least squares in conjunction with a conventional T test), or if we test H_0: β_3=0 using a more robust hypothesis testing method derived for the least squares estimator that is based on a modified percentile bootstrap method (Wilcox, 2003), or when using various robust estimators (such as an M-estimator with Schweppe weights or when using the Coakley-Hettmansperger estimator), we reject. But an issue is whether we reject because there is indeed an interaction, or because the model provides an inadequate representation of the data. And another concern is that by using an invalid model, an interaction might be masked.

A more general and more flexible approach when investigating interactions is to test the hypothesis that there exists some functions f_1 and f_2 such that

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \varepsilon. \qquad (2)$$

Equation (2) is called a generalized additive model, a general discussion of which can be found in Hastie and Tibshirani (1990). A special case is where $f_1(X_1) = \beta_1 X_1$, $f_2(X_2) = \beta_2 X_2$, but (2) allows situations where the regression surface is not necessarily a plane, even when there is no interaction. If the model represented by (2) is true, then there is no interaction in the following sense. Pick any two values for X_2,

say 6 and 8. Then no interaction means that the regression line between Y and X_1, given that $X_2 = 6$, is parallel to the regression line between Y and X_1, given that $X_2 = 8$.

For completeness, Barry (1993) derived a method for testing the hypothesis of no interaction assuming an ANOVA-type decomposition where

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) + \varepsilon,$$

in which case the hypothesis of no interaction is

$$H_0 : f_3(X_1, X_2) \equiv 0.$$

Barry (1993) used a Bayesian approach assuming that the (conditional) mean of Y is to be estimated and that prior distributions for f_1, f_2 and f_3 can be specified. The goal in this article is to investigate the small-sample properties of a non-Bayesian method where the mean is replaced by some robust measure of location (cf. Samarov, 1993).

Methodology

There are, in fact, many approaches that might be used that are based on combinations of existing statistical techniques. The problem is finding a combination of methods that controls the probability of a Type I error in simulations even when the sample size is relatively small. One possibility is to use some extension of the method in Dette (1999), this was considered, but in simulations no variation was found that performed well in terms of controlling the probability of a Type I error.

Only one method was found that performs well in simulations; it is based on a combination of methods summarized in Wilcox (2005). The approach is outlined here, and the computational details are relegated to Appendices A and B. Briefly, the method begins by fitting the model given by (2) using the so-called backfitting algorithm (Hastie & Tibshirani, 1990) in conjunction with a what is called a running interval smoother. Generally, smoothers are methods for approximating regression lines without forcing them to have a particular shape such as a straight line. As with most smoothers, the running interval smoother is based in part on something called a span, κ, which plays a role when determining whether the value X is close to a particular value of X_1 (or X_2). Details are provided in Appendix A.

There are many ways of fitting the model given by (2). Here, the focus is on a method where the goal is to estimate a robust measure of location associated with Y, given (X_1, X_2), because of the many known advantages such measures have (e.g., Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 2003, 2005). Primarily for convenience, the focus is on a 20% trimmed mean, but various robust M-estimators are certainly a possibility. The advantages associated with robust measures of location include an enhanced ability to control the probability of a Type I error in situations where methods based on means are known to fail, and substantial gains in power, over methods based on means, even under slight departures from normality. (Comments about using the mean, in conjunction with the proposed method, are made in the final section of this paper.) Here, the main reason for not using a robust M-estimator (with say, Huber's Ψ), is that this estimator requires division by the median absolute deviation (MAD) statistic, and in some situations considered here, when the sample size is small, MAD is zero.

The running interval smoother provides a predicted value for Y, given (X_{i1}, X_{i2}), say $\hat{Y}_i$; see Appendix A. Next, compute the residuals $r_i = Y_i - \hat{Y}_i$. If the model given by (2) is true, meaning that there is no interaction, then the regression surface when predicting r, given (X_1, X_2), should be a horizontal plane. The hypothesis that this regression surface is indeed a horizontal plane can be tested using the method derived by Stute et al. (1998). The details can be found in Appendix B.

Results

Simulations were conducted as a partial check on the ability of the method, just outlined, to control the probability of a Type I error. Values for X_1, X_2 and ε were generated from four types of distributions: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. For non-normal distributions, observations were generated from a g-and-h distribution which is described in Appendix C. The goal was to check on how the method performs under normality, plus what would seem like extreme departures from normality, with the idea that if good performance is obtained under extreme departures from normality, the method should perform reasonably well with data encountered in practice. The correlation between X_1 and X_2 was taken to be either ρ=0 or ρ=.5.

Initial simulation results revealed that the actual probability of a Type I error, when testing at the .05 level, is sensitive to the span, κ. (Härdle & Mammen, 1993, report a similar result for a method somewhat related to the problem at hand.) If the span is too large, the actual Type I error probability can drop well below the nominal level. When testing at the .05 level, simulations were used to approximate a reasonable choice for κ. Here, the span corresponding to the sample sizes 20, 30, 50, 80 and 150 are taken to be .4, .36, .18, .15 and .09, respectively. It is suggested that when $20 \leq n \leq 150$, interpolation based on these values be used, and for $n>150$ use a span equal to .09. For $n>150$ and sufficiently large, perhaps the actual Type I error probability is well below the nominal level, but exactly how the span should be modified when $n>150$ is an issue that is in need of further investigation.

Table 1 contains $\hat{\alpha}$, the estimated probability of making a Type I error when testing at the .05 level. n=20, and when $Y=\varepsilon$ or $Y = X_1 + X_2^2 + \varepsilon$. (The g and h values are explained in Appendix C.) Simulations were also run when $Y = X_1 + X_2 + \varepsilon$, the results were very similar to the case $Y=\varepsilon$, so for brevity they are not reported. No situation was found where the estimated probability of a Type I error exceeded the nominal .05 level. The main difficulty is that when marginal distributions have a skewed, heavy-tailed distribution, ρ=.5, and there is curvature, the estimated probability of a Type I error dropped below .01. This situation corresponds to what would seem like an extreme departure from normality as indicated in Appendix C.

An Illustration

Returning to the cannabis data described in the introduction, the hypothesis of no interaction is rejected at the .05 level when testing the model given by (2). (The test statistic described in Appendix B is D=3.37 and the .05 critical value is 1.79.) To provide some overall sense of the association, Figure 1 shows an approximation of the regression surface based on a smooth derived by Cleveland and Devlin (1988) called loess. (Using the robust smooth in Wilcox, 2003, section 14.2.3, gives similar results when the span is set to 1.2.) Note the nonlinear appearance of the surface. Also, there appears to be little or no association over some regions of the X_1 and X_2 values.

Figure 2 shows the plot based on X_1 and X_2 versus the residuals corresponding to the generalized additive model given by (2). This plot should be a horizontal plane if there is no interaction. As is evident, the surface appears to be nonlinear, at least to some degree

To further explore the nature of the interaction, first it is noted that the quartiles associated with X_2 (alcohol use) are -0.732, -0.352 and 0.332. The left panel of Figure 3 shows three smooths between Y and X_1; they are the smooths between Y and X_1 given that $X_2 = -0.73$, X_2=-0.352 and $X_3 = 0.332$. (These smooths were created using a slight generalization of the kernel regression estimator in Fan, 1993; see R or S-PLUS function kercon in Wilcox, 2005, Ch. 11.)

Table 1: Estimated probability of a Type I error, *n*=20.

				$Y=\varepsilon$		$Y = X_1 + X_2^2 + \varepsilon$	
g	h	g	h	ρ=0	ρ=.5	ρ=0	ρ=.5
0.0	0.0	0.0	0.0	.033	.034	.047	.035
		0.0	0.5	.039	.034	.026	.031
		0.5	0.0	.045	.043	.045	.034
		0.5	0.5	.037	.035	.035	.032
0.0	0.5	0.0	0.0	.031	.032	.019	.015
		0.0	0.5	.032	.024	.020	.012
		0.5	0.0	.033	.031	.016	.013
		0.5	0.5	.029	.024	.023	.013
0.5	0.0	0.0	0.0	.029	.022	.036	.022
		0.0	0.5	.031	.020	.032	.014
		0.5	0.0	.040	.039	.037	.028
		0.5	0.5	.029	.027	.025	.020
0.5	0.5	0.0	0.0	.028	.024	.024	.003
		0.0	0.5	.026	.017	.015	.003
		0.5	0.0	.035	.029	.014	.006
		0.5	0.5	.020	.015	.015	.007

Figure 1: An approximation of the regression surface based on the smoother loess.

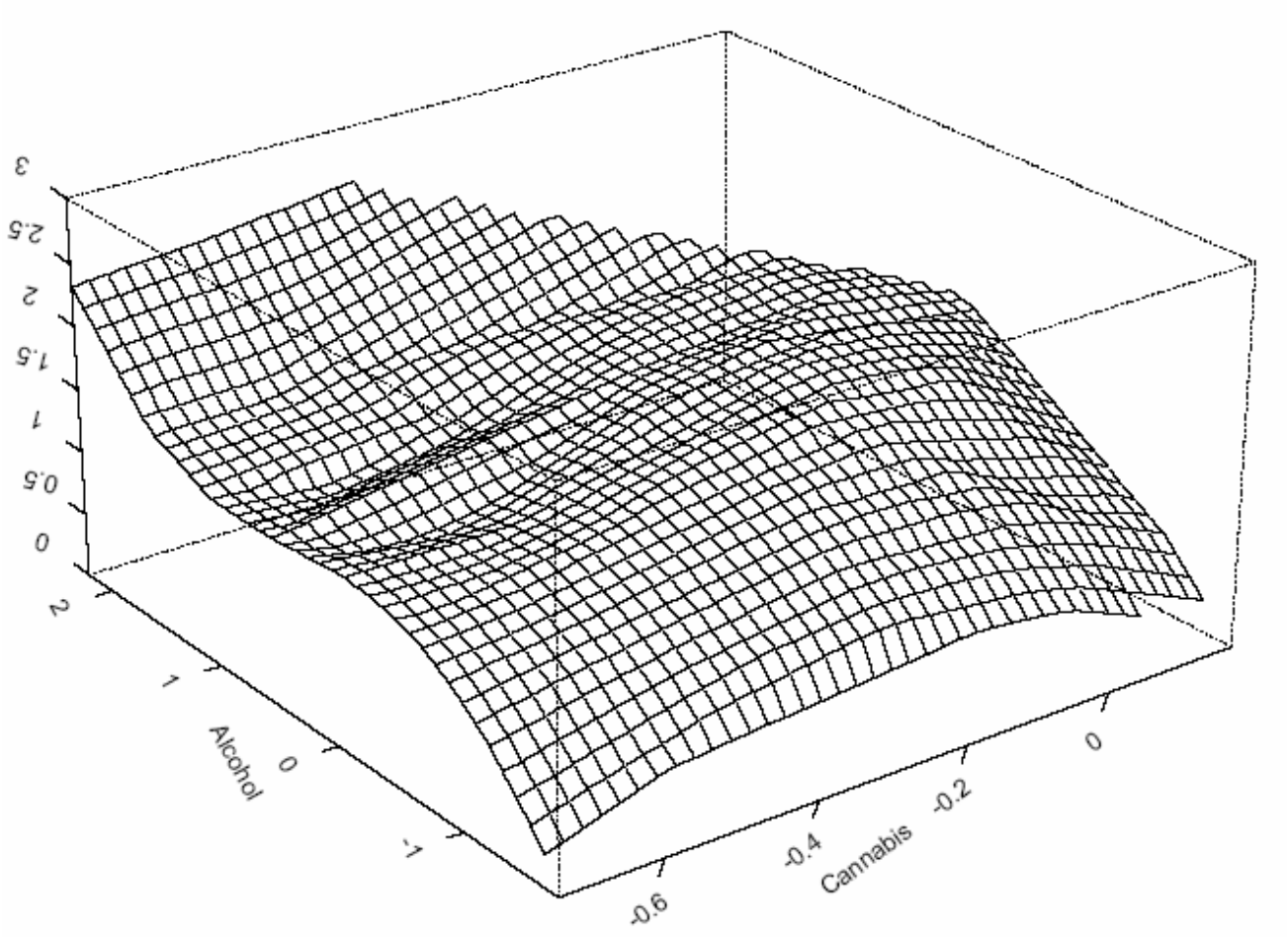
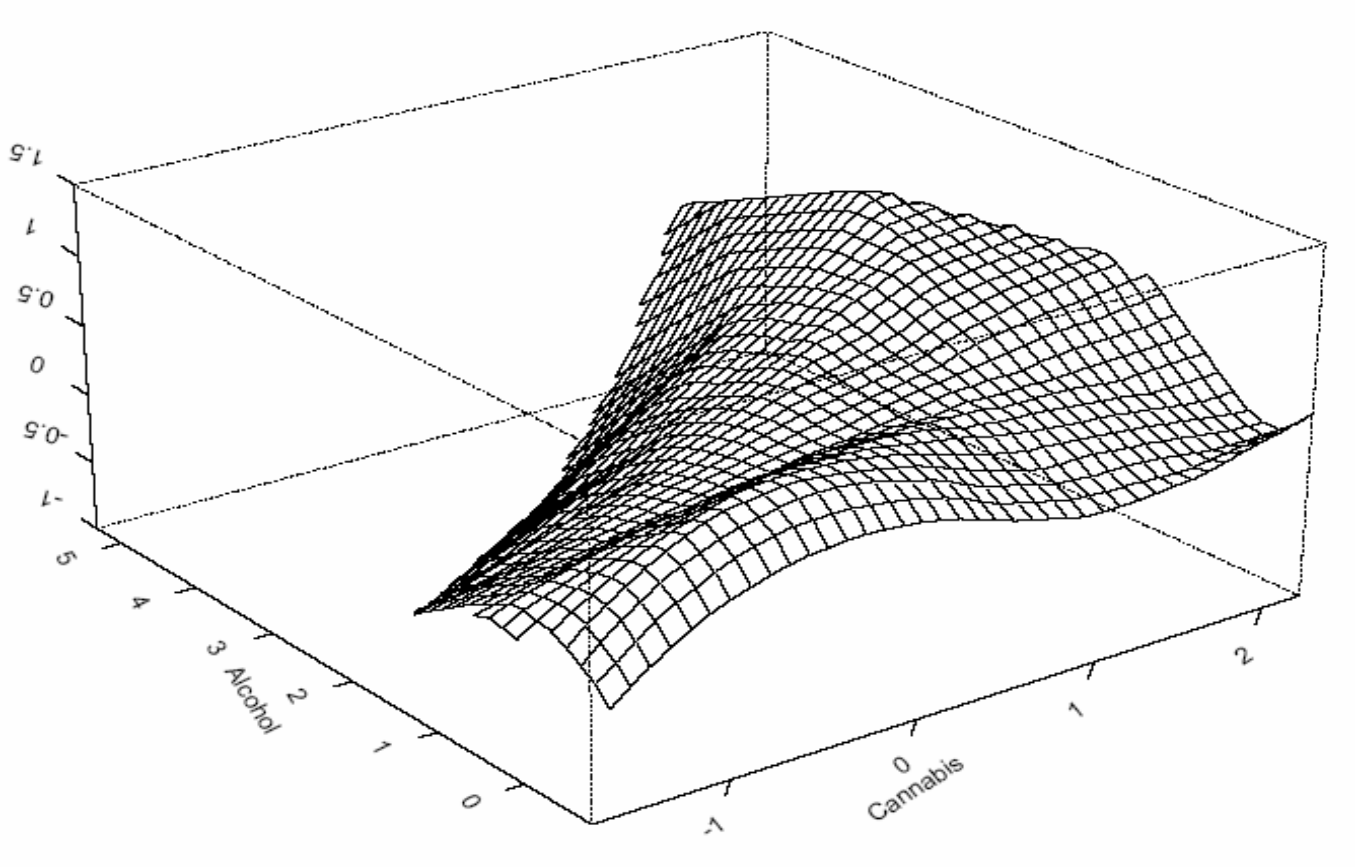


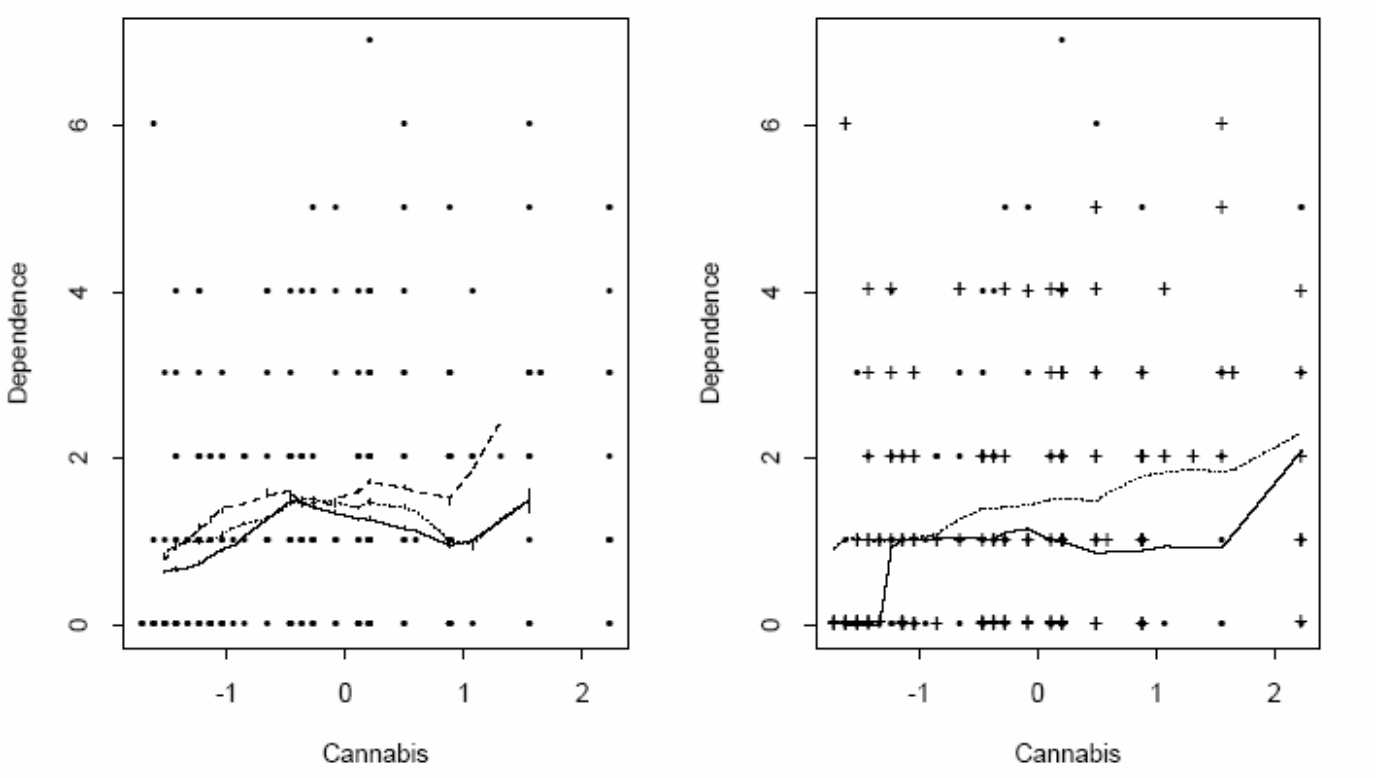
Figure 2: A smooth of the residuals stemming from the generalized additive model versus the two predictors.



When there is no interaction, all three regression lines should be approximately parallel which is not the case. The regression lines corresponding to $X_2 = -0.73$ and -0.352 are reasonably parallel, and they are approximately horizontal suggesting that there is little association between Y and X_1 for these special cases.

But for $X_3 = 0.332$, the association changes, particularly in the right portion of Figure 1 where the association becomes more positive. If the data are split into two groups according to whether X_{i2} is less than the median of the values $X_{12}, ..., X_{n2}$, -0.352, and then create a smooth between Y and X_1, the result is shown in right panel of Figure 3.

Figure 3: Some smooths used to investigate interactions.



Conclusion

In principle, the method in this article can be used with any measure of location. It is noted, however, that if the 20% trimmed mean is replaced by the sample mean, poor and unstable control over the probability of a Type I error results.

Finally, all of the methods used in this paper are easily applied using the S-PLUS or R functions in Wilcox (2005). (These functions can be downloaded as described in chapter 1.) Information about S-PLUS can be obtained from www.insightful.com, and R is a freeware variant of S-PLUS that can be downloaded from www.R-project.org. For convenience, the relevant functions for the problem at hand have been combined into a single function called adtest. If, for example, the *X* values are stored in an S-PLUS matrix x, and the *Y* values are stored in y, the command adtest(x,y) tests the hypothesis that the model given by (2) is true.

References

Barry, D. (1993). Testing for additivity of a regression function. *Annals of Statistics, 21*, 235-254.

Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, 83*, 596-610.

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88*, 872-880.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variances estimator. *Annals of Statistics, 27*, 1012-1040.

Fan, J. (1993). Local linear smoothers and their minimax efficiencies. *The Annals of Statistics, 21*, 196-216.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.

Hardle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Hardle, W., & Mammen, E. (1993). Comparing non-parametric versus parametric regression fits. *Annals of Statistics, 21*, 1926-1947.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Underststanding Robust and Exploratory Data Analysis.* (pp. 297-336). New York: Wiley.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association, 88*, 836-847.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16*, 209-222.

Stute, W., Gonzlez-Manteiga, W. G. & Presedo-Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association, 93*, 141-149.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. San Diego, CA: Academic Press.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition. San Diego, CA: Academic Press.

Appendix A

We begin by describing how to compute a 20% trimmed mean based on a sample of *m* observations. Put the observations in ascending order yielding $W_{(1)} \leq \cdots \leq W_{(m)}$. Let ℓ=[.2*m*], where [.2*m*] means to round .2*m* down to the nearest integer. Then the 20% trimmed mean is

$$\bar{X}_t = \frac{1}{n-2\ell} \sum_{i=\ell+1}^{n-\ell} W_{(i)}.$$

In terms of efficiency (achieving a small standard error relative to the usual sample mean), 20% trimming performs very well under normality but continues to perform well in situations where the sample mean performs poorly (e.g., Rosenberger & Gasko, 1983).

Now, we describe the running interval smoother in the one-predictor case. Consider a random sample $(X_1, Y_1), ..., (X_n, Y_n)$ and let κ be some constant that is chosen in a manner to be described. The constant κ is called the span. The median absolute deviation (MAD), based on $X_1, ..., X_n$, is the median of the *n* values $|X_1 - M|, ..., |X_n - M|$, where *M* is the usual median. Let MADN=MAD/.6745. Under normality, MADN estimates σ, the standard deviation. Then the point *X* is said to be close to X_i if

$$|X_i - X| \leq \kappa \times MADN.$$

Thus, for normal distributions, *X* is close to X_i if *X* is within κ standard deviations of X_i. Then $\hat{Y}_i$ is the 20% trimmed mean of the Y_j values for which X_j is close to X_i. In exploratory work, a good choice for the span is often κ=.8 or 1, but for the situation at hand an alternative choice is needed.

Virtually any smoother, including the one used here, can be extended to the generalized additive model given by (2) using the backfitting algorithm in Hastie and Tibshirani (1990). Set *k*=0 and let f_j^0 be some

initial estimate of f_j (j=1, 2). Here, $f_j^0(X_j) = S_j(Y \mid X_j)$, where $S_j(Y|X_j)$ is the running interval smooth based on the jth predictor, ignoring the other predictor under investigation. Next, iterate as follows.

1. Increment k.
2. Let

$$f_1^k(X_1) = S_1(Y - f_2^{k-1}(X_2) \mid X_1)$$

and

$$f_2^k(X_2) = S_2(Y - f_1^{k-1}(X_1) \mid X_2).$$

3. Repeat steps 1 and 2 until convergence.

Finally, estimate β_0 with the 20% trimmed mean of the values $Y_i - \sum f_j^k(Y_i \mid X_{ij})$, i=1,...,n. The computations are performed by R or S-PLUS function adrun in Wilcox (2005).

Appendix B

This appendix describes the method for testing the hypothesis of no interaction. Fit the generalized additive model as described in Appendix A yielding $\hat{Y}_i$, and let $r_i = Y_i - \hat{Y}_i$, i=1,...,n. The goal is to test the hypothesis that the regression surface, when predicting the residuals, given (X_{i1}, X_{i2}), is a horizontal plane. This is done using the wild bootstrap method derived by Stute, González-Manteiga and Presedo-Quindimil (1998). Let $\overline{r}_t$ be the 20% trimmed mean based on the residuals $r_1,...,r_n$. Fix j and set $I_i = 1$ if simultaneously $X_{i1} \le X_{j1}$ and $X_{i2} \le X_{j2}$, otherwise $I_i = 0$.

Let

$$R_j = \frac{1}{\sqrt{n}} \sum I_i (r_i - \overline{r}_t)$$

$$= \frac{1}{\sqrt{n}} \sum I_i v_i, \qquad (3)$$

where $v_i = r_i - \overline{r}_t$.

The test statistic is the maximum absolute value of all the R_j values. That is, the test statistic is

$$D = max \mid R_j \mid . \qquad (4)$$

An appropriate critical value is estimated with the wild bootstrap method as follows. Generate $U_1,...,U_n$ from a uniform distribution and set

$$V_i = \sqrt{12}(U_i - .5),$$

$$v_i^* = v_i V_i,$$

and

$$r_i^* = \overline{r}_t + v_i^*.$$

Then based on the n pairs of points (X_1, X_2, r_1^*), ..., (X_n, X_n, r_n^*), compute the test statistic as described in the previous paragraph and label it D^*. Repeat this process B times and label the resulting (bootstrap) test statistics $D_1^*,...,D_B^*$. Here, B=500 is used. Finally, put these B values in ascending order yielding $D_{(1)}^* \le \cdots \le D_{(B)}^*$. Then the critical value is $D_{(u)}^*$, where u=(1-α)B rounded to the nearest integer. That is, reject if

$$D \ge D_{(u)}^*.$$

Appendix C

Details regarding the simulations are as follows. Observations were generated where the marginal distributions have a g-and-h distribution (Hoaglin, 1985) which includes the normal distribution as a special case. More precisely, observations Z_{ij}, (i=1,...,n; j=1, 2) were initially generated from a multivariate normal distribution having correlation ρ, then the marginal distributions were transformed to

$$X_{ij} = \begin{cases} \dfrac{exp(gZ_{ij})-1}{g} exp(hZ_{ij}^2/2), & \text{if } g>0 \\ Zexp(hZ_{ij}^2/2), & \text{if } g=0 \end{cases}$$

where g and h are parameters that determine the third and fourth moments. The four (marginal) g-and-h distributions examined were the standard normal (g=h=0), a symmetric heavy-tailed distribution (g=0, h=.5), an asymmetric distribution with relatively light tails (g=.5, h=0), and an asymmetric distribution with heavy tails (g=h=.5). Here, two choices for ρ were considered: 0 and .5.

Table 2 shows the theoretical skewness (κ_1) and kurtosis (κ_2) for each distribution considered. When g>0 and h>1/k, $E(X^k)$ is not defined and the corresponding entry in Table 2 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). Some of these distributions might appear to represent extreme departures from normality, but the idea is that if a method performs reasonably well in these cases, this helps support the notion that they will perform well under conditions found in practice.

Table 2: Some properties of the g-and-h distribution.

g	h	κ_1	κ_2
0.0	0.0	0.00	3.0
0.0	0.5	0.00	---
0.5	0.0	1.75	8.9
0.5	0.5	---	---

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 63-74

1538 – 9472/05/$95.00

Regression By Data Segments Via Discriminant Analysis

Stan Lipovetsky Michael Conklin
GfK Custom Research Inc
Minneapolis, Minnesota

It is known that two-group linear discriminant function can be constructed via binary regression. In this article, it is shown that the opposite relation is also relevant – it is possible to present multiple regression as a linear combination of a main part, based on the pooled variance, and Fisher discriminators by data segments. Presenting regression as an aggregate of the discriminators allows one to decompose coefficients of the model into sum of several vectors related to segments. Using this technique provides an understanding of how the total regression model is composed of the regressions by the segments with possible opposite directions of the dependency on the predictors.

Key words: Regression, discriminant analysis, data segments

Introduction

Linear Discriminant Analysis (LDA) was introduced by Fisher (1936) for classification of observations into two groups by maximizing the ratio of between-group variance to within-group variance (Rao, 1973; Lachenbruch, 1979; Hand, 1982; Dillon & Goldstein, 1984; McLachlan, 1992; Huberty, 1994). For two-group LDA, the Fisher linear discriminant function can be represented as a linear regression of a binary variable (groups indicator) by the predictors (Fisher, 1936; Anderson, 1958; Ladd, 1966; Hastie, Tibshirani & Buja, 1994; Ripley, 1996). Many-group LDA can be described in terms of the Canonical Correlations Analysis (Bartlett, 1938; Kendall & Stuart, 1966; Dillon & Goldstein, 1984; Lipovetsky, Tishler, & Conklin, 2002). LDA is used in various applications, for example, in marketing research (Morrison, 1974; Hora & Wilcox, 1982; Lipovetsky & Conklin, 2004).

Considered in this article is the possibility of presenting a multiple regression by segmented data as a linear combination of the Fisher discriminant functions. This technique is based on the relationship between total and pooled variances. Using this approach, we can interpret regression as an aggregate of discriminators, that allows us to decompose the coefficients of regression into a sum of vectors related to the data segments. Such a decomposition helps explain how a regression by total data could have the opposite direction of the dependency on the predictors, in comparison with the coefficients related to each segment.

These effects correspond to well-known Simpson's and Lord's paradoxes (Blyth, 1972; Holland & Rubin, 1983; Good & Mittal, 1987; Pearl, 2000; Rinott & Tam, 2003; Skrondal & Rabe-Hesketh, 2004; Wainer & Brown, 2004), and to treatment and causal effects in the models (Arminger, Clogg & Sobel, 1995; Rosenbaum, 1995; Winship & Morgan, 1999).

The article is organized as follows. Linear discriminant analysis and its relation to binary regression are first described. The next section considers regression by segmented data and its decomposition by Fisher discriminators, followed by a numerical example and a summary.

Stan Lipovetsky joined GfK Custom Research as a Research Manager in 1998. His primary areas of research are multivariate statistics, multiple criteria decision making, econometrics, and marketing research. Email him at: slipovetsky@gfkcustomresearch.com. Michael Conklin is Senior Vice-President, Analytic Services for GfK Custom Research. His research interests include Bayesian methods and the analysis of categorical and ordinal data.

Methodology

Consider the main features of LDA. Denote X a data matrix of n by p order consisting of n rows of observations by p variables $x_1, x_2, \ldots, x_p$. Also denote y a vector of size n consisting of binary values 1 or 0 that indicate belonging of each observations to one or another class. Suppose there are n_1 observations in the first class (y =1), n_2 observations in the second class (y =0), and total number of observations $n=n_1+n_2$. Construct a linear aggregate of x-variables:

$$z = X\,a\,, \qquad (1)$$

where a is a vector of p-th order of unknown parameters, and z is an n-th order vector of the aggregate scores. Averaging scores z (1) within each group yields two aggregates:

$$z^{(1)} = m^{(1)}a\,, \quad z^{(2)} = m^{(2)}a\,, \qquad (2)$$

where $m^{(1)}$ and $m^{(2)}$ are vectors of p-th order of mean values $m_j^{(1)}$ and $m_j^{(2)}$ of each j-th variable x_j within the first and second group of observations, respectively. The maximum squared distance between two groups $\|z^{(1)}-z^{(2)}\|^2 = \|(m^{(1)}-m^{(2)})a\|^2$ versus the pooled variance of scores $a'S_{pool}\,a$ defines the objective for linear discriminator:

$$F = \frac{a'(m^{(1)}-m^{(2)})(m^{(1)}-m^{(2)})'a}{a'S_{pool}\,a}\,, \qquad (3)$$

with elements of the pooled matrix defined by combined cross-products of both groups:

$$(S_{pool})_{jk} = \sum_{i=1}^{n_1}(x_{ji}-m_j^{(1)})(x_{ki}-m_k^{(1)}) + \sum_{i=1}^{n_2}(x_{ji}-m_j^{(2)})(x_{ki}-m_k^{(2)})\,. \qquad (4)$$

Equation (3) can represent as a conditional objective:

$$F = a'(m^{(1)}-m^{(2)})(m^{(1)}-m^{(2)})'a - \lambda(a'S_{pool}\,a-1)\,, \qquad (5)$$

where λ is Lagrange multiplier. The first-order condition $\partial F/\partial a = 0$ yields:

$$(m^{(1)}-m^{(2)})(m^{(1)}-m^{(2)})'a = \lambda\,S_{pool}\,a\,, \qquad (6)$$

that is a generalized eigenproblem. The matrix at the left-hand side (6) is of the rank one because it equals the outer product of a vector of the group means' differences. So the problem (6) has just one eigenvalue different from zero and can be simplified. Using a constant of the scalar product $c = (m^{(1)}-m^{(2)})'a$, reduces (6) to the linear system:

$$S_{pool}\,a = q\,(m^{(1)}-m^{(2)})\,, \qquad (7)$$

where $q=c/\lambda$ is another constant. The solution of this system is:

$$a = S_{pool}^{-1}\,(m^{(1)}-m^{(2)})\,, \qquad (8)$$

that defines Fisher famous two-group linear discriminator (up to an arbitrary constant).

The same Fisher discriminator (8) can be obtained if instead of the pooled matrix (4) the total matrix of second-moments defined as a cross-product $X'X$ of the centered data is used, so the elements of this matrix are:

$$(S_{tot})_{jk} = \sum_{i=1}^{n}(x_{ji}-m_j)(x_{ki}-m_k)\,, \qquad (9)$$

where m_j corresponds to mean value of each x_j by total sample of size n. Similarly to transformation known in the analysis of variance, consider decomposition of the cross-product (9) into several items when the total set of n observations is divided into subsets with sizes n_t with t = 1, 2, …, T :

$$\begin{aligned}(S_{tot})_{jk} &= \sum_{i=1}^{n}(x_{ji}-m_j)(x_{ki}-m_k) = \sum_{t=1}^{T}\sum_{i=1}^{n_t}[(x_{ji}^{(t)}-m_j^{(t)})+(m_j^{(t)}-m_j)]x_{ki}^{(t)} \\ &= \sum_{t=1}^{T}\sum_{i=1}^{n_t}(x_{ji}^{(t)}-m_j^{(t)})x_{ki}^{(t)} + \sum_{t=1}^{T}\sum_{i=1}^{n_t}(m_j^{(t)}-m_j)x_{ki}^{(t)} \\ &= \sum_{t=1}^{T}\sum_{i=1}^{n_t}(x_{ji}^{(t)}-m_j^{(t)})(x_{ki}^{(t)}-m_k^{(t)}) + \sum_{t=1}^{T}n_t(m_j^{(t)}-m_j)(m_k^{(t)}-m_k)\,.\end{aligned} \qquad (10)$$

The obtained double sum equals the pooled second moment (4) for T groups, and the last sum corresponds to a total (weighted by sub-sample sizes) of the second moment of group means centered by the total means of the variables. So (10) can be rewrote in a matrix form as:

$$S_{tot} = S_{pool} + \sum_{t=1}^{T} n_t (m^{(t)} - m)(m^{(t)} - m)', \quad (11)$$

where $m^{(t)}$ is a vector of mean values $m_j^{(t)}$ of each j-th variable within t-th group, and m is a vector of means for all variables by the total sample.

Consider the case of two groups, T=2. Then (11) can be reduced to

$$\begin{aligned} S_{tot} &= S_{pool} + n_1 \left(m^{(1)} - \frac{n_1 m^{(1)} + n_2 m^{(2)}}{n_1 + n_2} \right) \\ &\cdot \left(m^{(1)} - \frac{n_1 m^{(1)} + n_2 m^{(2)}}{n_1 + n_2} \right)' \\ &+ n_2 \left(m^{(2)} - \frac{n_1 m^{(1)} + n_2 m^{(2)}}{n_1 + n_2} \right) \quad (12) \\ &\cdot \left(m^{(2)} - \frac{n_1 m^{(1)} + n_2 m^{(2)}}{n_1 + n_2} \right)' \\ &= S_{pool} + h(m^{(1)} - m^{(2)})(m^{(1)} - m^{(2)})', \end{aligned}$$

where $h = n_1 n_2/(n_1+n_2)$ is a constant of the harmonic sum of sub-sample sizes. In place of the pooled matrix S_{pool} let us use the total matrix S_{tot} (12) in the LDA problem (7):

$$\begin{aligned} &\left(S_{pool} + h(m^{(1)} - m^{(2)})(m^{(1)} - m^{(2)})' \right) a \\ &= q(m^{(1)} - m^{(2)}) \end{aligned} . \quad (13)$$

Applying a known Sherman-Morrison formula (Rao, 1973; Harville, 1997)

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1} uv' A^{-1}}{1 + u' A^{-1} v}, \quad (14)$$

where A is a non-singular square n-th order matrix, u and v are vectors of n-th order, the matrix in the left-hand side (13) is inverted and solution obtained:

$$\begin{aligned} a &= S_{tot}^{-1}(m^{(1)} - m^{(2)})\, q \\ &= \frac{q}{1 + h(m^{(1)} - m^{(2)})' S_{pool}^{-1}(m^{(1)} - m^{(2)})} \quad (15) \\ &\cdot S_{pool}^{-1}(m^{(1)} - m^{(2)}). \end{aligned}$$

Comparison of (8) and (15) shows that both discriminant functions coincide (up to unimportant in LDA constant in the denominator (15)), so we can use S_{tot} instead of S_{pool} .

This feature of proportional solutions for the pooled or total matrices holds for more than two classification groups as well. Consider a criterion of maximizing ratio (3) of between-group to the within-group variances for many groups. Using the relation (11) yields:

$$\begin{aligned} F &= \frac{a'(S_{tot} - S_{pool})a}{a' S_{pool}\, a} \\ &= \frac{a' \left(\sum_{t=1}^{T} n_t (m^{(t)} - m)(m^{(t)} - m)' \right) a}{a' S_{pool}\, a} \end{aligned} . \quad (16)$$

Similarly to derivation (5)-(6), (16) is reduced to an eigenproblem:

$$\left(\sum_{t=1}^{T} n_t (m^{(t)} - m)(m^{(t)} - m)' \right) a = \lambda S_{pool}\, a, \quad (17)$$

that is a generalized eigenproblem for the many groups. Denoting the scalar products at the left-hand side (17) as some constants $c_t = (m^{(t)} - m)'a$, the solution of (22) via a linear combination of Fisher discriminators is presented:

$$a = \sum_{t=1}^{T} c_t n_t S_{pool}^{-1}(m^{(t)} - m). \quad (18)$$

In the case of two groups we have simplification (12) that reduces the eigenproblem (17) to the solution (8). But the discriminant functions in

multi-group LDA with the pooled matrix or the total matrix in (17) are the same (up to a normalization) – a feature similar to two group LDA (15). To show this, rewrite (17) using (16) in terms of these two matrices as a generalized eigenproblem:

$$(S_{tot} - S_{pool})a = \lambda S_{pool}\, a\,. \qquad (19)$$

Multiplying S_{pool}^{-1} by the relation (19) reduces it to a regular eigenproblem $(S_{pool}^{-1}S_{tot})\,a = (\lambda+1)\,a$. Taking the objective (16) with the total matrix in denominator, another generalized eigenproblem is obtained:

$$(S_{tot} - S_{pool})b = \mu S_{tot}\, b\,, \qquad (20)$$

with eigenvalues μ and eigenvectors b in this case. Multiplying S_{pool}^{-1} by the relation (20), it is represented as $(S_{pool}^{-1}S_{tot})\,b = (1/(1-\mu))\,b$. Both problems (19) and (20) are reduced to the eigenproblem for the same matrix $S_{pool}^{-1}S_{tot}$ with the eigenvalues connected as $(1+\lambda)(1-\mu)=1$ and with the coinciding eigenvectors a and b.

Now, consider some properties of linear regression related to discriminant analysis. Multiple regression can be presented in a matrix form as a model:

$$y = X\,a + \varepsilon\,, \qquad (21)$$

where Xa is a vector of theoretical values of the dependent variable y (corresponding to the linear aggregate z (1)), and ε denotes a vector of errors. The Least Squares objective for minimizing is:

$$LS = \|\varepsilon\|^2 = (y - Xa)'(y - Xa) = y'y - 2a'X'\,y + a'X'Xa\,. \qquad (22)$$

The condition for minimization $\partial LS / \partial a = 0$ yields a normal system of equations:

$$(X'X)a = X'y\,, \qquad (23)$$

with the solution for the coefficients of the regression model:

$$a = (X'X)^{-1}X'y\,. \qquad (24)$$

Matrix of the second moments $X'X$ in (23) for the centered data is the same matrix S_{tot} (9). If the dependent variable y is binary, then the vector $X'y$ is proportional to the vector of differences between mean values by two groups $m^{(1)} - m^{(2)}$, and solution (24) is proportional to the solution (15) for the discriminant function defined via S_{tot}. As it was shown in (15), the results of LDA are essentially the same with both S_{tot} or S_{pool} matrices. Although the Fisher discriminator can be obtained in regular linear regression of the binary group indicator variable by the predictors, a linear regression with binary output can also be interpreted as a Fisher discriminator. Predictions $z=Xa$ (21) by the regression model are proportional to the classification (1) by the discriminator (15).

Regression as an Aggregate of Discriminators

Now, the regression is described by data segments presented via an aggregate of discriminators. Suppose the data are segmented; for instance, the segments are defined by clustering the independent variables, or by several intervals within a span of the dependent variable variation. Identify the segments by index t =1,…,T to present the total second-moment matrix $S_{tot} = X'X$ as the sum (11) of the pooled second-moment matrix S_{pool} and the total of outer products for the vectors of deviations of each segment's means from the total means. Using the relation (11), the normal system of equations (23) for linear regression is represented as follows:

$$\left(S_{pool} + \sum_{t=1}^{T} n_t (m^{(t)} - m)(m^{(t)} - m)'\right) a = X'y\,. \qquad (25)$$

where the pooled cross-product is defined due to (10)-(11) as:

$$S_{pool} = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (x_{ji}^{(t)} - m_j^{(t)})(x_{ki}^{(t)} - m_k^{(t)}) \equiv \sum_{t=1}^{T} S_t \,, \tag{26}$$

where S_t are the matrices of second moments within each t-th segment. Introducing the constants

$$c_t = (m^{(t)} - m)' a \,, \tag{27}$$

defined similarly to those in derivation (17)-(18), reducing the system (25) to:

$$S_{pool}\, a = X'y - \sum_{t=1}^{T} n_t c_t (m^{(t)} - m) \,. \tag{28}$$

Then solution of (28) is:

$$\begin{aligned} a &= S_{pool}^{-1} X'y - \sum_{t=1}^{T} n_t c_t\, S_{pool}^{-1} (m^{(t)} - m) \\ &\equiv a_{pool} - \sum_{t=1}^{T} n_t c_t\, a_t \end{aligned} \,. \tag{29}$$

In (29) the notations used are:

$$a_{pool} = S_{pool}^{-1} X'y \,, \quad a_t = S_{pool}^{-1} (m^{(t)} - m) \,, \tag{30}$$

so the vector a_{pool} corresponds to the main part of the total vector in (29) of the regression coefficients defined via the pooled matrix (26), and additional vectors a_t correspond to Fisher discriminators (8) between each t-th particular segment and total data set. Decomposition (29) shows that regression coefficients a consist of the part a_{pool} and a linear aggregate (with weights $n_t c_t$) of Fisher discriminators a_t of the segments versus total data. It is interesting to note that if to increase number of segments up to the number of observations ($T=n$, with only one observation in each segment) then each variable's mean in any segment coincides with the original observation itself, $m_k^{(t)} = x_{ki}^{(t)}$, so $S_{pool} = 0$ in (26). In this case the sum in (25) coincides with the total second-moment matrix, so the regular regression solution can be seen as an aggregate of the discriminators by each observation versus total vector of means.

The obtained decomposition (29) is useful for interpretation, but it still contains the unknown parameters c_t (27) that need to be estimated. First, notice that the Fisher discriminators a_t (30) of each segment versus entire data, are restricted by the relation:

$$\begin{aligned} &\sum_{t=1}^{T} n_t\, a_t = \sum_{t=1}^{T} n_t S_{pool}^{-1} (m^{(t)} - m) \\ &= S_{pool}^{-1} \sum_{t=1}^{T} n_t (m^{(t)} - m) \\ &= S_{pool}^{-1} \left(\sum_{t=1}^{T} n_t m^{(t)} - m \sum_{t=1}^{T} n_t \right) = 0 \end{aligned} \,. \tag{31}$$

Thus, for T segments there are only T-1 independent discriminators.

Consider a simple case of two segments in data. In difference to the described two-group LDA problem (12)-(15) and its relation to the binary linear regression (24), we can have a non-binary output, for instance, a continuous dependent variable. Using the derivation (12)-(15) for the inversion of the matrix of the normal system of equations (25), the solution (29) is obtained for two-segment linear regression in explicit form:

$$\begin{aligned} a &= S_{tot}^{-1} X'y \\ &= \left(S_{pool}^{-1} - \frac{h S_{pool}^{-1} (m^{(1)} - m^{(2)})(m^{(1)} - m^{(2)})' S_{pool}^{-1}}{1 + h (m^{(1)} - m^{(2)})' S_{pool}^{-1} (m^{(1)} - m^{(2)})} \right) X'y \\ &= S_{pool}^{-1} X'y - \left(\frac{h (m^{(1)} - m^{(2)})' S_{pool}^{-1} X'y}{1 + h (m^{(1)} - m^{(2)})' S_{pool}^{-1} (m^{(1)} - m^{(2)})} \right) \\ &\quad \cdot S_{pool}^{-1} (m^{(1)} - m^{(2)}) \,, \end{aligned} \tag{32}$$

where h is the same constant as in (12). It can be seen that the vector of coefficients for two-segment regression, similarly to the general solution (29), equals the main part a_{pool} (30) minus a constant (in the parentheses at the right-hand side (32) multiplied by the discriminator (8).

Another analytical result can be obtained for three segments in data, when a general solution (29) contains two discriminators. For this case we extended the Sherman-Morrison formula (14) to the inversion of a matrix $A + u_1 v_1' + u_2 v_2'$, where *A* is a non-singular matrix and $u_1 v_1' + u_2 v_2'$ are two outer products of vectors. The derivation for the inverted matrix of such a structure is given in the Appendix. In this case, the system (25) can be presented in the notations:

$$A = S_{pool}, \quad u_1 = v_1 = \sqrt{n_1}(m^{(1)} - m), \quad u_2 = v_2 = \sqrt{n_2}(m^{(2)} - m) \qquad (33)$$

Applying the formula (A16) with definitions (33), we obtain solution of the system (25) for three segments. In accordance with the relations (29)-(31), this solution is expressed via the vector a_{pool} and two Fisher discriminators.

In a general case of any number *T* of segments, the parameters c_t in the decomposition (29) can be obtained in the following procedure. Theoretical values of the dependent variable are predicted by the regression model (28) as follows:

$$\tilde{y} = X\,a = XS_{pool}^{-1}\,X'y + \sum_{t=1}^{T-1} c_t [n_t X\,S_{pool}^{-1}(m - m^{(t)})] \equiv \tilde{y}_{pool} + \sum_{t=1}^{T-1} c_t \tilde{y}_t \,, \qquad (34)$$

where a predicted vector $\tilde{y}$ is decomposed to the vector $\tilde{y}_{pool}$ defined via the pooled variance and the items $\tilde{y}_t$ related to the Fisher discriminator functions in the prediction:

$$\tilde{y}_{pool} = XS_{pool}^{-1}\,X'y, \quad \tilde{y}_t = n_t XS_{pool}^{-1}(m - m^{(t)}). \qquad (35)$$

All the vectors in (35) can be found from the data, so using $\tilde{y}$ (34) in the regression (21), the model is reduced to:

$$\Delta y = \sum_{t=1}^{T-1} c_t \tilde{y}_t + \varepsilon, \qquad (36)$$

where $\Delta y = y - \tilde{y}_{pool}$ is a vector of difference between empirical and predicted by pooled variance theoretical values of the dependent variable. The relation (36) is also a model of regression of the dependent variable Δy by the new predictors - the Fisher classifications $\tilde{y}_t$ (35). This regression can be constructed in the Least Squares approach (22)-(24). In difference to the regression (21) by possibly many independent *x* variables, the model (36) contains just a few regressors $\tilde{y}_t$, because a number of segments is usually small.

Regression decomposition (25)-(35) uses the segments within the independent variables, that is expressed in presentation of the total second-moment matrix of *x*-s at the left-hand side (25) via the pooled matrix of *x*-s (26). However, there is also a vector *X'y* of the *x*-s cross-products with the dependent variable *y* at the right-hand side of normal system of equations (25). The decomposition of this vector can also be performed by the relations (10)-(11). Suppose, we use the same segments for all *x*-s and *y* variables, then:

$$X'y \equiv (X'y)_{tot} = (X'y)_{pool} + \sum_{t=1}^{T} n_t (m^{(t)} - m)(\bar{y}^{(t)} - \bar{y}), \qquad (37)$$

where $\bar{y}^{(t)}$ and $\bar{y}$ are the mean values of the dependent variable in each *t*-th segment and the total mean. The elements of the vector $(X'y)_{pool}$ in (37) are defined due to (10)-(11) as:

$$(x_j' y)_{pool} = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (x_{ji}^{(t)} - m_j^{(t)})(y_i^{(t)} - \bar{y}^{(t)}), \qquad (38)$$

where x_j is a column of observations for the *j*-th variable in the *X* matrix. Using the presentation (37)-(38) in place of the vector *X'y* in (29)-(30) yields a more detailed decomposition of the vector a_{pool} by the segments within the dependent variable data. In the other relations (32), or (34)-(35), this further decomposition can be used as

well. In a more general case we can consider different segments for the independent variables and for the dependent variable *y*.

If *y* is an ordinal variable, and the segments are chosen by its levels, then within each segment there are zero equaled deviations $y_i^{(t)} - \bar{y}^{(t)} = 0$. Thus, in (38) the values $(x_j' y)_{pool} = 0$, and the decomposition (37) does not contain the pooled vector $(X'y)_{pool}$. Solution (29) can then be given as:

$$a = S_{pool}^{-1} \begin{pmatrix} \sum_{t=1}^{T} n_t (m^{(t)} - m)(\bar{y}^{(t)} - \bar{y}) \\ -\sum_{t=1}^{T} n_t c_t (m^{(t)} - m) \end{pmatrix} = \sum_{t=1}^{T} \gamma_t a_t , \tag{39}$$

where the vectors by segments and the constants are defined as:

$$a_t = S_{pool}^{-1}(m^{(t)} - m) , \quad \gamma_t = n_t(\bar{y}^{(t)} - \bar{y} - c_t) . \tag{40}$$

Thus, the solution (29)-(30) is in this case reduced to the linear combination of discriminant functions a_t with the weights γ_t, without the a_{pool} input. This solution corresponds to the classification (18) by several groups in discriminant analysis. The parameters γ_t can be estimated as it is described in the procedure (32)-(36). If we work with a centered data, a vector of total means by *x*-variables $m = 0$ and the mean value $\bar{y} = 0$, so these items can be omitted in all the formulae.

A useful property of the solution (30) consists in the inversion of the pooled matrix S_{pool} instead of inversion of the total matrix $S_{tot} = X'X$ as in (24). If the independent variables are multicollinear, their covariance or correlation matrix is ill-conditioned or close to a singular matrix. The condition number, defined as ratio between the biggest and the smallest eigenvalues, is large for the ill-conditioned matrices and even infinite for a singular matrix. For such a total matrix *X'X* there could be a problem with its inversion. At the same time the pooled matrix obtained as a sum of segmented matrices (26), is usually less ill-conditioned. The numerical simulations showed that the condition numbers of the pooled matrices are regularly many times less than these values of the related total second-moment matrices. It means that working with a pooled matrix in (30) yields more robust results, not as prone to multicollinearity effects as in a regular regression approach.

Numerical example

Consider an example from a real research project with 550 observations, where the dependent variable is customer overall satisfaction with a bank merchant's services, and the independent variables are: x_1 – satisfaction with the account set up; x_2 – satisfaction with communication; x_3 – satisfaction with how sales representatives answer questions; x_4 – satisfaction with information needed for account application; x_5 – satisfaction with the account features; x_6 – satisfaction with rates and fees; x_7 – satisfaction with time to deposit into account. All variables are measured with a ten-point scale from absolutely non-satisfied to absolutely satisfied (1 to 10 values). The pair correlations of all variables are positive. The data is considered in three segments of non-satisfied, neutral, and definitely satisfied customers, where the segments correspond to the values of the dependent variable from 1 to 5, from 6 to 9, and 10, respectively.

Consider the segments' contribution into the regression coefficients and into the total model quality. The coefficients of regression for the standardized variables are presented in the last column of Table 1.

The coefficient of multiple determination for this model is R^2=0.485, and *F*-statistics equals 73.3, so the quality of the regression is good. The first four columns in Table 1 present inputs to the coefficients of regression from the pooled variance of the independent variables combined with the pooled variance of the dependent variable and three segments (37)-(38). The sum of these items in the next column comprises the pooled subtotal a_{pool} (30).

Table 1. Regression Decomposition by the Items of Pooled Variance and Discriminators.

	Pooled Variance of Predictors				Fisher Discriminators			Regression
Variable	Pooled Dependent	Segment 1	Segment 2	Segment 3	Pooled Subtotal	Segment 1	Segment 3	Total
x_1	.116	.026	.015	.064	.222	-.011	-.044	.166
x_2	.007	.149	.001	.049	.206	-.064	-.034	.108
x_3	.008	.232	-.006	.048	.282	-.100	-.033	.149
x_4	-.035	.005	.021	.077	.068	-.002	-.053	.013
x_5	.039	.101	-.016	-.028	.096	-.044	.019	.072
x_6	.054	.325	.012	.142	.533	-.141	-.098	.294
x_7	.048	.102	.018	.095	.262	-.044	-.065	.153

Table 2. Regression Decomposition by Segments.

	Core Input		Segment 1		Segment3		Regression Total	
Variable	Coefficient	Net Effect	Coefficient	Net Effect	Coefficient	Net Effect	Coefficient	Net Effect
x_1	.131	.072	.015	.008	.020	.011	.166	.091
x_2	.008	.005	.084	.046	.015	.008	.108	.059
x_3	.003	.001	.131	.069	.015	.008	.149	.078
x_4	-.014	-.006	.003	.001	.024	.011	.013	.006
x_5	.023	.008	.057	.020	-.009	-.003	.072	.025
x_6	.066	.037	.184	.103	.044	.025	.294	.165
x_7	.065	.026	.058	.023	.030	.012	.153	.061
R^2	.143		.271		.071		.485	
R^2 share	29%		56%		15%		100%	

The next two columns present the Fisher discriminators (30) for the first and the third segments. It is interesting to note that the condition numbers of the predictors total and pooled matrices of second moments equal 19.7 and 11.9, so the latter one is much less ill-conditioned. Adding the pooled subtotal a_{pool} and Fisher discriminators yields the total coefficients of regression in the last column of Table 1.

Combining some columns of the first table, Table 2 of the main contributions to the coefficients of regression is obtained. Table 2 consists of doubled columns containing coefficients of regression and the corresponded net effects. In Table 2, the core input coefficients equal the sum of pooled dependent and the segment-2 columns from Table 1. Segment-1 coefficients in Table 2 equal the sum of two columns related to Segment-1 from Table 1, and similarly for the Segment-3 coefficients.

Summing all three of these columns of coefficients in Table 2 yields the total coefficients of regression. Considering coefficients in the columns of Table 2 in a way similar to factor loadings in factor analysis, we can identify which variables are more important in each segment of the total coefficients of regression. For instance, comparing coefficients in each row across three first columns in Table 2, we see that the variables x_1 and x_7 have the bigger values in the core input than in segments, satisfaction with account set up and with time to deposit into account play a basic role in the customer overall satisfaction.

Segment-1 has bigger coefficients by the variables x_2, x_3, x_5, and x_6, and the Segment-3 has a bigger coefficient by the variable x_4, so the corresponded attributes play the major roles in creating customers dissatisfaction or delight, respectively. It is interesting to note that this approach produces similar results to another technique developed specifically for the customer satisfaction studies (Conklin, Powaga & Lipovetsky, 2004).

Besides the coefficients of regression, Table 2 presents the net effects, or the characteristics of comparative influence of the regressors in the model (for more on this topic, see Lipovetsky &Conklin, 2001). Quality of regression can be estimated by the coefficient of multiple determination defined by the scalar product of the standardized coefficients of regression a_j and the vector of pair correlations r_{yj} of the dependent variable and each j-th independent variables, so $r_{yj}=(X'y)_j$. Items $r_{yj}a_j$ in total R^2 are called the net effects of each predictor: $R^2 = r_{y1}a_1 + r_{y2}a_2 + ...r_{yn}a_n$. The net effects for core, two segment items, and their total (that is equal to the net effects obtained by the total coefficients of regression) are shown in Table 2.

The net effects can be also used for finding the important predictors in each component of total regression. Summing net effects within their columns in Table 2 yields a splitting of total R^2 =.485 into its core (R^2 =.143), segment-1 (R^2 =.271), and segment-3 (R^2 =.071) components. In the last row of Table 2 we see that the core and two segments contribute to total coefficient of multiple determination by 29%, 56%, and 15%, respectively. Thus, the main share in the regression is produced by segment-1 of the dissatisfaction influence.

Conclusion

Relations between linear discriminant analysis and multiple regression modeling were considered using decomposition of total matrix of second moments of predictors into pooled matrix and outer products of the vectors of segment means. It was demonstrated that regression coefficients can be presented as an aggregate of several items related to the pooled segments and Fisher discriminators. The relations between regression and discriminant analyses demonstrate how a total regression model is composed of the regressions by the segments with possible opposite directions of the dependency on the predictors. Using the suggested approach can provide a better understanding of regression properties and help to find an adequate interpretation of regression results.

References

Anderson T. W. (1958) *An introduction to multivariate statistical analysis*. New York: Wiley and Sons.

Arminger, G., Clogg C. C., & Sobel M. E., (Eds.) (1995). *Handbook of statistical modeling for the social and behavioral sciences.* London: Plenum Press.

Bartlett M. S. (1938). Farther aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society*, 34, 33-40.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67, 364-366.

Conklin, M., Powaga, K., & Lipovetsky, S. (2004). Customer satisfaction analysis: identification of key drivers. *European Journal of Operational Research, 154/3*, 819-827.

Dillon, W. R., & Goldstein, M. (1984) . *Multivariate analysis: methods and applications.* New York: Wiley and Sons.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Good, I. J., & Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics, 15*, 694-711.

Hand, D. J. (1982). *Kernel discriminant analysis*. New York: Research Studies Press.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association, 89*, 1255-1270.

Harville, D. A. (1997) *Matrix algebra from a statistician' perspective*. New York: Springer.

Holland, P. W., & Rubin D. B. (1983). On Lord's paradox. In: H. Wainer & S. Messick (Eds.) *Principals of Modern Psychological Measurement*, 3-25. Hillsdale, NJ: Lawrence Earlbaum.

Hora, S. C., & Wilcox, J. B. (1982). Estimation of error rates in several-population discriminant analysis. *Journal of Marketing Research, 19*, 57-61.

Huberty, C. H. (1994). *Applied discriminant analysis.* New York: Wiley and Sons.

Kendall, M. G. & Stuart, A. (1966). *The advanced theory of statistics*, vol. III. London: Griffin.

Lachenbruch, P. A. (1979). Discriminant analysis. *Biometrics, 35*, 69-85.

Ladd, J. W. (1966) Linear probability functions and discriminant functions. *Econometrica*, 34, 873-885.

Lipovetsky, S., & Conklin M. (2001). Analysis of regression in game theory approach,. *Applied Stochastic Models in Business and Industry, 17*, 319-330.

Lipovetsky, S., & Conklin, M. (2004). Decision making by variable contribution in discriminant, logit, and regression analyses. *Information Technology and Decision Making, 3*, 265-279.

Lipovetsky, S., Tishler A., & Conklin. M. (2002). Multivariate least squares and its relation to other multivariate techniques. *Applied Stochastic Models in Business and Industry, 18*, 347-356.

McLachlan, G. J. (1992) *Discriminant analysis and statistical pattern recognition.* New York: Wiley and Sons.

Morrison, D. G. (1974). Discriminant analysis. In Ferber R. (ed.) *Handbook on marketing research*. New York: McGraw-Hill.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley and Sons.

Rinott, Y., & Tam, M. (2003). Monotone regrouping, regression, and Simpson's paradox. *The American Statistician, 57*, 139-141.

Ripley, B. D. (1996). *Pattern recognition and neural networks.* London: Cambridge University Press.

Rosenbaum, P. R. (1995) *Observational studies.* New York: Springer-Verlag.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. New York: Chapman & Hall.

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data. *The American Statistician, 58*, 117-123.

Winship, C., & Morgan, L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-707.

Appendix:

The Sherman-Morrison formula

$$(A+uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+u'A^{-1}v} \qquad \text{(A1)}$$

is well known in various theoretical and practical statistical evaluations. It is convenient to use when the inverted matrix A^{-1} is already known, so the inversion of $A+uv'$ can be expressed via A^{-1} due to the formula (A1).

We extend this formula to the inversion of a matrix with two pairs of vectors. Consider a matrix $A+u_1v_1'+u_2v_2'$, where A is a square non-singular matrix of n-th order, and $u_1v_1'+u_2v_2'$ is a matrix of the rank 2, arranged via two outer products u_1v_1' and u_2v_2' of the vectors of n-th order. Suppose we need to invert such a matrix to solve a linear system:

$$(A+u_1v_1'+u_2v_2')a = b, \qquad \text{(A2)}$$

where a is a vector of unknown coefficients and b is a given vector. Opening the parentheses, we get an expression:

$$Aa + u_1k_1 + u_2k_2 = b, \qquad \text{(A3)}$$

where k_1 and k_2 are unknown parameters defined as scalar products of the vectors:

$$k_1 = (v_1'a), \quad k_2 = (v_2'a), \qquad \text{(A4)}$$

Solution a can be found from (A3) as:

$$a = A^{-1}b - k_1A^{-1}u_1 - k_2A^{-1}u_2. \qquad \text{(A5)}$$

Substituting the solution (A5) into the system (A2) and opening the parentheses yields a vector equation:

$$\begin{aligned} &k_1u_1 + k_2u_2 + k_1q_{11}u_1 + k_2q_{12}u_1 \\ &+k_1q_{21}u_2 + k_2q_{22}u_2 = c_1u_1 + c_2u_2 \end{aligned}, \qquad \text{(A6)}$$

where the following notations are used for the known constants defined by the bilinear forms:

$$\begin{aligned} q_{11} &= v_1'A^{-1}u_1, \quad q_{12} = v_1'A^{-1}u_2, \\ q_{21} &= v_2'A^{-1}u_1, \quad q_{22} = v_2'A^{-1}u_2, \\ c_1 &= v_1'A^{-1}b, \quad c_2 = v_2'A^{-1}b. \end{aligned} \qquad \text{(A7)}$$

Considering equations (A6) by the elements of vector u_1 and by the elements of vector u_2, we obtain a system with two unknown parameters k_1 and k_2:

$$\begin{cases} (1+q_{11})k_1 + q_{12}k_2 = c_1 \\ q_{21}k_1 + (1+q_{22})k_2 = c_2 \end{cases}. \qquad \text{(A8)}$$

So the solution for the parameters (A4) is:

$$\begin{aligned} k_1 &= (c_1 + q_{22}c_1 - q_{12}c_2)/\Delta, \\ k_2 &= (c_2 + q_{11}c_2 - q_{21}c_1)/\Delta, \end{aligned} \qquad \text{(A9)}$$

with the main determinant of the system:

$$\begin{aligned} \Delta &= (1+q_{11})(1+q_{22}) - q_{12}q_{21} \\ &= (1+v_1'A^{-1}u_1)(1+v_2'A^{-1}u_2) - (v_1'A^{-1}u_2)(v_2'A^{-1}u_1) \end{aligned}. \qquad \text{(A10)}$$

Using the obtained parameters (A9) in the vector a (A5), we get:

$$a = \left\{ A^{-1} - \frac{\begin{array}{c} A^{-1}u_1v_1'A^{-1}(1+q_{22}) + A^{-1}u_2v_2'A^{-1}(1+q_{11}) \\ -A^{-1}u_1v_2'A^{-1}q_{12} - A^{-1}u_2v_1'A^{-1}q_{21} \end{array}}{\Delta} \right\} b \qquad \text{(A11)}$$

with the constants defined in (A7).

The expression in the figure parentheses (A11) defines the inverted matrix of the system (A2). It can be easily proved by multiplying the matrix in (A2) by the matrix in (A11), that yields the uniform matrix. In a simple case when both pairs of the vectors are equal, or $u_1v_1' = u_2v_2'$, they can be denoted as $u_1v_1' = u_2v_2' = 0.5uv'$, and the expression (A12) reduces to the formula (A1). We can explicitly present the inverted matrix (A11) as follows:

$$(A+u_1v_1'+u_2v_2')^{-1} = A^{-1} - \frac{A^{-1}u_1v_1'A^{-1} + A^{-1}u_2v_2'A^{-1}}{\Delta} + \frac{\begin{pmatrix} A^{-1}u_1v_1'A^{-1}u_2v_2'A^{-1} + A^{-1}u_2v_2'A^{-1}u_1v_1'A^{-1} \\ -A^{-1}u_1v_2'A^{-1}u_2v_1'A^{-1} - A^{-1}u_2v_1'A^{-1}u_1v_2'A^{-1} \end{pmatrix}}{\Delta}. \quad \text{(A12)}$$

For the important case of a symmetric matrix A, each of the bilinear forms (A7) can be equally presented by the transposed expression, for instance,

$$\begin{aligned} q_{11} &= v_1'A^{-1}u_1 = u_1'A^{-1}v_1, \\ q_{12} &= v_1'A^{-1}u_2 = u_2'A^{-1}v_1, \\ q_{21} &= v_2'A^{-1}u_1 = u_1'A^{-1}v_2, \\ q_{22} &= v_2'A^{-1}u_2 = u_2'A^{-1}v_2. \end{aligned} \quad \text{(A13)}$$

Using the property (A13) we simplify the numerator of the second ratio in (A12) to following:

$$\begin{aligned} &A^{-1}u_1u_2'A^{-1}v_1v_2'A^{-1} + A^{-1}u_2u_1'A^{-1}v_2v_1'A^{-1} \\ &-A^{-1}u_1u_2'A^{-1}v_2v_1'A^{-1} - A^{-1}u_2u_1'A^{-1}v_1v_2'A^{-1} \\ &= A^{-1}(u_1u_2' - u_2u_1')A^{-1}(v_1v_2' - v_2v_1')A^{-1}. \end{aligned} \quad \text{(A14)}$$

So the formula (A12) for a symmetric matrix A can be represented as:

$$(A+u_1v_1'+u_2v_2')^{-1} = A^{-1} - \frac{\begin{pmatrix} A^{-1}(u_1v_1'+u_2v_2')A^{-1} \\ -A^{-1}(u_1u_2'-u_2u_1')A^{-1}(v_1v_2'-v_2v_1')A^{-1} \end{pmatrix}}{\Delta}, \quad \text{(A15)}$$

with the determinant defined in (A10).

In a special case of the outer products of each vector by itself, when $u_1 = v_1$ and $u_2 = v_2$, the formula (A15) transforms into:

$$(A+u_1u_1'+u_2u_2')^{-1} = A^{-1} - \frac{\begin{pmatrix} A^{-1}(u_1u_1'+u_2u_2')A^{-1} \\ -A^{-1}(u_1u_2'-u_2u_1')A^{-1}(u_1u_2'-u_2u_1')A^{-1} \end{pmatrix}}{(1+u_1'A^{-1}u_1)(1+u_2'A^{-1}u_2) - (u_1'A^{-1}u_2)^2}. \quad \text{(A16)}$$

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 75-80

1538 – 9472/05/$95.00

Local Power For Combining Independent Tests in The Presence of Nuisance Parameters For The Logistic Distribution

W. A. Abu-Dayyeh Z. R. Al-Rawi M. MA. Al-Momani
Department of Statistics, Faculty of Science
Yarmouk University Irbid-Jordan

Four combination methods of independent tests for testing a simple hypothesis versus one-sided alternative are considered viz. Fisher, the logistic, the sum of P-values and the inverse normal method in case of logistic distribution. These methods are compared via local power in the presence of nuisance parameters for some values of α using simple random sample.

Key words: combination method; independent tests; logistic distribution; local power; simple random sample; nuisance parameter.

Introduction

Combining independent tests of hypotheses is an important and popular statistical practice. Usually, data about a certain phenomena comes from different sources in different times, so we want to combine these data to study such phenomena. Many authors have considered the problem of combining (n) independent tests of hypotheses. For simple null hypotheses, Little and Folks (1971), studied four methods for combining a finite number of independent tests. They found that the Fisher method is better than the other three methods via Bahadur efficiency. Again, Little and Folks (1973) studied all methods of combining a finite number of independent tests and thy found that the Fisher's method is optimal under some mild conditions.

W. A. Abu-Dayyeh, Department of Mathematical Sciences, Dhahran, Saudi Arabia M. MA. Al-Momani, Department of Mathematical Sciences, Dhahran, Saudi Arabia Z. R. Al-Rawi Chairman, Department of Statistics Yarmouk University, Irbid, Jordan For correspondence regarding this article, send E-mail to alrawiz@yu.edu.jo. This work was carried out with financial support from the Yarmouk University Research Council.

Brown, Cohen and Strawderman (1976) have shown that such all tests form a complete class. Abu-Dayyeh and Bataineh (1992) showed that the Fisher's method is strictly dominated by the sum of P-values method via Exact Bahadur Slop in case of combining an infinite number of independent shifted exponential tests when the sample size remains finite. Also, Abu-Dayyeh (1992) showed that under certain conditions that the local limit of the ratio of the Exact Bahadur efficiency of two tests equivalent to the Pitman efficiency between the two tests where these tests are based on sum of iid *r.v's*. Again Abu-Dayyeh and El-Masri (1994) studied the problem of combining (n) independent tests as $(n \to \infty)$ in case of triangular distribution using six methods viz. sum of P-values, inverse normal, logistic, Fisher, minimum of P-values and maximum of P-values. They showed that the sum of P-values is better than all other methods.

Abu-Dayyeh (1997) extended the definition of the local power of tests to the case of having nuisance parameters. He derived the local power for any symmetric test in the case of a bivariate normal distribution with known correlation coefficient, and then he applied it to the combination methods.

Specific Problem

Suppose there is (n) simple hypotheses:

$$H_0^{(i)} : \theta_i = \theta_{0i} \quad \text{vs} \quad H_1^{(i)} : \theta_i > \theta_{0i} \quad i=1,2,\ldots,n \tag{1}$$

Where θ_{0i} is known for $i=1,2,\ldots,n$ and $H_0^{(i)}$ is rejected for sufficiently large values of some continuous real valued test statistic $T^{(i)}$, $i=1,2,\ldots,n$ and we want to combine the (n) hypotheses into one hypothesis as follows:

$$H_0: (\theta_1, \theta_2, \ldots, \theta_n) = (\theta_{01}, \theta_{02}, \ldots, \theta_{0n})$$

vs

$$H_1: \theta_i \geq \theta_{0i} \text{ for all i, and } \theta_i > \theta_{0i} \text{ for some } i,\ i=1,2,\ldots,n \quad (2)$$

Many methods have been used for combining several tests of hypotheses into one overall test. Among these methods are the non-parametric (omnibus) methods that combine the P-values of the different tests. The P-value of the *i-th* hypothesis is given by:

$$P_i = \underset{H_0^{(i)}}{P}(T^{(i)} \geq t) = 1 - \underset{H_0^{(i)}}{F}(t) \quad (3)$$

where $F_{H0}^{(i)}(t)$ is the *cdf* of $T^{(i)}$ under $H_0^{(i)}$. Note that $P_i \sim U(0,1)$ under $H_0^{(i)}$.

Considered in this article is the case of $\theta_i^* = \gamma\theta_i$, where $\theta_1, \theta_2, \ldots, \theta_r \geq 0$ fixed constants and γ is the unknown parameter. Then $T^{(1)}, T^{(2)}, \ldots, T^{(r)}$ are independent *r.v's* such that for $i = 1,2,\ldots,r$ and we want to test

$$H_0 : \gamma = 0 \quad \text{vs}$$
$$H_1 : \gamma > 0 \ ` \quad (4)$$

and therefore considered is the problem of combining a finite number of independent tests by looking at the Local Power of tests which is defined for a test φ by:

$$L_P(\varphi) = \inf_{\theta} \frac{\partial}{\partial\gamma} E_{\gamma\theta}(\varphi) \Big|_{\gamma=0} \quad (5)$$

where

$\gamma \geq 0, \theta = (\theta_1, \theta_2, \ldots, \theta_r), \theta_i \geq 0, i = 1,2,\ldots,r$, in case of logistic distribution. Compared (5) for the four methods of combining tests for the location family of distributions when $r = 2$ and $r = 3$. These methods are: Fisher, logistic, the sum of p-values and the inverse normal methods.

Methodology

Now we will find expressions for the Local Power of the four combination methods of tests then compare them via the Local Power.

Lemma 1

Let X_1, X_2 be independent *r.v's* such that $X_i \sim Logistic(\gamma\theta_i, 1)$ for $i = 1,2$. Then

A(1)

$$\frac{\partial}{\partial\gamma} E_{\gamma(\theta_1,\theta_2)}(\varphi_F) \Big|_{\gamma=0} = K_F(\theta_1 + \theta_2), \text{where}$$

$$K_F = \int_1^a \left(1 - e^{-c/2} y\right) \frac{2-y}{y^3} dy, \quad a = e^{c/2} \quad \text{and}$$

$$c = \chi^2_{(4),(1-\alpha)}$$

A(2)

$$\frac{\partial}{\partial\gamma} E_{\gamma(\theta_1,\theta_2)}(\varphi_L) \Big|_{y=0} = K_L(\theta_1 + \theta_2), \quad \text{where}$$

$$K_L = \int_1^{\infty} \frac{(y-2)(y-1)}{(y-1+e^{-c})y^3} dy, \text{ and } c \text{ satisfies the following } 1-\alpha = \frac{1-e^{-c}(c+1)}{(1-e^{-c})^2}.$$

A(3)

$$\frac{\partial}{\partial\gamma} E_{\gamma(\theta_1,\theta_2)}(\varphi_S) \Big|_{\gamma=0} = K_S(\theta_1 + \theta_2), \text{where}$$

$$K_S = \frac{c^2(3-2c)}{6}, \text{ and } c = \sqrt{2\alpha}.$$

A(4)

$$\frac{\partial}{\partial\gamma}\mathrm{E}_{\gamma(\theta_1,\theta_2)}(\varphi_N)\mid_{\gamma=0}=\mathrm{K}_N(\theta_1+\theta_2),\text{where}$$

$$K_N=-\int_1^a\left(1-\Phi\left(-c-\Phi^{-1}\left(\frac{1}{y}\right)\right)\right)\frac{y-2}{y^3}dy,$$

$$a=\frac{1}{\Phi(-c)}\text{ and }c=\sqrt{2}\,\Phi^{-1}(1-\alpha).$$

Proofs of the previous lemma are similar to proofs of lemma 2, so we will not write it.

Lemma 2

Let X_1,X_2,X_3 be independent *r.v's* such that $X_i\sim Logistic(\gamma\theta_i,1)$ for $i=1,2,3$. Then

B(1) $$\frac{\partial}{\partial\gamma}E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_F)\mid_{\gamma=0}=K_F\sum_{i=1}^{3}\theta_i,$$
$$=K_F(\theta_1+\theta_2+\theta_3)\qquad\text{where}$$

$$K_F=\int_1^a\left[1-e^{-c/2}\,y\left(1+\frac{c}{2}-\ln(y)\right)\right]\frac{y-2}{y^3}dy$$

, $a=e^{c/2}$, and $c=\chi^2_{(6),(1-\alpha)}$.

`B(2)

, where

$$K_L=-\int_1^\infty\int_1^\infty\frac{(u-1)(v-1)(2-v)}{(u-1)(v-1)+e^{-c}}\frac{1}{u^2}\frac{1}{v^3}du\,dv,$$

and *c* satisfies the following:

$$1-\alpha=\int_1^\infty\int_1^\infty\frac{(u-1)(v-1)}{(u-1)(v-1)+e^{-c}}\frac{1}{u^2}\frac{1}{v^2}du\,dv$$

B(3)

$$\frac{\partial}{\partial\gamma}E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_S)\mid_{\gamma=0},$$
$$=K_S\sum_{i=1}^{3}\theta_i=K_S(\theta_1+\theta_2+\theta_3)\text{ where}$$

$$K_S=\frac{c^3(2-c)}{12}\text{ and }c=\sqrt[3]{6\alpha}\ .$$

B(4)

$$\frac{\partial}{\partial\gamma}E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_N)\mid_{\gamma=0}=K_N\sum_{i=1}^{3}\theta_i,$$
$$=K_N(\theta_1+\theta_2+\theta_3)$$

where

$$K_N=$$
$$-\int_1^a\int_1^b\left(1-\Phi\left(-c-\Phi^{-1}(u)-\Phi^{-1}(v)\right)\right)(1-2v)\,du\,dv$$

$a=\Phi(-c)$, $b=\Phi\left(-c-\Phi^{-1}(v)\right)$, and $c=\sqrt{3}\,\Phi^{-1}(1-\alpha)$.

Now, we will prove just B(1), because the proof of the others can be done in the same way.

Proof of B(1):

$$E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_F)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi_F\prod_{i=1}^{3}f(x_i-\gamma\theta_i)\,dx_i,$$

where $f(x_i-\gamma\theta_i)$ is the $p.d.f$ of $Logistic(\gamma\theta_i,1)$ $for\,i=1,2,3$

It easy to show that: $E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_F)$

$$=1-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(1-\phi_F)\prod_{i=1}^{3}f(x_i-\gamma\theta_i)\,dx_i$$

so,

$$\frac{\partial}{\partial\gamma}E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_F)$$

$$=\frac{\partial}{\partial\gamma}\left[1-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(1-\phi_F)\prod_{i=1}^{3}f(x_i-\gamma\theta_i)\,dx_i\right]$$

$$\frac{\partial}{\partial\gamma}E_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_F)\;\Big|_{\gamma=0}=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(1-\varphi_F)\times$$

$$\left\{\begin{array}{l} f(x_1)\,f(x_2)\,f'(x_3)+f(x_1)\,f'(x_2)\,f(x_3)\\ +f'(x_1)\,f(x_2)\,f(x_3)\end{array}\right\}$$

$$dx_1dx_2dx_3$$

such that when

$$\gamma=0\ ,\ f'(x_i)=\frac{\theta_i\left(e^{-x_i}-e^{-2x_i}\right)}{\left(1+e^{-x_i}\right)^3}\ \ for\ i=1,2,3.$$

By symmetric of x_i we have

$$\frac{\partial}{\partial\gamma}\mathrm{E}_{\gamma(\theta_1,\theta_2,\theta_3)}(\varphi_\mathrm{F})\;\Big|_{\gamma=0}=\left(\sum_{\mathrm{i}=1}^{3}\theta_\mathrm{i}\right)\mathrm{K_F}\ ,\ \text{where}$$

$K_F=$

$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(1-\varphi_F)\frac{e^{-x_1}}{\left(1+e^{-x_1}\right)^2}\frac{e^{-x_2}}{\left(1+e^{-x_2}\right)^2}\frac{\left(e^{-x_3}-e^{-2x_3}\right)}{\left(1+e^{-x_3}\right)^3}dx_1dx_2dx_3$$

where $1-\varphi_F=\begin{cases}1, & -2\sum_{i=1}^{3}\ln(p_i)\le c\\ 0, & o.w\end{cases}$,

$$p_i=\frac{1}{1+e^{x_i}},\ i=1,2,3$$

$-2\ln(p_1)-2\ln(p_2)-2\ln(p_3)\le c$ implies

that $x_1\le\ln\left(\frac{e^{c/2}}{\left(e^{x_2}+1\right)\left(e^{x_3}+1\right)}-1\right)$,

also

$-2\ln(p_2)-2\ln(p_3)\le c$ and

$-2\ln(p_3)\le c$ implies

that $x_2\le\ln\left(\frac{e^{c/2}}{\left(e^{x_3}+1\right)}-1\right)$

and $x_3\le\ln\left(e^{c/2}-1\right)$ respectively.

Let $a=\ln\left(e^{c/2}-1\right)$,

$b=\ln\left(\frac{e^{c/2}}{\left(e^{x_3}+1\right)}-1\right)$ and let

$d=\ln\left(\frac{e^{c/2}}{\left(e^{x_2}+1\right)\left(e^{x_3}+1\right)}-1\right)$, then we will get

$K_F=$

$$-\int_{-\infty}^{a}\int_{-\infty}^{b}\int_{-\infty}^{d}\frac{e^{-x_1}}{\left(1+e^{-x_1}\right)^2}\frac{e^{-x_2}}{\left(1+e^{-x_2}\right)^2}\frac{\left(e^{-x_3}-e^{-2x_3}\right)}{\left(1+e^{-x_3}\right)^3}dx_1dx_2dx_3\ .$$

Let $I_1=\int_{-\infty}^{d}\frac{e^{-x_1}}{\left(1+e^{-x_1}\right)^2}dx_1$, then put

$u=1+e^{-x_1}$ to get that $I_1=\frac{1}{1+e^{-d}}$,

so, $I_1=1-e^{-c/2}\left(e^{x_2}+1\right)\left(e^{x_3}+1\right)$.

$\therefore K_F=$

$$-\int_{-\infty}^{a}\int_{-\infty}^{b}\frac{e^{-x_2}}{\left(1+e^{-x_2}\right)^2}\frac{\left(e^{-x_3}-e^{-2x_3}\right)}{\left(1+e^{-x_3}\right)^3}\left(1-e^{-c/2}\left(e^{x_2}+1\right)\left(e^{x_3}+1\right)\right)dx_1dx_2$$

Also, let

$$I_2=\int_{-\infty}^{b}\frac{e^{-x_2}}{\left(1+e^{-x_2}\right)^2}\left(1-e^{-c/2}\left(e^{x_2}+1\right)\left(e^{x_3}+1\right)\right)dx_2$$

$$= \int_{-\infty}^{b} \frac{e^{-x_2}}{\left(1+e^{-x_2}\right)^2} dx_2$$

$$-e^{-c/2}\left(e^{x_3}+1\right)\int_{-\infty}^{b} \frac{1}{\left(1+e^{-x_2}\right)} dx_2$$

$$=1-e^{-c/2}\left(e^{x_3}+1\right)$$

$$-e^{-c/2}\left(e^{x_3}+1\right)\left(\frac{c}{2}-\ln\left(e^{x_3}+1\right)\right)$$

$$=1-e^{-c/2}\left(e^{x_3}+1\right)\left(1+\frac{c}{2}-\ln\left(e^{x_3}+1\right)\right)$$

$$\therefore K_F =$$

$$-\int_{-\infty}^{a} \frac{e^{-x_3}-e^{-2x_3}}{\left(1+e^{-x_3}\right)}\left[1-e^{-c/2}\left(e^{x_3}+1\right)\left(1+\frac{c}{2}-\ln\left(e^{x_3}+1\right)\right)\right] dx_3 .$$

Finally put $y = 1+e^{x_3}$ we get

$$K_F = \int_{1}^{d}\left(1-e^{-c/2}y\left(1+\frac{c}{2}-\ln(y)\right)\right)\frac{y-2}{y^3}dy ,$$

$$d = e^{c/2}$$

$$\alpha = P_0\left(-2\sum_{i=1}^{3}\ln(p_i)\geq c\right) = 1-P_0\left(-2\sum_{i=1}^{3}\ln(p_i)\leq c\right),$$

because $-2\sum_{i=1}^{3}\ln(p_i) \sim \chi^2_{(6)}$

under H_0, then $c = \chi^2_{(6),(1-\alpha)}$, which completes the proof.

Also, here for the logistic distribution we will compare the Local Power for the previous four tests numerically. So from tables (1) and (2) when $\alpha = 0.01$ and $r = 2$ the sum of p-values method is the best method followed by the inverse normal method, the logistic method and Fisher method respectively, but for all of the other values of α and r the inverse normal method is the best method followed by the sum of p-values method followed by logistic method and the worst method is Fisher method.

References

Abu-Dayyeh, W. A. (1989). Bahadur exact slope, pitman efficiency and local power combining independent tests. Ph.D. Thesis, University of Illinois at Urbana – Champaign.

Abu-Dayyeh, W. A. (1992). Exact bahadur efficiency. *Pakistan Journal of Statistics*, *8* (2), 53-61.

Abu-Dayyeh, W. A., & Bataineh (1992). Comparing the exact Bahadur of the Fisher and sum of P-values methods in case of shifted exponential distribution. Mu'tah slopes. *Journal for Research and Studies*, *8*, 119-130.

Abu-Dayyeh, W. A., & El-Masri. (1994). Combining independent tests of triangular distribution. *Statistics & Probability Letters*, *21*, 195-202.

Abu-Dayyeh, W. A. (1997). Local power of tests in the prescience of nuisance parameters with an application. *The Egyptian Statistical Journal ISSR*, *41*, 1-9.

Little, R. C., & Folks, L. J. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, *66*, 802-806.

Little, R. C., & Folks, L. J. (1973). Asymptotic optimality of Fisher's method of combining Independent Tests II. *Journal of the American Statistical Association*, *68*, 193-194.

The following tables explain the term K_A where $A \in \{F, L, S, N\}$ for the logistic distributions.

Table (1): Local power for the logistic distribution when $(r = 2)$

α	K_F	K_L	K_S	K_N
0.010	0.0073833607	0.0081457298	0.0090571910	0.0089064740
0.025	0.0174059352	0.0192749938	0.0212732200	0.0214554551
0.050	0.0326662436	0.0361783939	0.0394590744	0.0415197403

Table (2): Local power for the logistic distribution when $(r = 3)$

α	K_F	K_L	K_S	K_N
0.010	0.0062419188	0.0071070250	0.0080425662	0.0083424342
0.025	0.0144747833	0.0165023359	0.0183583839	0.0199610766
0.050	0.0267771426	0.0304639648	0.0332641762	0.0381565019

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 81-89

1538 – 9472/05/$95.00

Effect Of Position Of An Outlier On The Influence Curve Of The Measures Of Preferred Direction For Circular Data

B. Sango Otieno
Department of Statistics
Grand Valley State University

Christine M. Anderson-Cook
Statistical Sciences Group
Los Alamos National Laboratory

Circular or angular data occur in many fields of applied statistics. A common problem of interest in circular data is estimating a preferred direction and its corresponding distribution. It is complicated by the wrap-around effect on the circle, which exists because there is no natural minimum or maximum. The usual statistics employed for linear data are inappropriate for directional data, as they do not account for its circular nature. The robustness of the three common choices for summarizing the preferred direction (the sample circular mean, sample circular median and a circular analog of the Hodges-Lehmann estimator) are evaluated via their influence functions.

Key words: Circular distribution, directional data, influence function, outlier

Introduction

The notion of preferred direction in circular data is analogous to the center of a distribution for data on a linear scale. Unlike in linear data where a center always exists, if data are uniformly distributed around the circle, then there is no natural preferred direction. Therefore, it is appropriate and desirable that all sensible measures of preferred direction are undefined if the sample data are equally spaced around the circle. This article considers estimating the preferred direction for a sample of unimodal circular data. Three choices for summarizing the preferred direction are the mean direction, the median direction (Fisher 1993) and the Hodges-Lehmann estimate (Otieno & Anderson-Cook, 2003a).

B. Sango Otieno is Assistant Professor at Grand Valley State University. He is a member of Institute of Mathematical Statistics and Michigan Mathematics Teachers Association. Email: otienos@gvsu.edu. Christine Anderson-Cook is a Statistician at Los Alamos National Laboratory. She is a member of the American Statistical Association, and the American Society of Quality. Email: c-and-cook@lanl.gov.

The sample mean direction is a common choice for moderately large samples, because when combined with a measure of sample dispersion, it acts as a summary of the data suitable for comparison and amalgamation with other such information. The sample mean is obtained by treating the data as vectors of length one unit and using the direction of their resultant vector. Given a set of circular observations $\theta_1, \ldots, \theta_n$, each observations is measured as a unit vector with coordinates from the origin of $(\cos(\theta_i), \sin(\theta_i))$, i = 1, . . ., n. The resultant vector of these n unit vectors is obtained by summing them componentwise to get the resultant vector $R = \left(\sum_{i=1}^{n} \cos(\theta_i), \sum_{i=1}^{n} \sin(\theta_i) \right) = (C, S)$, say. The sample circular mean is the angle corresponding to the mean resultant vector $\overline{R} = \frac{R}{n} = \left(\frac{C}{n}, \frac{S}{n} \right) = (\overline{C}, \overline{S})$. That is, the angle corresponding to the mean resultant length $|\overline{R}| = \sqrt{(\overline{C}^2 + \overline{S}^2)}$.

Jamalamadaka and SenGupta (2001), show that the sample circular mean direction is location invariant, that is, if the data are shifted by a certain amount, the value of the sample

circular mean direction also changes by that amount.

An alternative, the sample median, can be thought of as the location of the circumference of the circle that balances the number of observations on the two halves of the circle, Otieno and Anderson-Cook (2003b). The sample median direction $\tilde{\theta}$ of angles $\theta_1, \ldots, \theta_n$, is defined to be the point P on the circumference of the circle that satisfies the following two properties: (a) The diameter PQ through P divides the circle into two semi-circles, each with an equal number of observed data points and, (b) the majority of the observed data are closer to P than to the anti-median Q. See Mardia (1972, p.28-30) and Fisher (1993, p. 35-36).

Note, the antimedian can be thought of as the meeting point of the two tails of the distribution on the opposite side of the circle. Intuitively, fewer observations are expected at the tails. As with the linear case, for odd size samples the median is an observation, while for even sized samples the median is the midpoint of two adjacent observations. Observations directly opposite each other do not contribute to the preferred direction, since in such a case the observations balance each other for all possible choices of medians. The procedure for finding the circular median has the flexibility to find a balancing point for situations involving ties, by mimicking the midranking idea for linear data.

Otieno and Anderson-Cook (2003b) describe a strategy for more efficiently dealing with non-unique circular median estimates especially for small samples, which are commonly encountered in circular data. Note that the angle $\tilde{\theta}$ which has the smallest circular mean deviation given by $d(\tilde{\theta}) = \pi - \frac{1}{n}\sum_{i=1}^{n}\left|\pi - \left|\theta_i - \tilde{\theta}\right|\right|$ is the circular median, Fisher (1993).

A third measure of preferred direction for circular data is the circular Hodges-Lehmann estimate of preferred direction, subsequently referred to as HL. This is the circular median of all pairwise circular means of the data (Otieno & Anderson-Cook, 2003a). As with the linear case, there are three possible methods for calculating this quantity based on which pairs of observations are considered.

The three possible methods involve using the circular means of all distinct pairs of observations, all distinct pairs of observations plus the individual observations (which are essentially pairwise circular means of individual observations with themselves), and all possible pairwise circular means. The estimates obtained by all the three methods, divide the obtained pairwise circular means evenly on the two semicircles. All these estimates of preferred direction are location invariant, since they satisfy the definition of the circular median, which is also location invariant. The approach used is feasible regardless of sample size or the presence of ties. Note that no ranking is used in computing the new measure, since on the circle there is no uniquely defined natural minimum/maximum. Simulation results show that the three HL measures tend towards being asymptotically identical, Otieno (2002), as is the case of linear data, Huber (1981).

Three choices are presented for estimating preferred direction for a single population of circular measures, and study their robustness via their influence curve. As with linear data, where the mean and the median represent different types of centers for data sets, the three estimates of preferred direction also have relative trade-offs for what they are trying to estimate as well as how they deal with lack of symmetry and outliers. The following data set is considered, which give a small overview of the types of data that may be encountered in practice by say, biologists. The data given in Table 1, relates the homing ability of the Northern cricket frog, *Acris crepitans*, as studied by Ferguson, et. al. (1967).

Table 1: Frog Data-Angles in degrees measured due North.

104	110	117	121	127	130	136	145	152
178	184	192	200	316				

Methodology

A circular distribution (CD) is a probability distribution whose total probability is concentrated on the circumference of a unit circle. A set of identically distributed independent random variables from such a distribution is referred to as a random sample from the CD. See Jammalamadaka & SenGupta (2001, p. 25-63) for a detailed discussion of circular probability distributions. Two frequently used families of distributions for circular data include the von Mises and the Uniform distribution.

The von Mises distribution VM (μ, κ), is a symmetric unimodal distribution characterized by a mean direction μ,and concentration parameter κ, with probability density function $f(\theta) = [2\pi I_0(\kappa)]^{-1} \exp[\kappa \cos(\theta - \mu)]$, $0 \le \theta, \mu < 2\pi$ and $0 \le \kappa < \infty$, where $I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa xos(\phi)]d\phi = \sum_{j=0}^{\infty} -\frac{\kappa^{2j}}{4^j j^2}$ is the modified Bessel function of order zero.

The concentration parameter, κ, quantifies the dispersion. If κ is zero, $f(\theta) = \frac{1}{2\pi}$ and the distribution is uniform with no preferred direction. As κ increase from zero, $f(\theta)$ peaks higher about μ. The von Mises is symmetric since it has the property $f(\mu + \theta) = f(\mu - \theta)$, for all θ, where addition or subtraction is modulo 2π. With the uniform or isotropic distribution, however, the total probability is spread out uniformly on the circumference of a circle; that is, all directions are equally likely. It thus represents the state of no preferred direction.

The von Mises is similar in importance to the Normal distribution on the line, (Mardia,1972). When $\kappa \ge 2$, the von Mises distribution VM (μ, κ), can be approximated by the Wrapped Normal distribution WN(μ, ρ), which is a symmetric unimodal distribution obtained by wrapping a normal N(μ, σ^2) distribution around the circle. A circular r.v θ is said to have a wrapped normal (WN) distribution if its pdf is $f_W(\theta) = (2\pi)^{-1} + \pi^{-1} \sum_{p=1}^{\infty} \rho^{p^2} \cos[p(\theta - \mu)]$, $0 \le \mu \le 2\pi$, $0 \le \rho \le 1$, where μ and $\rho = \exp\left(\frac{-1}{2}\sigma^2\right)$ are the mean direction and mean resultant length respectively. The value of $\rho = 0$ corresponds to the circular uniform distribution, and as ρ increases to 1, the distribution concentrates increasingly around μ. Stephens (1963) matched the first trigonometric moments of the von Mises and wrapped normal distributions, that is,

$$\rho = \exp\left(\frac{-1}{2}\sigma^2\right) = A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)},$$

establishing that the two have a close relationship, where

$$I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa \cos(\theta)]d\theta = \sum_{j=0}^{\infty} \frac{1}{(j!)^2} \left(\frac{\kappa^2}{4}\right)^j$$

and

$$I_1(\kappa) = \sum_{r=0}^{\infty} [(r+1)!r!]^{-1} \left(\frac{1}{2}\kappa\right)^{2r+1}$$

are the modified Bessel functions of order zero and order one, respectively. Based on the difficulty in distinguishing the two distributions, Collett and Lewis (1981) concluded that decision on whether to use a von Mises model or a Wrapped Normal model, depends on which of the two is most convenient.

The Wrapped Normal distribution WN (μ, ρ) is obtained by wrapping the N(μ, σ^2) distribution onto the circle, where $\sigma^2 = -2\log\rho$, which implies that, $\rho = \exp\left[\frac{-\sigma^2}{2}\right]$. But for large κ, in particular $\kappa \ge 2$, (Fisher, 1987), VM(μ, κ) is approximately equivalent to N$\left(\mu, \frac{1}{\kappa}\right)$, which in turn is approximately equivalent to

$\mathrm{WN}\left(\mu, \exp\left[\frac{-1}{2\kappa}\right]\right)$. This approximation is very accurate for $\kappa > 10$ (Mardia & Jupp, 2000).

Note $\hat{\sigma}^2 = -2\log A(\kappa)$ and $\hat{\sigma}^2 = \frac{1}{\kappa}$ are the estimates of σ^2 when $\mathrm{VM}(\mu, \kappa)$ is approximated by $\mathrm{WN}(\mu, \rho)$ and $\mathrm{N}\left(\mu, \frac{1}{2\kappa}\right)$ respectively. Figure 1 shows how the WN and N approximations are related for various values of concentration parameter, κ, using the following approximation,

$$A(\kappa) \approx 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} - \ldots,$$

Jammalamadaka & SenGupta (2001, p. 290).

The circular median is rotationally invariant as shown by Ackermann (1997). Lenth (1981), and, Wehrly and Shine (1981) studied the robustness properties of both the circular mean and median using influence curves, and revealed that the circular mean is quite robust, in contrast to the mean for linear data on the real line. Durcharme and Milasevic (1987), show that in the presence of outliers, the circular median is more efficient than the mean direction. Many authors, including He and Simpson (1992), advocate the use of circular median as an estimate of preferred direction, especially in situations where the data are not from the von Mises distribution.

The Hodges-Lehmann estimator, on the other hand is a compromise between the occasionally non-robust circular mean and the more robust circular median. Unlike the circular median which downweights outliers significantly but is sensitive to rounding and grouping (Wehrly & Shine, 1981), the HL estimate downweights outliers more sparingly and is more robust to rounding and grouping. The circular HL estimator has comparable efficiency to mean and is superior to median; see Otieno and Anderson-Cook (2003a). Other properties of this estimate are explored and compared to those of circular mean and circular median in Otieno and Anderson-Cook (2003a). S-Plus or R functions for computing this estimate are available by request from the authors.

Consider a circular distribution F which is unimodal and symmetric about the unknown direction μ_0. The influence function (IF) for the circular mean direction is given by $IF(\theta) = \frac{\sin(\theta - \mu_0)}{\rho}$, where the mean resultant length is given by $\rho = \exp\left(\frac{-1}{2}\sigma^2\right) = A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$. For any given value of ρ, this influence function and its derivative are bounded by $\pm\rho^{-1}$, see Wehrly and Shine (1981). Another result due to Wehrly and Shine (1981) is the influence function of the circular median. Without loss of generality for notational simplicity, assume that $\mu \in [0, \pi]$.

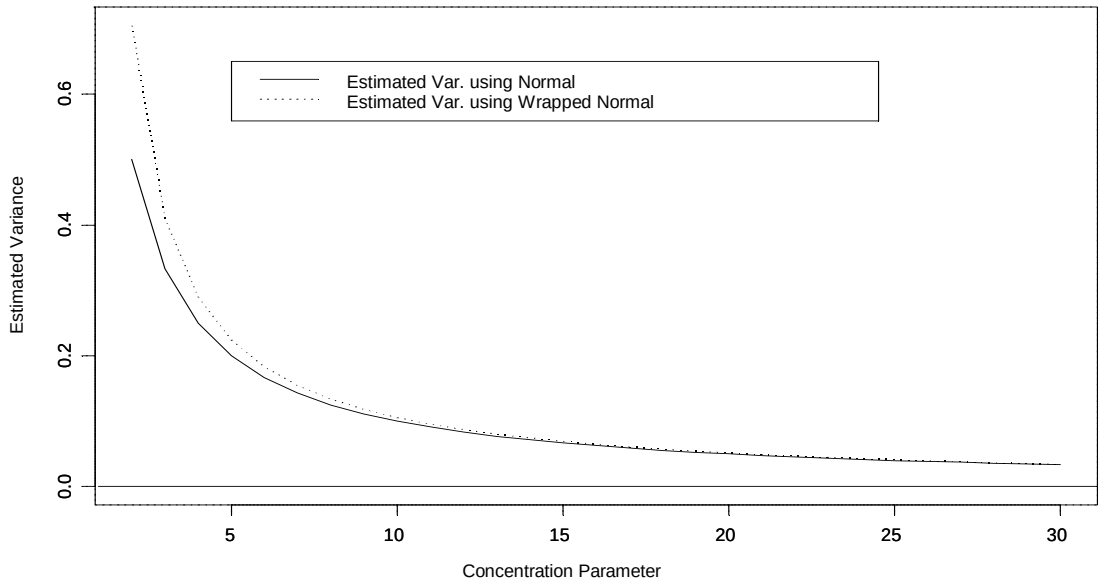
The influence function for the circular median direction is given by

$$IF(\theta) = \frac{\frac{1}{2}\mathrm{sgn}(\theta - \mu_0)}{[f(\mu_0) - f(\mu_0 + \pi)]},$$

$(\mu_0 - \pi < \theta < \mu_0 + \pi)$, where $f(\mu_0)$ is the probability density function of the underlying distribution of the data at the hypothesized mean direction μ_0, and sgn(x) = 1, 0, or -1 as x > 0, x = 0, or x < 0, respectively.

Wehrly and Shine (1981) and Watson (1986) evaluated the robustness of the circular mean via an influence function introduced by Hampel (1968, 1974) and concluded that the estimator is somewhat robust to fixed amounts of contamination and to local shifts, since its influence function is bounded. The influence curve for the circular median, however, has a jump at the antimode. This implies that the circular median is sensitive to rounding or grouping of data (Wehrly & Shine, 1981).

Figure 1: Plot of $\hat{\sigma}^2 = \left[-2\log A\left[\frac{1}{\kappa}\right]\right]$, and $\hat{\sigma}^2 = \left[\frac{1}{\kappa}\right]$ verses Concentration Parameter (κ) for a single observation.



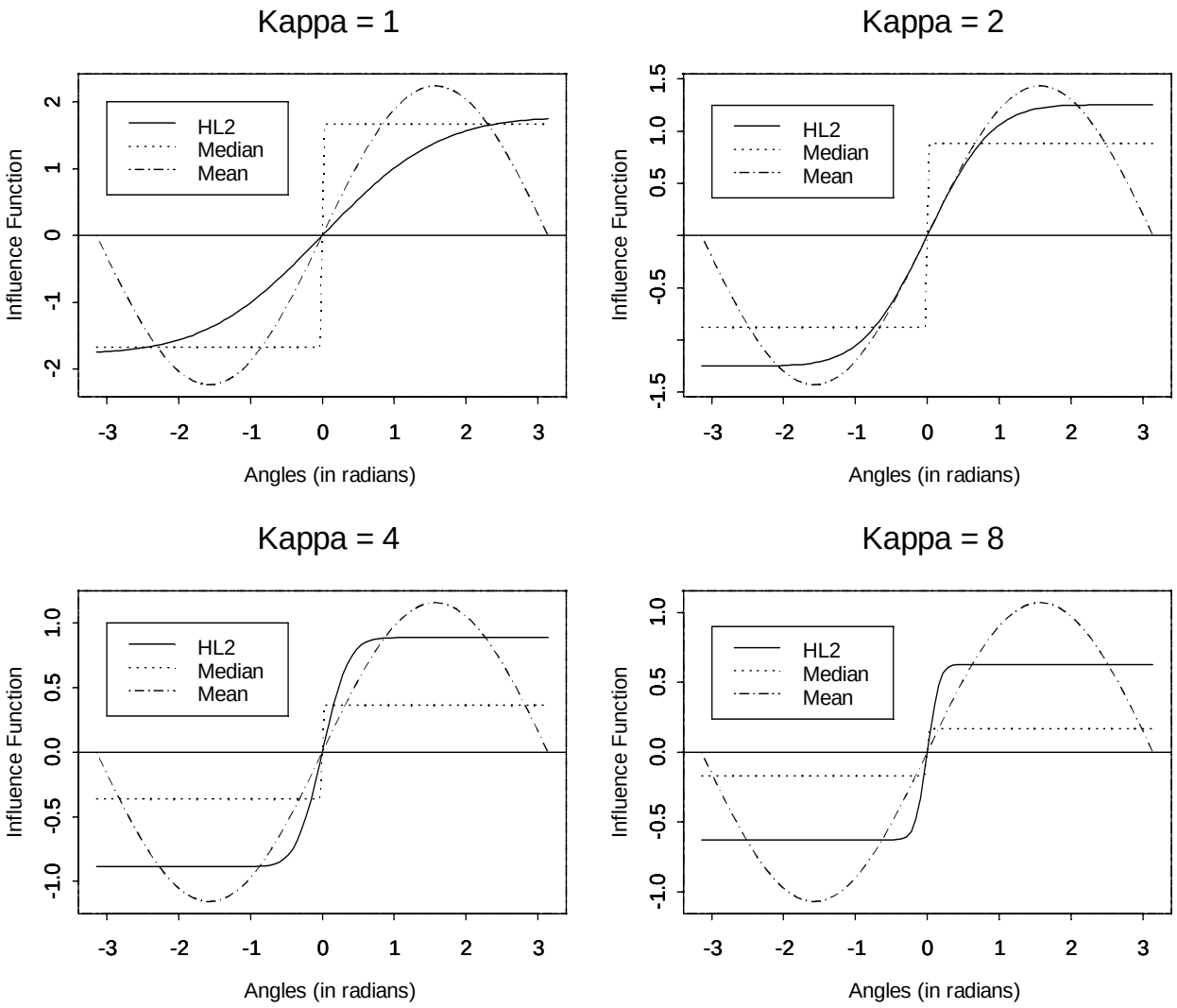
Assume that θ_i and θ_j are iid, with distribution function $F(\theta)$. Let $\Phi = \frac{(\theta_i + \theta_j)}{2}$, $i \le j$. Φ is equivalent to the pairwise circular mean of θ_i and θ_j, Otieno and Anderson-Cook,(2003a). The functional of the circular Hodges-Lehmann estimator $\hat{\theta}^c{}_{HL}$ is the Pseudo-Median Locational functional $F = F^{*-1}\left(\frac{1}{2}\right)$, where $F(\phi) = P(\Phi \le \phi) = \int F(2\phi - \theta)h(\theta)d\theta$, Hettmansperger & McKean (1998, p.3,10-11). For a sample from a von Mises distribution with a limited range of concentrated parameter values, $\kappa \ge 2$, the influence function of the circular HL estimator $\hat{\theta}^c{}_{HL}$ is given by

$$IF(\theta) = \frac{F(\theta) - \frac{1}{2}}{\left(\frac{\kappa}{4\pi}\right)^{\frac{1}{2}}},$$

where F(.) is the cumulative density function of $\theta_1, \ldots, \theta_n$. Note that this influence function is a centered and scaled cdf and is therefore bounded. Note that, it is also discontinuous at the antimode, like the influence function of the circular median.

Figure 2 are plots of the influence functions of the circular mean, circular median and the circular HL estimators for preferred direction for various concentration parameters. The range of the data values is $\frac{-\pi}{2}$ radians to $\frac{\pi}{2}$ radians and 4 dispersion values ranging from κ =1 to 8.

Figure 2: Influence Functions for measures of preferred direction, $(1 \leq \kappa \leq 8)$.



Notice that all the estimators have curves which are bounded. Also, as the data becomes more concentrated (with κ increasing), the influence function of the circular median changes least followed by the circular HL estimator. This is similar to the linear case.

Also, as κ increases, the bound for the influence function for all the three measures decreases, however, overall the bound of the influence function for the mean is largest for angles closest to $\frac{\pi}{2}$ radians from the preferred direction. The maximum influence for the mean

occurs at $\frac{\pi}{2}$ or $\frac{-\pi}{2}$ from the mode for all κ, while for both the median and HL, the maximum occurs uniformly for a range away from the preferred direction. Overall, HL seems like a compromise between the mean and the median.

A Practical Example

Consider the following example of Frog migration data Collett (1980), shown in Figure 3. The data relates the homing ability of the Northern cricket frog, *Acris crepitans*, as studied by Ferguson, et. al.(1967). A number of frogs were collected from mud flats of an abandoned stream meander and taken to a test pen lying to the north of the collection point. After 30 hours enclosure within a dark environmental chamber, 14 of them were released and the directions taken by these frogs recorded (taking 0^0 to be due North), Table 1.

In order to compute the sample mean of these data, consider them as unit vectors, the resultant vector of these 14 unit vectors is obtained by summing them componentwise to get $R = \left(\sum_{i=1}^{n} \cos(\theta_i), \sum_{i=1}^{n} \sin(\theta_i) \right) = (C, S)$, say. The sample circular mean is the angle corresponding to the mean resultant vector $\overline{R} = \frac{R}{n} = \left(\frac{C}{n}, \frac{S}{n} \right) = (\overline{C}, \overline{S})$. That is, the angle corresponding to the mean resultant length $|\overline{R}| = \sqrt{(\overline{C}^2 + \overline{S}^2)}$.For this data the circular mean is -0.977 (124^0), the mean resultant length, $|\overline{R}| = 0.725$, thus, the estimate of the concentration parameter, $\hat{\kappa} = 2.21$ for the best fitting von Mises.(Table A.3, Fisher, 1993, p. 224).

The circular median is -0.816 (133.25^0) and circular Hodges-Lehmann is -0.969 (124.5^0). Using $\hat{\kappa} = 2.21$, Figure 4 gives the influence curves of the mean, median and HL. Note that the measure least influenced by observation x, a presumed outlier, is the circular mean, since x is nearer to the antimode. However, the circular median is influenced most by observations nearest the center of the data followed by HL. The influence of an outlier on the sample circular median is bounded at either a constant positive or a constant negative value, regardless of how far the outlier is from the center of the data. On the other hand, the HL estimator is influenced less by observations near the center, and reflects the presence of the outlier. The influence curve for the circular mean is similar to that of the redescending Φ function (See Andrews et. al., 1972 for details).

Conclusion

Like in the linear case, it is helpful to decide what aspects of the data are of interest. For example, in the case of distributions that are not symmetric or have outliers, like in the case of the Frog migration data, the circular mean and circular median are measuring different characteristics of the data. Hence one needs to choose which aspect of the data is of most interest. For data that are close to uniformly distributed or have rounding or grouping, it is wise to avoid the median since its estimate is prone to undesirable jumps. Either of the other two measures perform similarly. For data spread on a smaller fraction of the circle, with a natural break in the data, the median is least sensitive to outliers. The mean is typically most responsive to outliers, while HL gives some, but not too much weight to outliers.

Overall, the circular HL is a good compromise between circular mean and circular median, like its counterpart for linear data. The HL estimator is less robust to outliers compared to the median, however it is an efficient alternative, since it has a smaller circular variance, Otieno and Anderson-Cook, (2003a). The HL estimator also provides a robust alternative to the mean especially in situations where the model of choice of circular data (the von Mises distribution) is in doubt. Overall, the circular HL estimate is a solid alternative to the established circular mean and circular median with some of the desirable features of each.

Figure 3:The Orientation of 14 Northern Cricket Frogs

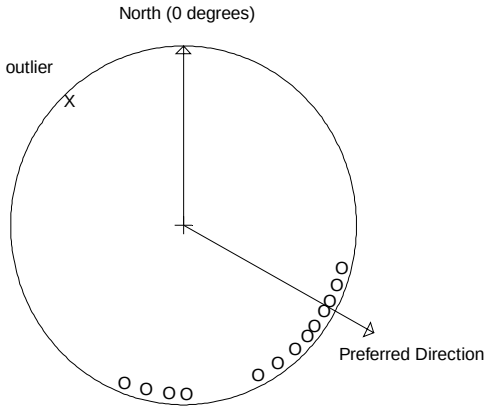
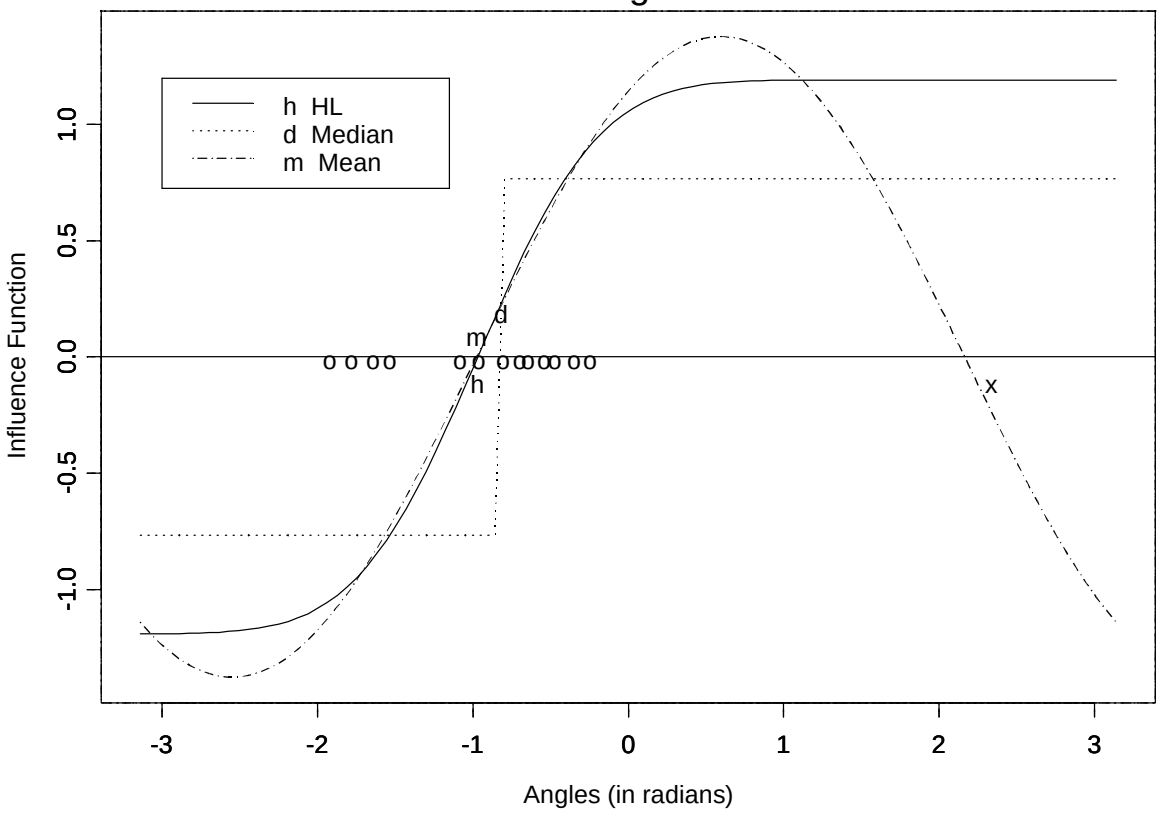


Figure 4: Influence curves for the three measures for data with a single outlier



References

Ackermann, H. (1997). A note on circular nonparametrical classification. *Biometrical Journal*, *5*, 577-587.

Andrews, D. R., Bickel, P. J., Hampel,F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances.* New Jersey: Princeton University Press.

Collett, D. (1980). Outliers in circular data. *Applied Statistics*, *29*, 50-57.

Collett, D., & Lewis, T. (1981). Discriminating between the von Mises and wrapped normal distributions. *Australian Journal of Statistics*, 23, 73-79.

Ferguson, D. E., Landreth, H. F., & McKeown, J. P. (1967). Sun compass orientation of northern cricket frog, Acris crepitans. *Animal Behaviour*, *15*, 45-53.

Fisher, N. I. (1993). *Statistical analysis of circular data.* Cambridge University Press.

Ko, D., & Guttorp P. (1988). Robustness of estimators for directional data. *Annals of Statistics*, *16*, 609-618.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York: Wiley.

Hettmansperger, T. P., & McKean, J. W. (1998). *Robust nonparametric statistical methods*. New York: Wiley.

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics, world scientific*. New Jersey.

Mardia, K. V. (1972) *Statistics of directional data.* London: Academic Press.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester: Wiley.

Otieno, B. S. (2002) *An alternative estimate of preferred direction for circular data*. Ph.D Thesis., Department of Statistics, Virginia Tech. Blacksburg: VA.

Otieno, B. S., & Anderson-Cook, C. M. (2003a). Hodges-Lehmann estimator of preferred direction for circular. *Virginia Tech Department of Statistics Technical Report*, 03-3.

Otieno, B. S., & Anderson-Cook, C. M. (2003b). A More efficient way of obtaining a unique median estimate for circular data. *Journal of Modern Applied Statistical Methods*, *3*, 334-335.

Rao, J. S. (1984) Nonparametric methods in directional data. In P. R. Krishnaiah and P. K. Sen. (Eds.), *Handbook of Statistics*, 4, pp. 755-770. Amsterdam: Elsevier Science Publishers.

Stephens, M. A. (1963). Random walk on a circle. *Biometrika*, *50*, 385-390.

Watson, G. S. (1986). Some estimation theory on the sphere. *Annals of the Institute of Statistical Mathematics*, 38, 263-275.

Wehrly, T., & Shine, E. P. (1981). Influence curves of estimates for directional data. *Biometrika*, *68*, 334-335.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 90-99

1538 – 9472/05/$95.00

Bias Of The Cox Model Hazard Ratio

Inger Persson
TFS Trial Form Support AB
Stockholm, Sweden

Harry Khamis
Statistical Consulting Center
Wright State University

The hazard ratio estimated with the Cox model is investigated under proportional and five forms of nonproportional hazards. Results indicate that the highest bias occurs for diverging hazards with early censoring, and for increasing and crossing hazards under a high censoring rate.

Key words: censoring proportion, proportional hazards, random censoring, survival analysis, type I censoring

Introduction

In recent decades, survival analysis techniques have been extended far beyond the medical, biomedical, and reliability research areas to fields such as engineering, criminology, sociology, marketing, insurance, economics, etc. The study of survival data has previously focused on predicting the probability of response, survival, or mean lifetime, and comparing the survival distributions. More recently, the identification of risk and/or prognostic factors related to response, survival, and the development of a certain condition has become equally important (Lee, 1992).

Conventional statistical methods are not adequate to analyze survival data because some observations are censored, i.e., for some observations there is incomplete information about the time to the event of interest. A common type of censoring in practice is Type I censoring, where the event of interest is observed only if it occurs prior to some pre-specified time, such as the closing of the study or the end of the follow-up. The most common approach for modeling covariate effects in survival data uses the *Cox Proportional Hazards Regression Model* (Cox, 1972), which takes into account the effect of censored observations. As the name indicates, the Cox model relies on the assumption of proportional hazards, i.e., the assumption that the effect of a given covariate does not change over time. If this assumption is violated, then the Cox model is invalid and results deriving from the model may be erroneous.

A great number of procedures, both numerical and graphical, for assessing the validity of the proportional hazards assumption have been proposed over the years. Some of the procedures require partitioning of failure time, some require categorization of covariates, some include a spline function, and some can be applied to the untransformed data set.

However, no method is known to be definitively better than the others in determining nonproportionality. Some authors recommended using numerical tests, e.g., Hosmer and Lemeshow (1999). Others recommended graphical procedures, because they believe that the proportional hazards assumption only approximates the correct model for a covariate and that any formal test, based on a large enough sample, will reject the null hypothesis of proportionality (Klein & Moeschberger, 1997, p. 354).

Power studies to compare some numerical tests have been performed; see, e.g.,

Harry Khamis, Statistical Consulting Center Wright State University, Dayton, OH. 45435. Email him: harry.khamis@wright.edu.

Ng'andu, 1997; Quantin, et al., 1996; Song & Lee, 2000, and Persson, 2002. The goal of this article is to assess the bias of the Cox model estimate of the hazard ratio under different censoring rates, sample sizes, types of nonproportionality, and types of censoring. The second section reviews the Cox regression model and the proportional hazards assumption. The average hazard ratio, the principal criterion against which the Cox model estimates are compared, is described in the third section. The fourth section presents the simulation strategy. The results and conclusions are given in the remaining two sections.

Cox proportional hazards model

A central quantity in the Cox regression model is the hazard function, or the hazard rate, defined by:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t \mid T \ge t]}{\Delta t},$$

where T is the random variable under study: time until the event of interest occurs. Thus, for small Δt, $\lambda(t)\Delta t$ is approximately the conditional probability that the event of interest occurs in the interval $[t, t + \Delta t]$, given that it has not occurred before time t.

There are many general shapes for the hazard rate; the only restriction is $\lambda(t) \ge 0$. Models with increasing hazard rates may arise when there is natural aging or wear. Decreasing hazard functions are less common, but may occur when there is a very early likelihood of failure, such as in certain types of electronic devices or in patients experiencing certain types of transplants.

A bathtub-shaped hazard is appropriate in populations followed from birth. During an early period deaths result, primarily from infant diseases, after which the death rate stabilizes, followed by an increasing hazard rate due to the natural aging process. Finally, if the hazard rate is increasing early and eventually begins declining, then the hazard is termed "hump-shaped." This type of hazard rate is often used in modeling survival after successful surgery, where there is an initial increase in risk due to infection or other complications just after the procedure followed by a steady decline in risk as the patient recovers (see, e.g., Kline & Moeschberger, 1997).

In the Cox model, the relation between the distribution of event time and the covariates $\mathbf{z}$ (a p x 1 vector) is described in terms of the hazard rate for an individual at time t:

$$\lambda(t,\mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}), \qquad (1)$$

where $\lambda_0(t)$ is the baseline hazard rate, an unknown (arbitrary) function giving the value of the hazard function for the standard set of conditions $\mathbf{z} = \mathbf{0}$, and $\boldsymbol{\beta}$ is a p x 1 vector of unknown parameters. The partial likelihood estimate of $\boldsymbol{\beta}$ is asymptotically consistent (Andersen & Gill, 1982; Cox, 1975, and Tsiatis, 1981).

The ratio of the hazard functions for two individuals with covariate values $\mathbf{z}$ and $\mathbf{z}^*$ is $\lambda(t,\mathbf{z})/\lambda(t,\mathbf{z}^*) = \exp[\boldsymbol{\beta}'(\mathbf{z} - \mathbf{z}^*)]$, an expression that does not depend on t. Thus, the hazard functions are proportional over time. The factor $\exp(\boldsymbol{\beta}'\mathbf{z})$ describes the hazard ratio for an individual with covariates $\mathbf{z}$ relative to the hazard at a standard $\mathbf{z} = \mathbf{0}$. The usual interpretation of the hazard ratio, $\exp(\boldsymbol{\beta}'\mathbf{z})$, requires that (1) holds. There is no clear interpretation if the hazards are not proportional.

Of principal interest in a Cox regression analysis is to determine whether a given covariate influences survival, i.e. to estimate the hazard ratio for that covariate. The behavior of the hazard ratio estimated with the Cox model when the underlying assumption of proportional hazards is false (i.e., when the hazards are not proportional) is investigated in this paper. To assess the Cox estimates under nonproportional hazards, the estimates are compared to an exact calculation of the geometric average of the hazard ratio described in the next section. An average hazard ratio does not reflect the truth exactly since the hazard ratio is changing with time when the proportionality assumption is not in force. However, it can provide an approximate standard against which to compare the Cox model estimates. Because the estimation of the hazard ratio from the Cox model cannot be done analytically (Klein & Moeschberger, 1997), the comparison is made by simulations.

Average hazard ratio

The average hazard ratio (AHR) is defined as (Kalbfleisch & Prentice, 1981):

$$\theta(W) = -\int_0^\infty [\lambda_1(t)/\lambda_2(t)]dW(t), \quad (2)$$

where $\lambda_1(t)$ and $\lambda_2(t)$ are the hazard functions of two groups and W(t) is a survivor or weighting function. The weight function can be chosen to reflect the relative importance attached to hazard ratios in different time periods. Here, W(t) depends on the general shape of the failure time distribution and is defined as $W(t) = S_1^{\varepsilon}(t)S_2^{\varepsilon}(t)$, where $S_1(t)$ and $S_2(t)$ are the survivor functions (i.e., one minus the cumulative distribution function) for the two groups, and $\varepsilon > 0$. The value $\varepsilon = ½$ weights the hazard ratio at time t according to the geometric average of the two survivor functions. Values of $\varepsilon > ½$ will assign greater weight to the early times while $\varepsilon < ½$ assigns greater weight to later times. Here, $\varepsilon = ½$ will be used.

For Weibull distributed lifetimes with scale parameter α and shape parameter γ, the survival function is $S(t) = \exp[-(\alpha t)\gamma]$ and the AHR estimator (2) can be written

$$\theta(W) = -\int_0^\infty [(\gamma_1\alpha_1{}^{\gamma 1})/(\gamma_2\alpha_2{}^{\gamma 2})]d\{\exp[-½((\alpha_1 t)^{\gamma 1} + (\alpha_2 t)^{\gamma 2})]\}.$$

When the parametric forms of the survivor functions are unknown, the AHR (2) can still be used; in this case, the Kaplan-Meier product-limit estimates for the two groups are used as the survivor functions (Kaplan & Meier, 1958). However, (2) then only holds for uncensored data. The AHR function for censored data can be found in Kalbfleisch and Prentice, 1981.

Methodology

Simulation strategy

The hazard ratio estimates from the Cox model are evaluated under six scenarios:
(1) proportional hazards, (2) increasing hazards, (3) decreasing hazards, (4) crossing hazards, (5) diverging hazards, and (6) converging hazards. The AHR is compared in the two-sample case, corresponding to two groups with different hazard functions.

Equal sample sizes of 30, 50, and 100 observations per group are used along with average censoring proportions of 10, 25, and 50 percent. Type I censoring is used along with early and late censoring. The number of repetitions used in each simulation is 10,000. For a given sample size, censoring proportion, and type of censoring (random, early, late), the mean Cox estimate is calculated for all scenarios except converging hazards. Because of the asymmetry in the distribution of values in the case of converging hazards, the median estimate is used. For interpretation purposes, the percent bias of the mean or median Cox estimate relative to the AHR is reported in tables.

For the case of random censoring, random samples of survival times t_s are generated from the Weibull distribution. The hazard function for the Weibull distribution is $\lambda(t) = \alpha\gamma(\alpha t)^{\gamma-1}$. The censoring times t_c are generated from the exponential distribution with hazard function $\lambda(t) = \beta$, where the value of β is adjusted to achieve the desired censoring proportions. The time on study t is defined as:

$$t = \begin{cases} t_s & \text{if } t_s \le t_c \\ t_c & \text{if } t_s > t_c \end{cases}$$

The event indicator is denoted by d:

$$d = \begin{cases} 0, & \text{if the observation is censored} \\ 1, & \text{if the event has occurred} \end{cases}$$

For early censoring, a percentage of the lifetimes are randomly chosen and multiplied by a random number generated from the uniform distribution. The percentage chosen is the same as the censoring proportion. The parameters of the uniform distribution are chosen so that the censoring times are short in order to achieve the effect of early censoring. For late censoring, a percentage of the longest lifetimes are chosen; this percentage is slightly larger than the censoring proportion. Of those lifetimes, a percentage corresponding to the censoring time

is the lifetime, t_s, minus a random number generated from the uniform distribution. The parameters of the uniform distribution are now chosen so that the random numbers are relatively small in order to achieve the effect of late censoring.

Results

For each of the six scenarios concerning the hazard rates of the two groups, comparisons of the estimated hazards ratio from the Cox model to the AHR is made for random, early, and late censoring and for selected sample sizes and censoring rates. The comparison is made based on the percent difference (bias) between the average Cox hazard ratio estimate and the AHR; [(average Cox estimate – AHR)/AHR] x 100.

Proportional Hazards

Survival times are generated from the Weibull distribution where γ=1, α=1 for group 1, and γ=1, α=2 for group 2. The AHR is 2.0 for this situation. The percent of the bias for the mean Cox model estimate relative to the AHR is given in Table 1.

Under proportional hazards, the Cox model is correct. So, the estimated hazard ratio from the Cox model should be close to 2.0 in all cases. Table 1 reveals that the Cox estimate is slightly biased. This bias grows with decreasing sample size or increasing censoring proportion. Early censoring produces a more biased estimate than random or late censoring, especially for high censoring proportions.

Increasing Hazards

Survival times are generated from the Weibull distribution where γ=1.5, α=2 for group 1, and γ=2, α=2 for group 2. The AHR is 1.2 for this situation. The percent of the bias for the mean Cox model estimate relative to the AHR is given in Table 2.

The Cox estimates fall below the AHR for increasing hazards. The estimates closest to the AHR correspond to early censoring; these estimates are relatively stable regardless of censoring proportion or sample size. For random and late censoring the estimate decreases (higher bias) with increasing censoring proportion but remains stable relative to sample size. For early censoring the estimate is generally unbiased regardless of sample size or censoring proportion.

Decreasing Hazards

Survival times are generated from the Weibull distribution where γ=0.9, α=1 for group 1, and γ=0.75, α=3 for group 2. The AHR is 0.44 for this situation. The percent of the bias for the mean Cox model estimate relative to the AHR is given in Table 3.

The Cox estimates fall below the AHR. These estimates decrease slightly with increasing censoring proportion. The estimates for early censoring are slightly less biased than for random or late censoring at the higher censoring proportions. The bias is not heavily influenced by sample size.

Crossing Hazards

Survival times are generated from the Weibull distribution where γ=2.5, α=0.3 for group 1, and γ=0.9, α=2 for group 2. The AHR is 15.4 for this situation. The percent of the bias for the mean Cox model estimate relative to the AHR is given in Table 4.

The bias of the Cox estimates tends to be much smaller for 10% and 25% censoring proportions compared to the 50% censoring proportion. For 50% censoring, the Cox model tends to overestimate the AHR. The bias decreases with increasing sample size, especially for high censoring proportions.

Diverging Hazards

Survival times are generated from the Weibull distribution where γ=0.9, α=1.0 for group 1, and γ=1.5, α=2 for group 2. The AHR is 0.536 for this situation. The percent of the bias for the mean Cox model estimate relative to the AHR is given in Table 5.

The Cox estimates are larger for random and late censoring than for early censoring at the highest censoring proportion. Generally, the sample size has little effect on the bias. For early censoring, the percent bias is approximately 20% and is not strongly affected by sample size or censoring proportion.

Table 1. Proportional Hazards: percent bias of Cox model estimates relative to average hazard rate of 2.0.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	5.5	4.0	2.0
	25%	8.0	4.5	2.0
	50%	11.0	5.5	3.0
Early	10%	7.0	5.0	2.5
	25%	10.0	6.0	3.5
	50%	19.5	11.0	7.0
Late	10%	5.5	4.0	2.0
	25%	7.0	4.0	2.5
	50%	10.5	6.5	3.5

Table 2. Increasing Hazards: percent bias of Cox model estimates relative to average hazard rate of 1.20.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	- 6.7	- 7.5	- 8.3
	25%	- 9.2	-10.8	-10.8
	50%	-15.0	-17.5	-18.3
Early	10%	- 4.2	- 5.8	- 6.7
	25%	- 4.2	- 5.8	- 5.8
	50%	- 1.7	- 5.0	- 5.8
Late	10%	- 7.5	- 9.2	-10.0
	25%	-12.5	-14.2	-15.0
	50%	-20.8	-22.5	-23.3

Table 3. Decreasing Hazards: percent bias of Cox model estimates relative to average hazard rate of 0.441.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	- 2.0	- 3.2	- 3.2
	25%	- 4.3	- 5.7	- 5.9
	50%	- 9.5	-11.3	-12.2
Early	10%	- 1.4	- 2.5	- 2.3
	25%	- 2.7	- 3.6	- 3.6
	50%	- 5.4	- 5.9	- 6.6
Late	10%	- 2.0	- 3.4	- 3.6
	25%	- 4.9	- 6.8	- 7.3
	50%	-10.9	-12.9	-13.8

Table 4. Crossing Hazards: percent bias of Cox model estimates relative to average hazard rate of 15.4.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	5.8	- 7.1	-14.9
	25%	19.5	4.5	- 5.2
	50%	73.3	52.6	34.4
Early	10%	1.3	-11.0	-18.8
	25%	9.1	- 5.2	-15.6
	50%	32.5	8.4	- 6.5
Late	10%	- 1.9	-12.9	-19.5
	25%	- 0.6	- 5.8	- 8.4
	50%	100.6	81.8	67.5

Table 5. Diverging Hazards: percent bias of Cox model estimates relative to average hazard rate of 0.536.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	-16.2	-18.3	-19.2
	25%	-10.4	-12.9	-14.2
	50%	7.8	3.7	1.1
Early	10%	-19.0	-20.9	-22.0
	25%	-19.0	-21.3	-22.6
	50%	-18.8	-21.8	-23.7
Late	10%	-16.4	-18.5	-19.4
	25%	- 6.9	- 9.3	-10.4
	50%	18.5	13.9	12.3

Table 6. Converging Hazards: percent bias of Cox model estimates relative to average hazard rate of 7.15.

		Sample Size per Group		
Censoring	% Censored	30	50	100
Random	10%	- 8.9	-11.2	-12.2
	25%	- 5.6	- 8.3	- 9.4
	50%	4.0	1.9	- 0.6
Early	10%	- 9.4	-11.3	-12.4
	25%	- 6.2	- 8.8	-10.2
	50%	2.4	- 0.8	- 4.3
Late	10%	-10.2	-12.4	-13.1
	25%	- 7.3	- 8.4	- 8.1
	50%	10.5	9.2	8.1

Converging Hazards

Survival times are generated from the Weibull distribution where γ=0.9, α=6.0 for group 1, and γ=1.2, α=1 for group 2. The AHR is 7.15 for this situation. The percent of the bias for the median Cox model estimate relative to the AHR is given in Table 6. The median Cox estimate increases with increasing censoring proportion. The bias is not heavily influenced by sample size.

Conclusion

Just as with the classical maximum likelihood estimator, the maximum partial likelihood estimator is not unbiased, but it is asymptotically unbiased (Kotz & Johnson, 1985, p. 591-593). This behavior is evident in Table 1, where the Cox estimates can be seen to be larger than the AHR, but the bias decreases with increasing sample size regardless of the type of censoring or the censoring rate.

Table 7 shows those instances where the average percent bias exceeds 20% in absolute value; the entries are the percent bias averaged over sample size.

There is no serious bias for the proportional hazards case regardless of type of censoring or censoring rate. Similarly, there is no serious bias in the cases of decreasing or converging hazards.

Under-estimation occurs for increasing hazards at the 50% censoring rate with late censoring. It also occurs for diverging hazards with early censoring regardless of censoring rate. Over-estimation occurs for crossing hazards at the 50% censoring rate with random and late censoring.

One might suspect that late censoring would render the least biased estimates since such a data structure contains more information than early or random censoring. However, late censoring leads to severe bias for increasing and crossing hazards when the censoring proportion is high. For lower censoring proportions (25% or lower), there is no severe bias for any of the nonproportionality models except diverging hazards.

As a practical matter, one can obtain descriptive statistics from a given data set, including percent censored, sample sizes, and a plot of the hazard curves. From this information, one can approximate the magnitude and nature of the risk of biased estimation of the hazard ratio by the Cox model. Generally, the least biased estimates are obtained for the lower censoring proportions (10% and 25%) except for diverging hazards. In terms of bias, early censoring is problematic only for diverging hazards; late censoring is problematic for increasing and crossing hazards with the 50% censoring rate; and random censoring is problematic for crossing hazards with the 50% censoring rate. The case corresponding to the least occurrence of severe bias is the one involving random censoring with a censoring rate of 25% or less.

In practice, the experimenter typically has some control over sample size and perhaps the censoring proportion. For instance, the experimenter may be able to minimize censoring proportion, depending on the situation, through effective study design and experimental protocol. Minimizing the censoring rate is generally recommended, especially for increasing and crossing hazards. Early censoring is appreciably affected by censoring proportion only for constant and crossing hazards. Sample size has the strongest effect on constant and crossing hazards, especially at higher censoring proportions, where higher sample sizes lead to less biased estimates.

In practical applications, the proportional hazards assumption is never met precisely. If the deviation from the proportional hazards assumption is severe, then remedial measures should be taken. However, in many instances the model diagnostics reveal only a small to moderate deviation from the proportional hazards assumption. In these cases, the Cox model estimate of the hazard ratio is used for interpretation purposes in the presence of small to moderate assumption violations. This study characterizes the consequences of this interpretation in terms of bias, taking into account censoring rate, type of censoring, type of nonproportional hazards, and sample size. The general results indicate that the percent bias relative to AHR is under 20% in all but a few specific instances, as outlined above.

Table 7. Percent bias of the average Cox regression model estimates of the hazard ratio relative to the AHR averaged over sample size.

		Censoring		
Hazards	% censoring	random	early	late
constant	10	*	*	*
	25	*	*	*
	50	*	*	*
increasing	10	*	*	*
	25	*	*	*
	50	*	*	**-22**
decreasing	10	*	*	*
	25	*	*	*
	50	*	*	*
crossing	10	*	*	*
	25	*	*	*
	50	**53**	*	**83**
diverging	10	*	**-21**	*****
	25	*	**-21**	*
	50	*	**-21**	*****
converging	10	*	*	*
	25	*	*	*
	50	*	*	*

*under 20% in absolute value

References

Andersen, P. K., Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics, 10*, 1100-1120.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, 34*, 187-220.

Cox, D R., (1975). Partial likelihood. *Biometrika*, *62*, 269-276.

Hosmer, D. W., Lemeshow, S. (1999). Regression modeling of time to event data. New York: John Wiley & Sons.

Kalbfleisch. J. D., Prentice, R. L. (1981). Estimation of the average hazard ratio. Biometrika, *68*, 105-112.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457-481.

Klein, J. P., Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. Springer: New York.

Kotz, S., & Johnson, N. L., (1985). *Encyclopedia of Statistical Sciences*, Volume 6, New York: John Wiley & Sons.

Lee, E. T. (1992). Statistical methods for survival data analysis,(2nd Ed). Oklahoma City: John Wiley & Sons.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine*, *16*, 611- 626.

Persson, I. (2002). Essays on the assumption of proportional hazards in Cox regression. Acta Universitatis Upsaliensis. Unpublished Ph.D. dissertation.

Quantin, C., Moreau, T., Asselain, B., Maccario, J., & Lellouch, J. A. (1996). Regression survival model for testing the proportional hazards hypothesis. *Biometrics*, *52*, 874-885.

Song, H. H. & Lee, S. (2000). Comparison of goodness of fit tests for the Cox proportional hazards model. *Communications in Statistics – Simulation and Computation, 29*, 187-206.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics*, *9*, 93-108.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 100-105

1538 – 9472/05/$95.00

Bias Affiliated With Two Variants Of Cohen's *d* When Determining U_1 As A Measure Of The Percent Of Non-Overlap

David A. Walker
Educational Research and Assessment Department
Northern Illinois University

Variants of Cohen's d, in this instance d_t and d_{adj}, has the largest influence on U_1 measures used with smaller sample sizes, specifically when n_1 and n_2 = 10. This study indicated that bias for variants of *d*, which influence U_1 measures, tends to subside and become more manageable, in terms of precision of estimation, around 1% to 2% when n_1 and n_2 = 20. Thus, depending on the direction of the influence, both d_t and d_{adj} are likely to manage bias in the U_1 measure quite well for smaller to moderate sample sizes.

Key words: Non-overlap, effect size, Cohen's *d*

Introduction

In his seminal work on power analysis, Jacob Cohen (1969; 1988) derived an effect size measure, Cohen's *d*, as the difference between two sample means. Using n, M, and SD from two sample groups, *d* provided "score distances in units of variability" (p. 21), by translating the means into a common metric of standard deviation units pertaining to the degree of departure from the null hypothesis.

The common formula for Cohen's *d* (1988) is

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{pooled}} \qquad (1)$$

where $\bar{X}_1$ and $\bar{X}_2$ are sample means and

$$\hat{\sigma}_{pooled} = \frac{(n_1 - 1)(sd_1)^2 + (n_2 - 1)(sd_2)^2}{(n_1 + n_2 - 2)}$$

David Walker is an Assistant Professor at Northern Illinois University. His research interests include structural equation modeling, effect sizes, factor analyses, predictive discriminant analysis, predictive validity, weighting, and bootstrapping. Email: dawalker@niu.edu.

Cohen's *d* can be calculated if no n, M, or SD for two groups is reported via *t* values and degrees of freedom, termed d_t here, where it is assumed that n_1 and n_2 are equal (Rosenthal, 1991):

$$d_1 = \frac{2t}{\sqrt{df}} \qquad (2)$$

where *t* = *t* value, and df = $n_1 + n_2 - 2$

Kraemer (1983) noted that the distribution of Cohen's *d* was skewed and heavy tailed, and Hedges (1981) found that *d* was a positively biased effect size estimate. Hedges proposed an approximate, modified estimator of *d*, which will be termed d_{adj} here, where:

$$c(m) \approx 1 - \frac{3}{4m - 1} \qquad (3)$$

where $m = n_1 + n_2 - 2$.

Cohen (1969; 1988) revisited the idea of group overlap, which was studied by Tilton (1937), and the degree of overlap (O) between two distributions; and also in close proximity to the time of Cohen's initial work (i.e., 1969) by Elster and Dunnete (1971). This resulted in the U_1 measure, which was derived from *d* as a percent of non-overlap. As Cohen (1988) explained, "If we maintain the assumption that the populations being compared are normal and with equal variability, and conceive them further as equally numerous, it is possible to define measures of non-overlap (U_1) associated with *d*"

(p. 21).

Algebraically, U_1 is related to the cumulative normal distribution and is expressed as (Cohen, 1988):

$$U_1 = \frac{2P_{d/2} - 1}{P_{d/2}} = \frac{2U_2 - 1}{U_2} \quad (4)$$

where d = Cohen's *d* value, P = percentage of the area falling below a given normal deviate, and $U_2 = P_{d/2}$.

In SPSS (Statistical Package for the Social Sciences) syntax, U_1 is calculated using the following expressions:

```
Compute U = CDF.NORMAL((ABS(d)/2),0,1).
Compute U1 = (2*U-1)/U*100.
Execute.
```

where d = Cohen's d value, ABS = absolute value, CDF. NORMAL = cumulative probability that a value from a normal distribution where M = 0 and SD = 1 is < the absolute value of d/2.

Thus, the link between *d* and U_1 was seen by Cohen (1988) in that, "d is taken as a deviate in the unit normal curve and P [from expression 4] as the percentage of the area (population of cases) falling below a given normal deviate" (p. 23).

For Cohen (1998), non-overlap was the extent to which an experiment or intervention had had an effect of separating the two populations of interest. A high percentage of non-overlap indicated that the two populations were separated greatly. When $d = 0$, there was 0% overlap and $U_1 = 0$ also, or as Cohen (1988) noted "either population distribution is perfectly superimposed on the other" (p. 21). Therefore, the two populations were identical.

The assumptions for the percentage of population non-overlap are: 1) the comparison populations have normality and 2) equal variability. Further, Cohen (1988) added that the U_1 measure would also hold for samples from two groups if "the samples approach the conditions of normal distribution, equal variability, and equal sample size" (p. 68).

Cohen (1988, p. 22) went on to produce Table 2.2.1, which consisted of non-overlap percentages for values of *d*. Assuming a normal distribution, this table showed that, for example, a value of $d = .20$ would have a corresponding $U_1 = 14.7\%$, or a percentage of non-overlap of just over 14%. That is, the distribution of scores for the treatment group overlapped only a small amount with the distribution of scores for the non-treatment group, which was manifested in the small effect size of .20. As the value of *d* increased, so would the percentage of non-overlap between the two distributions of scores, which indicated that the two groups differed considerably.

Methodology

After an extensive review of the literature, it was found that very few studies included effect size indices with tests for statistical significance and none produced a U_1 measure when any of the variants of *d* were reported. Further, beyond studies, for example, by Hedges (1981) or Kraemer (1983) related to the upward bias and skewness associated with *d* in small samples, it appears in the scholarly literature that *d* as a percent of non-overlap has not been studied to evaluate any bias affiliated with variants of *d*, d_t and d_{adj}, substituted for it in the calculation of U_1, except for what has been provided by Cohen (1988).

Thus, the intent of this research was to examine U_1 under varying sizes of *d* and n (i.e., $n_1 = n_2$). That is, this research looked at *d* values of .2, .5, .8, 1.00, and 1.50, which represent in educational research typically small to extremely large effect sizes. The sizes of n were 10, 20, 40, 50, 80, and 120, which represent in educational research small to large sample sizes. It should be noted, though, as was first discussed by Glass, McGaw, and Smith (1981), and reiterated by Cohen (1988), about the previously-mentioned *d* effect size target values and their importance:

> these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in

> a research climate characterized by a neglect of attention to issues of magnitude (p. 532).

Using the work of Aaron, Kromrey, and Ferron (1998), this study's tables will display the bias and proportional bias found in each U_1 measure found via both d_t and d_{adj}. As noted in the Aaron et al. research, the current study defines bias as the difference between the tabled value of U_1, derived from the standard *d* formula and presented by Cohen (1988) as Table 2.2.1, and the presented U_1 value resultant from d_t and d_{adj}., respectively. Proportional bias, or the "size of [the] bias as a proportion of the actual effect size estimate" (Aaron et al., p. 9), will be defined as the bias found above divided by the presented estimate for U_1 derived from both d_t and d_{adj}, respectively (see Tables 1 and 2).

Results

Using syntax written in SPSS v. 12.0 to obtain the results of the study, Tables 1 and 2 indicated, as would be expected, that regardless of the variant of *d* used, as the value of *d* increased, the bias in U_1 decreased. For example, Table 1 shows that at a small value of $d = .2$, and also at a moderate value of $d = .5$, the bias for small to moderate sample sizes ranged from about 1% to over 4%. As the value of *d* increased into the large effect size range of $d = .8$ to 1.50, the bias for the same sample sizes ranged from about 3% to under 1%.

The bias related to the U_1 measure for both forms of *d* used in this study was similar with both variants of *d*, the bias was constant with small sample sizes having 3% to 4% bias, moderate sample sizes having about 1%, and large sample sizes having very small amounts of bias. More specifically, it did appear, though, that the bias related to d_{adj} decreased more readily after $d = .2$ than was seen with d_t. That is, when $d = .20$, the bias for $d_t = 4.5\%$ and the bias for $d_{adj} = 4.3$ %, which were very similar. However, when $d = .5$, d_t incurred a bias of 4.4%, while the bias for $d_{adj} = 3.5\%$. This trend continued to $d = 1.50$, with d_{adj} incurring less bias than d_t, or stated another way, d_t had more of a biased effect on U_1. d_t's over-estimation property was also noted by Thompson and Schumacker (1997) in a study that assessed the effectiveness of the binomial effect size display.

Conclusion

Finally, as was found by Aaron et al. (1998), Hedges (1981), and Kraemer (1983), this study added to the literature that the biases found in variants of *d*, in this instance d_t and d_{adj}, had the largest influence on U_1 measures used with smaller sample sizes, specifically when n_1 and $n_2 = 10$. Although not looking at U_1 measures per se, the Aaron et al., Hedges, and Kraemer studies showed the effect of small sample sizes on *d* and variants of *d* when n_1 and $n_2 = 5$ or 10.

The current study indicated that bias for variants of *d* tended to subside and become more manageable, in terms of precision of estimation, around 1% to 2% when n_1 and $n_2 = 20$, or beyond very small sample sizes of n_1 and $n_2 = 5$ and 10. This is favorable for educational and behavioral sciences research designs that contain sample sizes typically of less than 100 participants (Huberty & Mourad, 1980). Thus, both d_t and d_{adj} tended to manage bias in the U_1 measure quite well for smaller to moderate sample sizes.

Table 1: Bias Affiliated with Estimates of U_1 Derived from d_t

$n_1 = n_2$	d	U_1	U_1 via d_t	Bias ($U_1 - U_1$ d_t)	Proportional Bias (Bias / U_1 d_t)
10	.2	14.7	15.4	.7	.045
20	.2	14.7	15.1	.4	.026
40	.2	14.7	14.9	.2	.013
50	.2	14.7	14.8	.1	.007
80	.2	14.7	14.8	.1	.007
120	.2	14.7	14.7	0	0

Table 1 Continued.

$n_1 = n_2$	d	U_1	U_1 via d_t	Bias ($U_1 - U_1$ d_t)	Proportional Bias (Bias / U_1 d_t)
10	.5	33.0	34.5	1.5	.044
20	.5	33.0	33.7	.7	.021
40	.5	33.0	33.4	.4	.012
50	.5	33.0	33.3	.3	.009
80	.5	33.0	33.2	.2	.006
120	.5	33.0	33.1	.1	.003

$n_1 = n_2$	d	U_1	U_1 via d_t	Bias ($U_1 - U_1$ d_t)	Proportional Bias (Bias / U_1 d_t)
10	.8	47.4	49.2	1.8	.037
20	.8	47.4	48.3	.9	.019
40	.8	47.4	47.8	.4	.008
50	.8	47.4	47.7	.3	.006
80	.8	47.4	47.6	.2	.004
120	.8	47.4	47.5	.1	.002

$n_1 = n_2$	d	U_1	U_1 via d_t	Bias ($U_1 - U_1$ d_t)	Proportional Bias (Bias / U_1 d_t)
10	1.00	55.4	57.4	2.0	.035
20	1.00	55.4	56.4	1.0	.018
40	1.00	55.4	55.9	.5	.009
50	1.00	55.4	55.8	.3	.005
80	1.00	55.4	55.6	.2	.004
120	1.00	55.4	55.5	.1	.002

$n_1 = n_2$	d	U_1	U_1 via d_t	Bias ($U_1 - U_1$ d_t)	Proportional Bias (Bias / U_1 d_t)
10	1.50	70.7	72.7	2.0	.028
20	1.50	70.7	71.7	1.0	.014
40	1.50	70.7	71.2	.5	.007
50	1.50	70.7	71.1	.4	.006
80	1.50	70.7	70.9	.2	.003
120	1.50	70.7	70.8	.1	.001

Table 2: Bias Affiliated with Estimates of U_1 Derived from d_{adj}

$n_1 = n_2$	d	U_1	U_1 via d_{adj}	Bias ($U_1 - U_1$ d_{adj})	Proportional Bias (Bias / U_1 d_{adj})
10	.2	14.7	14.1	.6	.043
20	.2	14.7	14.4	.3	.021
40	.2	14.7	14.6	.1	.007
50	.2	14.7	14.6	.1	.007
80	.2	14.7	14.6	.1	.007
120	.2	14.7	14.7	0	0

$n_1 = n_2$	d	U_1	U_1 via d_{adj}	Bias ($U_1 - U_1$ d_{adj})	Proportional Bias (Bias / U_1 d_{adj})
10	.5	33.0	31.9	1.1	.035
20	.5	33.0	32.5	.5	.015
40	.5	33.0	32.8	.2	.006
50	.5	33.0	32.8	.2	.006
80	.5	33.0	32.9	.1	.003
120	.5	33.0	33.0	0	0

$n_1 = n_2$	d	U_1	U_1 via d_{adj}	Bias ($U_1 - U_1$ d_{adj})	Proportional Bias (Bias / U_1 d_{adj})
10	.8	47.4	45.9	1.5	.033
20	.8	47.4	46.7	.7	.015
40	.8	47.4	47.1	.3	.006
50	.8	47.4	47.1	.3	.006
80	.8	47.4	47.2	.2	.004
120	.8	47.4	47.3	.1	.002

$n_1 = n_2$	d	U_1	U_1 via d_{adj}	Bias ($U_1 - U_1$ d_{adj})	Proportional Bias (Bias / U_1 d_{adj})
10	1.00	55.4	53.8	1.6	.030
20	1.00	55.4	54.7	.7	.013
40	1.00	55.4	55.1	.3	.005
50	1.00	55.4	55.1	.3	.005
80	1.00	55.4	55.2	.2	.004
120	1.00	55.4	55.3	.1	.002

$n_1 = n_2$	d	U_1	U_1 via d_{adj}	Bias ($U_1 - U_1$ d_{adj})	Proportional Bias (Bias / U_1 d_{adj})
10	1.50	70.7	69.1	1.6	.023
20	1.50	70.7	69.9	.8	.011
40	1.50	70.7	70.3	.4	.006
50	1.50	70.7	70.4	.3	.004
80	1.50	70.7	70.5	.2	.003
120	1.50	70.7	70.6	.1	.001

Reference

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Elster, R. S., & Dunnette, M. D. (1971). The robustness of Tilton's measure of overlap. *Educational and Psychological Measurement, 31*, 685-697.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Huberty, C. J., & Mourad, S. A, (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8*, 93-101.

Rosenthal, R. (1991). (Series Ed.), *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics, 22*, 109-117.

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology, 28*, 656-662.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 106-119

1538 – 9472/05/$95.00

Some Guidelines For Using Nonparametric Methods For Modeling Data From Response Surface Designs

Christine M. Anderson-Cook
Statistical Sciences Group
Los Alamos National Laboratory

Kathryn Prewitt
Mathematics and Statistics
Arizona State University

Traditional response surface methodology focuses on modeling responses using parametric models with designs chosen to balance cost with adequate estimation of parameters and prediction in the design space. Using nonparametric smoothing to approximate the response surface offers both opportunities as well as problems. This article explores some conditions under which these methods can be appropriately used to increase the flexibility of surfaces modeled. The Box and Draper (1987) printing ink study is considered to illustrate the methods.

Key words: Edge-Corrections, data sparseness, bandwidth, lowess, Nadaraya-Watson

Introduction

In his review of the current status and future directions in response surface methodology, Myers (1999) suggests that one of the new frontiers is to utilize nonparametric methods for response surface modeling. Explored in this article are some of the key issues influencing the success of these methods used together. Combining nonparametric smoothing approaches, which typically depend on space-filling samples of points in the desired prediction region, with response surface designs, which primarily focus on an economy of points for adequate prediction of prespecified parametric models, presents some unique challenges. Nonparametric approaches are typically used either as an exploratory data analytic tool in conjunction with a parametric method or exclusively because a parametric model didn't provide the necessary sensitivity to curvature.

The number and location of design points impose a limitation on the order of the polynomial the parametric model can accommodate. This, in turn, imposes a limitation on the type of curvature of the fitted model. Standard response surface techniques using parametric models often assume a quadratic model. Nonparametric techniques assume a certain amount of smoothness, but do not impose a form for the curvature of the target function. Local polynomial models which fit a polynomial model within a window of the data can pick up important curvature, which a parametric fit typically cannot. Issues of what designs are suitable for utilizing nonparametric methods, appropriate choices of smoother types as well as bandwidth considerations will all be discussed. Important limitations exist for incorporating these methods into surface modeling, because ill-defined or nonsensical models can easily be generated without careful consideration of how to blend the method and design.

Vining and Bohn (1998) utilized the Gasser-Mueller estimator (G-M) (see Gasser & Mueller, 1984) which is a kernel based smoothing method to estimate the process variance for a dual response system for the Box and Draper (1987) printing ink study. In that study, a full 3^3 factorial design was used with three replicates per combination of factors. Each

Christine Anderson-Cook is a Technical Staff Member at Los Alamos National Laboratory. Her research interests include response surface methodology and design of experiments. Email: c-and-cook@lanl.gov. Kathryn Prewitt is Associate Professor at Arizona State University. Her research interests include nonparametric function estimation methods, time series and goodness-of-fit. Email: kathryn.prewitt@asu.edu

variable was considered in the range [-1, 1] for the coded variables. Dual models were developed to find an optimal location by modeling both the mean of the process, which has a desired target value of 500, and the variance of the process, which ideally would be minimized. Using a parametric model for the mean and a nonparametric model for the variance, Vining and Bohn (1998) obtained a location in the design space with a substantially improved estimate mean square error (MSE) over parametric models for both mean and variance presented by Del Castillo and Montgomery (1993) and Lin and Tu (1995). The estimated MSE was the chosen desirability function for simultaneously optimizing the mean and variances of the process.

The Box and Draper (1987) example has some interesting features that suggest consideration of nonparametric methods for modeling the variability of the data set. Because the mean response has been carefully studied and appears to be relatively straightforward to model, the focus is on the characteristics and modeling of the standard deviation. This is only half of the problem for the dual modeling approach, but the nonparametric issues here are many.

First, an overview of the characteristics of that part of the data set is provided. Figure 1 shows a plot of the 27 estimates of the standard deviation at the 3^3 factorial locations. Clearly, there is no easily discernible pattern in this response such as a simple function of the three factors. In addition, the range of the data should give us some concern. Within the range of the experimental design space, the standard deviation varies from a value of 0 (all three observations at each of (-1,-1,0) and (0,0,0) were measured to be exactly the same) to a value of 158.2 at (1,0,1).

This should alert one to a possible problem immediately as this range occurring in an actual process seems extreme. Figure 2 shows several different ranges of response standard deviations from 1 to 20. It is uncertain as to what a maximal proportional difference between minimum and maximum variance should be, however, a 1:10 or 1:20 ratio already seems excessive for most well-controlled industrial processes. Hence, one of the goals of the modeling should likely be to moderate this range of observed variability to more closely reflect what is believed to be realistic for the actual process.

If the modeling undersmoothes the data (approaching interpolating between observed points), a risk exists of basing the dual response optimization on non-reproducible idiosyncrasies of the data. If the data is oversmoothed, important curvature is flattened making it difficult to find the best location for the process. This perpetual problem of modeling is doubly important here as the results of the model are being used to determine weights for the modeling of the mean of the process as well as for the optimization of the global process through the dual modeling paradigm. Hence, as different models for the variability are considered, predicted ranges will be noted throughout the design space.

Reviewed in this article are some of the basics of nonparametric methods and their implications for the designed experiment are discussed with limited sample size and structured layout of design points. Then compared are different nonparametric approaches to the existing parametric choices and those presented in Vining and Bohn (1998) for this particular example, and conclude with some general recommendations for how to sensibly and appropriately use nonparametric methods for response surface designs

Smoothing Methods

Smoothing methods are distinct from traditional response surface parametric modeling in that they use different subsets of the data and different weightings for the selected points at different locations in the design space. There are several popular nonparametric smoothing methods such as the Nadaraya-Watson (Nadaraya, 1964) and Watson (1964) which fits a constant to the data in a window, the Gasser-Mueller (Gasser & Mueller, 1984) which is a convolution-type estimator, spline smoothing (Eubank, 1999), and local polynomial methods (Fan & Gijbels, 1996) which fit polynomials in the local data window.

Figure 1: Printing Ink standard deviation raw data

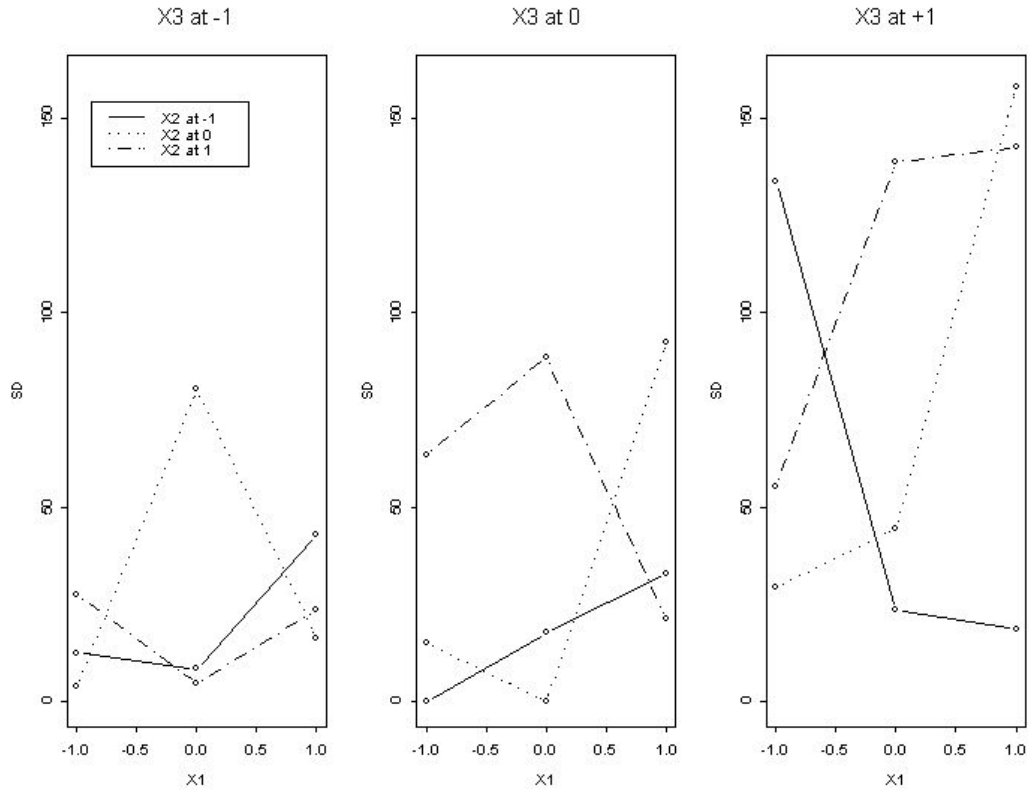
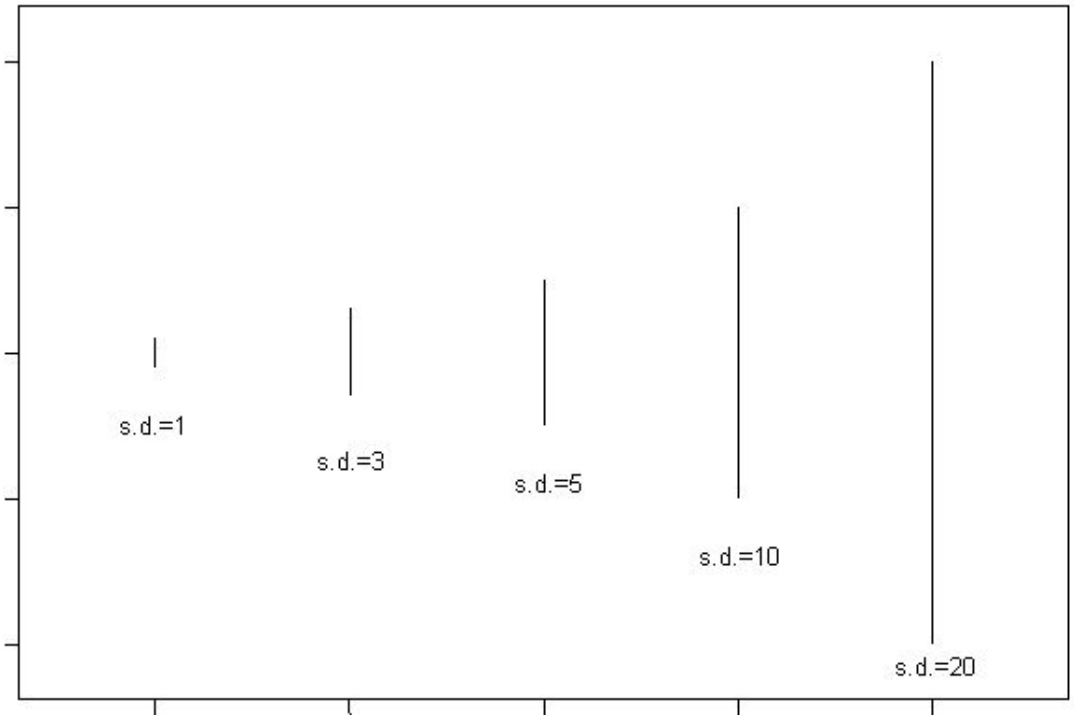


Figure 2: Range of observed responses likely with different values for a variety of standard deviations



The local polynomial methods (lowess) have several positive properties for this problem. Generally, these methods fit a polynomial within a window of data determined by a bandwidth with the kernel function providing weights for the points. The estimators are linear combinations of the responses just as the familiar parametric regression estimators. The Nadaraya-Watson estimator fits a constant in the window so it is a special case of the local polynomial method. When fitting polynomials of order greater than one, these estimators have been shown to naturally account for the bias issues on the boundary (Ruppert & Wand, 1994). In the problem, most of the data (26 of 27 locations) are on the edge, so the concern should be with the behavior of estimators on the boundary.

Typically the nonparametric literature differentiates between behavior on the boundary and behavior in the interior. The variance of this estimator conditioned on the data is unbounded. The unconditional variance of the estimator is actually infinite (Seifert & Gasser, 1996) if the number of points in the window is small (two or less in the local linear, univariate X case). Consequently, the number of points in the window should be greater than two in the univariate X case and in practice greater than the minimum necessary to calculate the estimator. The conditional unbounded variance is due to the fact that the coefficients of the Y_i's in the estimator can be positive or negative).

Problem can be envisioned as all data points are on the edges of a cube with the exception of the point (0, 0, 0). The minimum bandwidth which would include at least 4 data points would be larger than 1 (half of the range of each coded variable) otherwise the number of points in the window would be too small to allow estimation.

Observe *n* independent data points $(\boldsymbol{X_i}, Y_i)$ where $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \cdots, X_{ip})$ are the locations in the design space, and Y_i is the response. The model assuming homoscedastic error for nonparametric function estimation is:

$$Y_i = m(\boldsymbol{X_i}) + \sigma(\boldsymbol{X_i})\varepsilon_i \qquad (1)$$

where $E(\varepsilon_i) = 0$ with $Var(\varepsilon_i) = 1$ and $\sigma^2(x) = Var(Y \mid X = x)$. The smoothing function, *m(.)* is also called the regression function, $E(Y \mid X = x)$. It is assumed that the variance of the error term for the problem of modeling the printing ink standard deviations would be reasonably constant.

Kernel smoothing nonparametric methods involve the choice of a kernel function and a bandwidth as a smoothing parameter which determines the window of data to be utilized in the estimation process. The idea is to weight the data according to its closeness to the target location, hence to estimate $m(x_0)$, greater weight is given to the Y_i values with associated $\boldsymbol{X_i}$ values close to x_0.

Spline smoothing methods are categorized as a nonparametric technique and involve a smoothing parameter but no kernel function. One of the few references to an application of nonparametric methods to response surface problems is Hardy et al. (1997) who explored the use of R-splines with a significantly larger number of design points and with the goal of selecting variables for the regression model rather than obtaining a plausible curve.

There are special considerations when using nonparametric methods for the printing data problem which are next outlined. Most of the literature regarding nonparametric methods shows application to space-filling designs and a larger number of sample points. The printing example has 27 data points which is significantly smaller than the data typically seen in the smoothing literature. Most of these points are on the boundary or edge. It is known that nonparametric estimators can exhibit so-called boundary effects. If a method such as the Gasser-Mueller (Gasser & Mueller, 1984) is used, the bias is bounded but not decreasing with increased sample size as one would want unless kernel functions called boundary kernels are used. This means that a different kernel needs to be used when a point is on the boundary.

Local polynomial methods of order greater than 1 incorporate naturally the boundary kernels necessary. These methods are easily

explained by comparing them to a weighted least squares problem where the kernel function provides the weights and the estimate is provided by solving a familiar looking matrix operation. Most of the nonparametric methods literature provides results and examples for sample sizes much larger than the printing data and also provides leading terms of the bias and variance to describe the behavior of the estimator which implies that there are negligible terms as *n* grows large. The problem then is much different than has been addressed before: the sample size is small, the design is not space-filling and most of the points are on the edge.

Bandwidth issues

One of the most important choices to make when using a nonparametric method of function estimation is the smoothing parameter. For kernel methods, the bandwidth is such a parameter. Large bandwidths provide very smooth estimates and smaller bandwidths produce a more noisy summary of the underlying relationship. The reason for this behavior can be seen in the leading terms of the bias and variance for a point in the interior in the univariate explanatory variable case:

$$Bias(\hat{m}(x)) \approx \tfrac{1}{2} m''(x) b^2 \int K(u) u^2 du$$

and

$$Var(\hat{m}(x)) \approx \frac{\sigma^2 \int K(u)^2 du}{nbf(x)} \qquad (2)$$

where *f(x)* is the density of the **X** explanatory variable, *K(.)* the kernel function, and *b* the bandwidth. The effect of the bandwidth can be observed: large values of the bandwidth increase the bias and reduce the variance of the predicted function; small values decrease the bias and increase the variance. This difficulty is called the bias-variance tradeoff. Bandwidth selection methods can be local (potentially changing at each point at which the function is to estimated) or global (where a single bandwidth is used for the entire curve). Typically, the bandwidth is often chosen to minimize an optimality criterion, such as an estimate of the leading terms of the MSE or cross-validation (see Eubank, 1988, Fan & Gijbels, 1995; Prewitt & Lohr, 2002). The optimality quantities are more accurate when data sets are larger, i.e., the leading terms of the MSE leave out negligible terms which are often not negligible when *n* is small. In simulation studies with bivariate data, sample sizes of n < 50 are not seen. The current problem, on the other hand, involves multivariate data with three explanatory variables and one response with a total of only 27 data points. Minimizing a quantity such as *SSE* where

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{m}(\boldsymbol{X}_i))^2 \qquad (3)$$

cannot be used for the purpose of goodness of fit because without a parametric form for ***m(.)***, *SSE* is minimized with $Y_i = \hat{m}(\boldsymbol{X}_i)$, i.e. the curve estimate which minimizes this quantity is obtained by connecting the points. The purpose of the bandwidth selection method is essentially to solve the bias-variance tradeoff difficulty described previously. The second derivative in the bias term suggests that these estimators typically underestimate peaks and overestimate valleys which is sometimes an argument for using a local bandwidth choice since the expectation would be to use a smaller bandwidth in regions where there are more curvature.

Because the number of points in the problem is small, it would be more sensible to use a global bandwidth, one bandwidth for the entire curve. There are not enough points to justify accurate estimation of different local bandwidths. This is not to say that in the future it may be discovered that in fact different bandwidths should be used to estimate different portions of the surface, but existing methods (Fan & Gijbels, 1995; Prewitt, 2003) will not work. Methods for local bandwidth selection have relied on the fact that each candidate bandwidth for a particular point x_0 will incorporate additional data points as the bandwidth candidates become larger which may not be the case for the problem.

Methodology

Nonparametric Methods for the Sparse Response Surface Designs

Two methods were considered that take into account the special circumstances of the problem as previously outlined. The fitted constant (C) version of the local polynomial (Nadaraya, 1964; Watson, 1964) and the local linear version (LL) (see Fan & Gijbels, 1996) were used. The benefit of fitting these models is that curvature can be achieved without fitting higher order polynomials as is necessary when a completely parametric model is fit. There is the potential to capture different kinds of curvature consistent with what might be reasonable given the nature of the design implemented. The Epanechnikov kernel was used

$$K(u) = 0.75(1-u^2)I(|u| \le 1) \qquad (4)$$

which is simple and has optimal properties (Mueller, 1988).

At the point $\boldsymbol{x} = (x_1, x_2, x_3)$ the weighted least squares estimate with a kernel function as the weight. The two methods can be described as follows where $\hat{m}_C(\boldsymbol{x})$ is the local polynomial with fitted constant: Let the weight function be defined as:

$$K_b(x, X_1) = \tfrac{1}{b^3}\prod_{j=1}^{3} K\left(\frac{x_j - X_{ij}}{b}\right).$$

This is called a product kernel because it is the product of three univariate kernel functions. The kernel function equals zero when data points are outside the window defined by the bandwidth and has a nonzero weight when $\boldsymbol{X_i}$ is inside the window. It is appropriate to use the same bandwidth in each of the three directions because the scaling of the coded variables in the set-up of the response surface design makes units comparable in all directions. The definition below resembles a weighted least squares estimator when a constant is fit.

$$\hat{m}_{NW}(\boldsymbol{x}) = \arg\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_b(\mathbf{x}, \mathbf{X_i})$$
$$= \sum_{i=1}^{n} \frac{K_b(\mathbf{x}, \mathbf{X_i})Y}{\sum K_b(\mathbf{x}, \mathbf{X_i})} \qquad (5)$$

The second method considered is defined below and resembles a weighted least squares estimator where a plane is fit with the data centered at x so that the desired estimator is $\hat{\beta}_0$ and the "LL" stands for local linear with no higher order terms.

$$\hat{m}_{LL}(\boldsymbol{x})$$
$$= \arg\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(x_1 - \mathbf{X_{i1}})$$
$$-\beta_2(x_2 - \mathbf{X_{i2}}) - \beta_3(x_3 - \mathbf{X_{i3}}))^2 K_b(\mathbf{x}, \mathbf{X_i}) \qquad (6)$$

One can also think of the above estimators as motivated by a desire to estimate $m(\boldsymbol{x})$ by using the first few terms of its Taylor expansion, $m_{NW}(\boldsymbol{x})$ is constructed by considering an interval around x and estimating the first term of the Taylor expansion around *x* where $m_{LL}(\boldsymbol{x})$ uses estimates of first order terms of the Taylor expansion as an estimate of $m(\boldsymbol{x})$.

Printing Example Smoothing

It has already been noted that some particular issues concerning the application of nonparametric smoothing to a sparse small set of data with the vast majority of design locations on the edges. A related issue to consider is what type of surface is possible or likely. If the variability of the process can change very quickly and dramatically within the range of the design space, then the 3^3 factorial design is an inadequate choice and should be replaced by a much larger space filling design.

However, if the surface should change moderately slowly throughout the region, then the 3^3 design may be adequate. As well, if the surface is likely to be relatively smooth and undergoes changes slowly, then a nonparametric method should be selected and bandwidth that

uses information from several nearby points to estimate the surface locally. Examined now are some of the implications of choosing different bandwidths for this 3^3 factorial design.

The 3^3 factorial design is comprised of 27 locations on the cube: 8 corner points, 12 edge points, 6 face center points and one center point. Notably, all but one of the points are on the edge of the design space. This is standard practice for parametric estimation, because D- and G-efficiency both benefit from maximal spread of points to the edges of the design space.

However, this set-up coupled with the extreme small sample size is highly unusual for nonparametric approaches. One of the advantages of the structured locations selected for a response surface design is that allows the investigation of the characteristics of estimation for different nonparametric bandwidth choices. For example, using the Epanechnikov kernel weighting function, the number of design points can be specified that will be used for estimation at each of the four categories of design points.

Table 1 shows the effect of bandwidth on different locations as well as the range of the non-zero weights for particular bandwidths used for the local estimation. Bandwidths less than 0.5 of the total range of each variable use only the observation at that location, while a bandwidth of 1 uses all observations. The weights associated with each design location change for different weights. As the bandwidth increases, not only do more locations get used, but also their relative contributions to the estimate become more comparable. For example, for a bandwidth of 0.6 at one of the design points, the observation at the location to be estimated is weighted approximately 35 times more (1.25 / 0.036) than the most distant non-zero weighted observations. As well, for a design point and a bandwidth of 1, this ratio drops to 2.4 (0.75 / 0.316) and the points used are also further away.

Various authors considered different models for the standard deviation for this data set. Parametric models considered include a linear model in all three factors on log(standard deviation +1), shown in Figure 3(a) with an R^2 of 29.4 %. The transformation of the standard deviation was done to improve fit, and to avoid negative predicted values. The range of predicted standard deviation values back on the original scale for this model range from 5.0 to 113.5, which gives a ratio of maximum to minimum standard deviation of 22.7. A full quadratic model for log(standard deviation +1) yields an R^2 of 40.6 % and is shown in Figure 3(b). Here, the ratio of maximum to minimum standard deviation is 25.5 (145.1/5.7).

Fitting the constant (C) and local first-order polynomial (LL) methods for a variety of bandwidths to the data were also considered. Figures 4 (a), and (b), show predicted surfaces for the untransformed standard deviation with the constant C method and bandwidths of 0.8 and 1.0, respectively. Figures 5(a), (b) and (c), show the LL method for the same response and bandwidths of 0.6, 0.8 and 1.0. For each of the parts of the figures, three slices of the design space are shown, with the third factor, C, at the low, middle and high value. Figures 6(a), (b) and 7(a), (b) show the predicted surfaces when modeling using the log(standard deviation +1) response and bandwidths of 0.8 and 1.0. As the bandwidth increases, the surface becomes smoother, reflecting the idiosyncrasies of the data less.

Tables 2 and 3 summarize the ranges of predicted values throughout the design space observed for the different methods for both the untransformed and log(standard deviation +1) responses. The C method tends to moderate the range of the predicted values considerably more than either the parametric or the lowess models. This is due the relative lack of influence of edge effects with extreme values. The transformation to the log-scale does not have a consistent effect on the range of prediction for the different approaches, with it moderating the range of predicted values for only some of the bandwidths. The LL method is susceptible to prediction of larger values near the edges of the design space, with a seeming sensitivity to edge effects. Notably missing from this comparison is the best Vining and Bohn (1998) smoother (Gasser-Mueller), which uses a bandwidth of 0.3 of the total range. As noted in Table 1, this small bandwidth is essentially an interpolator with most regions having only a single observation used for the estimation.

Table 1: Number of points contributing to local estimation for 3^3 factorial, with Epanechnikov kernel.

	# of Points Used			Min. & Max. Weights		
Bandwidth	**(.25,,5)**	**(.5, 1)**	**1**	**.6**	**.8**	**1**
Location						
Corner	1	8	27	(0.036,1.250)	(0.212,0.938)	(0.316,0.750)
Edge	1	12	27			
Face Center	1	18	27			
Center	1	27	27			
(0,0,.5)	2	18	27	(0.096,1.033)	(0.042,0.846)	(0.185,0.703)
(0,.5,1)	2	12	27	(0.096,1.033)	(0.042,0.846)	(0.185,0.703)
(.5,.5,.5)	8	8	27	(0.705,0.705)	(0.002,0.689)	(0.063,0.618)

Table 2: Summary of Prediction Values for Lowess and Local Average on Untransformed Standard Deviations.

	Lowess			Local Average		
Bandwidth	**Minimum**	**Maximum**	**Ratio**	**Minimum**	**Maximum**	**Ratio**
.6	3.1	149.9	48.4	9.0	117.3	13.0
.8	4.8	145.5	30.3	14.1	100.8	7.1
1	7.2	147.3	20.5	15.9	94.8	6.0

Table 3: Summary of Prediction Values for Lowess and Local Average on Transformed Log (Standard Deviations+1).

	Lowess			Local Average		
Bandwidth	**Minimum**	**Maximum**	**Ratio**	**Minimum**	**Maximum**	**Ratio**
.8	5.5	192.0	34.9	7.4	70.3	9.5
1	6.5	227.2	35.0	8.0	60.9	7.6

Figure 3: Contour plots for best Linear and Quadratic parametric models based on the Box and Draper (1987) data for log (standard deviations + 1).

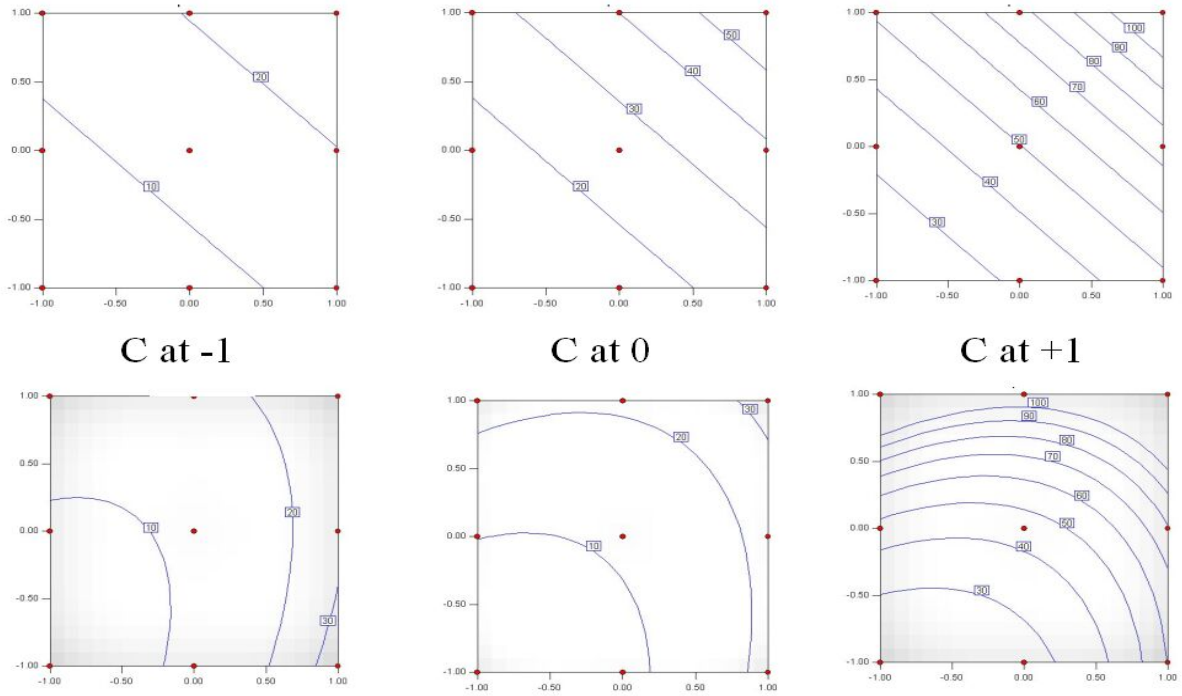
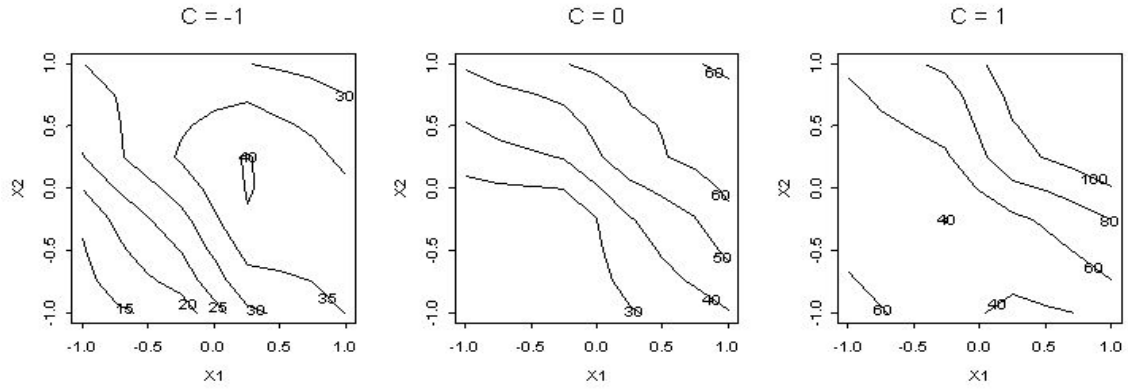


Figure 4: Contour plots of local average models for untransformed response with bandwidths 0.8 and 1.

(a)



(b)

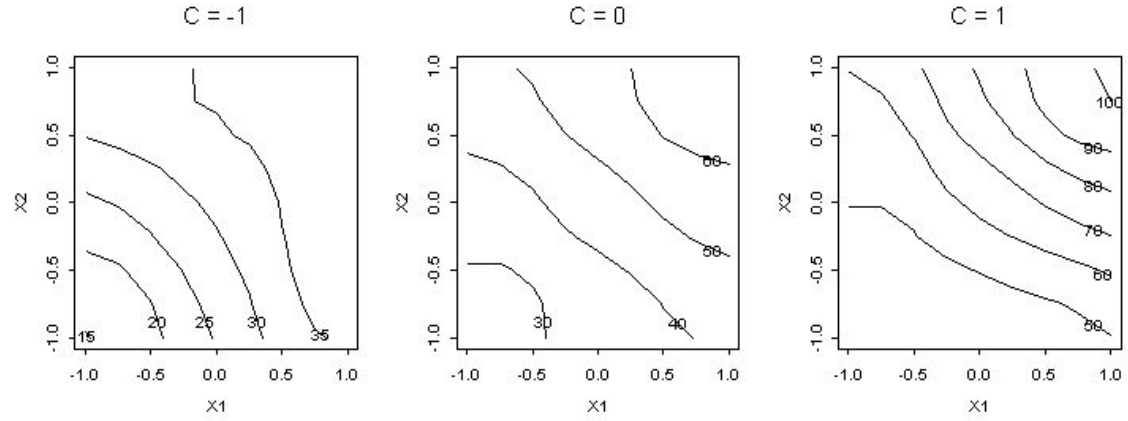
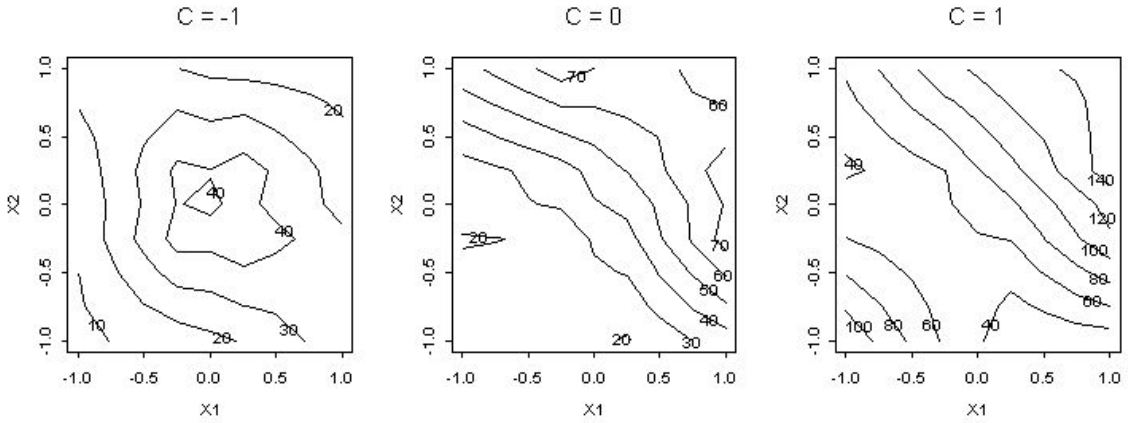
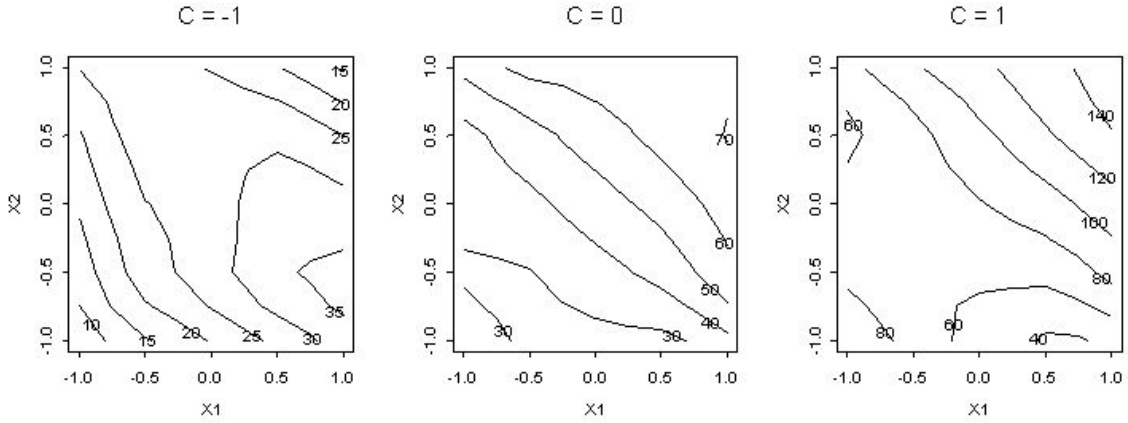


Figure 5: Contour plots of lowess models for untransformed response with bandwidths 0.6, 0.8 and 1.

(a)



(b)



(c)

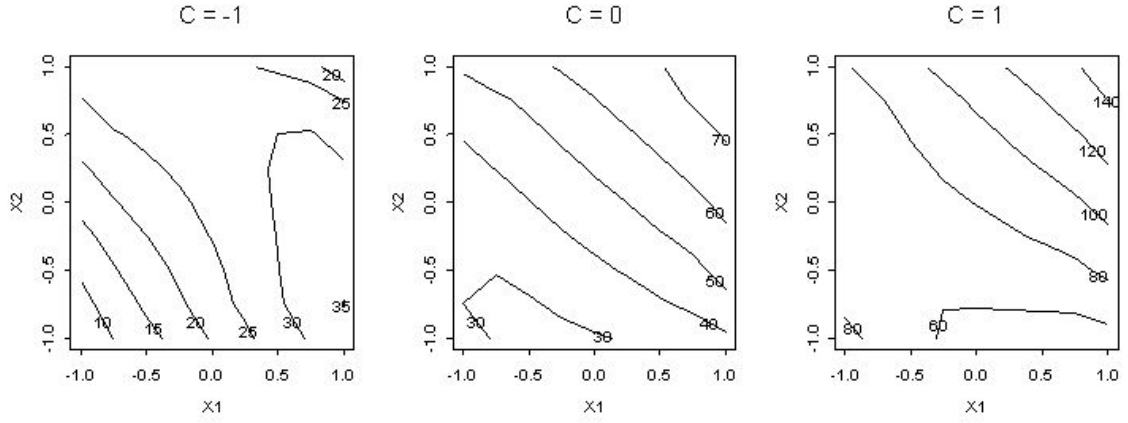
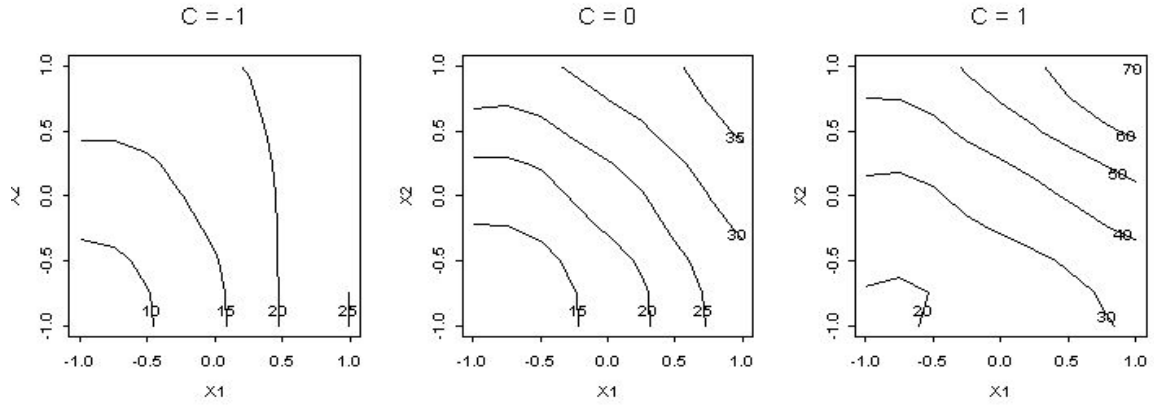


Figure 6: Contour plots of local average models for logarithm transformed response with bandwidths 0.8 and 1.

(a)



(b)

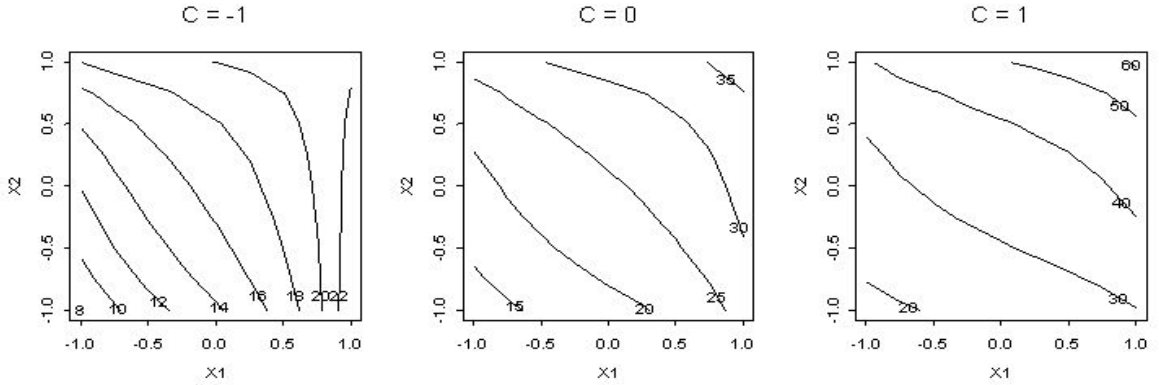
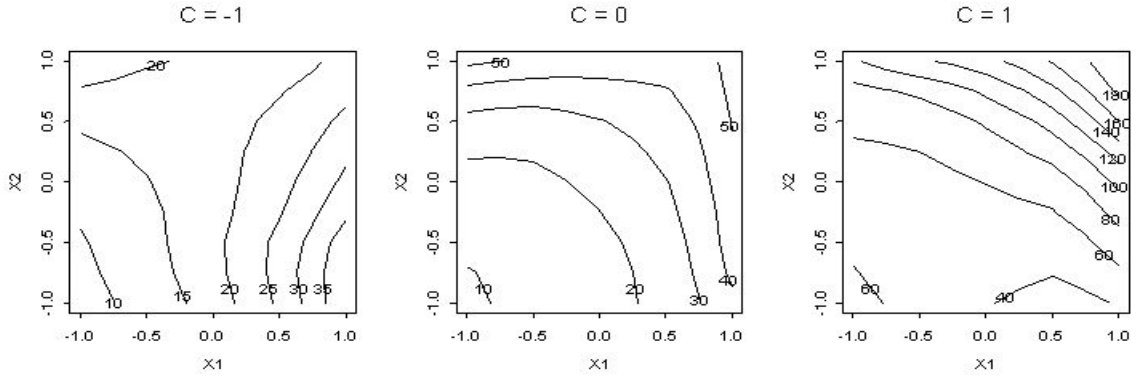
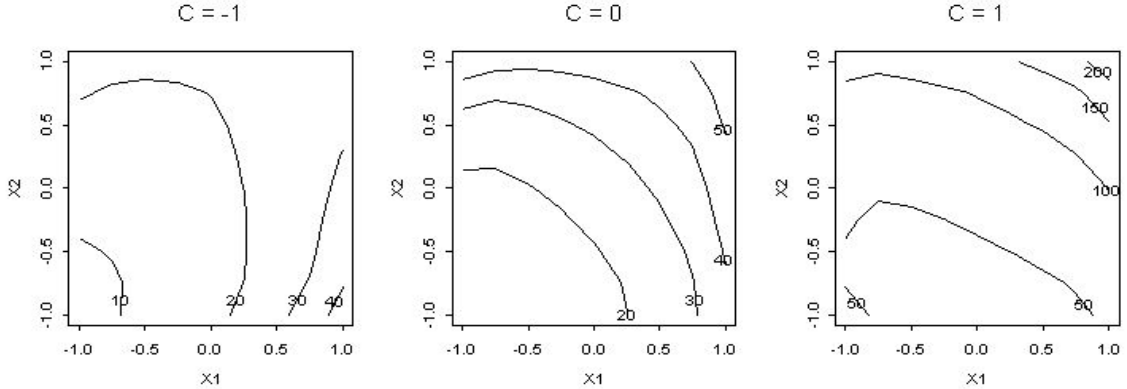


Figure 7: Contour plots of lowess models for logarithm transformed response with bandwidths 0.8 and 1.

(a)



(b)



The fit was next considered by comparing the R^2 values for the different methods in Table 4. Unlike parametric models, where minimizing the R^2 is desirable, here the goal is to obtain a good fit without merely interpolating between points. This is a particularly appropriate strategy given the extreme ranges of values for the standard deviation observed. Also reported are the cross-validation R^2 values which were obtained by removing a single observation, refitting the model with 26 points, and then calculating the difference between the predicted value and what was observed.

Typically in other regression settings, this is a way of measuring the robustness of the model for future prediction. However, in this case, with such a small sparse data set, removing a single point (almost all of which are on the edge of the design space) has the result of leading us to do extensive extrapolation to obtain the new predicted values. As a result, the values obtained were very discouraging. For a number of the cases, including the quadratic parametric model and the small bandwidth LL method, negative R^2 values were obtained, which imply that the model has predicted less well than just using a constant for the entire surface. Again, the structure of the data and the extreme amount of extrapolation involved in this calculation should be considered in interpreting these values. The 0.3 bandwidth Vining and Bohn (1998) smoother cannot be considered in this comparison, because an empty region in the design space was obtained for all of the points, which does not allow the cross-validation R^2 value to be calculated.

However, there are a few general conclusions that can be reached. First, one should be quite cautious with any of these models. Due to the sparsity of the data, they can be influenced considerably by a single value. Secondly, larger bandwidths give lower R^2 values, but generally perform better under the challenges of the cross-validation assessment. Finally, the LL method appears to outperform the C method for the R^2 values, but consistently underperform C for the cross-validation R^2. This reflects the sensitivity to edge effects of this method, which either yields good responsiveness if using the values near the edge, or wide extrapolation when this point is removed. This seems to imply some superiority for the C method, which outperforms both the parametric models, and appears to retain some useful predictive ability even when used for extrapolation.

Based on an overall assessment of all characteristics of the methods considered, the Nadaraya-Watson local averaging (C) method with bandwidth of either 0.8 or 1.0 emerge as leading choices. The bandwidth of 1.0 uses all of the data, with diminishing weights for more distant points. The 0.8 bandwidth excludes points on the opposite side of the design space for corner, edge and face-center points. Both of these models allow for greater flexibility than either of the parametric models, by allowing greater adaptability of the shape of the surface, while also utilizing a significant proportion of the data for estimation. They provide enough smoothing to produce a surface that likely is consistent with underlying assumptions of how the standard deviation of the process might vary across the range of the design space

Conclusion

Based on sparseness of the data sets typical for many response surface designs, it should be evident that the use of nonparametric methods must be used with care to avoid nonsensical results. However, the printing ink example has demonstrated that nonparametric models have real potential for helping with modeling responses, when the restrictions of a parametric model are too limiting. The ability to adapt the shape of the surface locally is desirable, and can be done even when there are only a small number of values observed across the range of each variable. It is particularly important to consider a priori what the surface, range and ratio of maximum to minimum predicted values reasonably might be. The chosen method should balance optimizing fit, while still maintaining characteristics of the appropriate shape.

Table 4: Fit of models to Log (Standard Deviation +1) response.

Model	R^2 (in %)			Cross-validation R^2 (in %)		
Linear	29.4			8.2		
Quadratic	40.6			**-47.0**		
Bandwidth	**.6**	**.8**	**1**	**.6**	**.8**	**1**
Lowess	83.6	67.6	61.8	**-11.0**	**-1.6**	1.0
Local Ave	65.7	42.5	35.7	12.7	13.4	12.8

Table 5: Number of points contributing to local estimation for different Central Composite Designs and widths.

Design	Bandwidth	Corner	Axial	Center Run ($n_c = 1$)
CCD **k=2**, $\alpha = \sqrt{2}$ **n=9**	<0.353	1	1	1
	0.353 < b < 0.5	4	3	5
	0.5 < b < 0.707	4	6	9
	0.707 < b < 1.0	7	6	9
CCD **k=3**, $\alpha = \sqrt{3}$ **n=15**	<0.289	1	1	1
	0.289 < b < 0.5	5	5	9
	0.5 < b < 0.578	5	10	15
	0.578 < b < 1.0	12	10	15
CCD **k=4**, $\alpha = 2$ **n=25**	< 0.25	1	1	1
	0.25 < b < 0.5	6	9	25
	0.5 < b < 1	21	16	25

Due to a large number of points on the edge of the design space, which is highly desirable for D- and G-efficiency when using a parametric model, a smoother which is insensitive to edge effects is recommended. The local averaging smoother (C) performed quite well although the (LL) supposedly has superior boundary capability in the bias term both in order and boundary kernel adjustment. The reason for this apparent contradiction may be again that the sample size is small and the boundary order results depend on larger sample sizes or as pointed out in Ruppert and Wand (1994) the boundary variance of the (LL) may be larger than the boundary variance of the (C) estimator. Consequently the local averaging (C) estimator is recommended for this problem.

The local first-order polynomial works well in many standard applications, where the proportion of edge points is small, but does not seem like a suggested choice for most response surface designs.

To avoid near-interpolation, a moderate to large bandwidth needs to be used. Table 5 considers perhaps the most popular class of response surface designs, the Central Composite Design. It gives the number of points used for estimation for the different types of points for a

number of different bandwidths. A bandwidth of size less than 0.5, or half the range of the coded variables, yield estimates for some of the points using only a small number of observations.

By coupling moderate to large bandwidths with the Epanechnikov kernel, it is possible to downweight but not eliminate the contribution of more distant points, and hence a balance between local adaptivity and moderating extreme values is retained.

Symmetric designs, such as 3^k factorials and Central Composite, are likely to perform better than non-symmetric designs, like Box-Behnken or fractional factorial designs (with some corners of the design space unexplored). While the non-symmetric design performs well for parametric models, the surface will be disproportionately poorly estimated in some regions.

Given the inherent different structure of response surface designs compared to more standard regression studies typically considered in the nonparametric smoothing literature, considerably more research is possible to determine not only reasonable, but optimal smoothing strategies in this context.

References

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response Surfaces*. Wiley.

Del Castillo, E., & Montgomery, D. C. (1993). A nonlinear programming solution to the dual response problem. *Journal of Quality Technology, 25*, 199-204.

Eubank, R. L. (1988). *Spline smoothing and nonparametrie regression.* Marcel Dekker.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing.* Marcel Dekker.

Fan, J., & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B, Methodological, 57*, 371-394.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications (ISBN 031298321)*. Chapman & Hall.

Gasser, T., & Mueller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics, 11*, 171-185.

Hardy, S. W., Nychka, D. W., Haaland, P. D., & O'Connell, M. (1997). Process modeling with nonparametric response surface methods. In *ASA Proceedings of the Section on Physical and Engineering Sciences*, pages 163-172. American Statistical Association (Alexandria, VA).

Lin, D. K. J., & Tu, W. (1995). Dual response surface optimization. *Journal of Quality Technology, 27*, 34-39.

Myers, R. H. (1999). Response surface methodology - Current status and future directions (pkg: P1-74). *Journal of Quality Technology, 31*, 30-44.

Nadaraya, E. (1964). Some new estimates for distribution functions. *Theory of Probability and its Applications (Transl of Teorija Verojatnostei iee Primenenija), 9*, 497-500.

Prewitt, K. (2003). Efficient bandwidth selection in non-parametric regression. *Scandinavian Journal of Statistics. 30, 75-92.*

Prewitt, K., & Lohr, S. (2002). Condition indices and bandwidth choice in local polynomial regression. Under Review.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics, 22*, 1346-1370.

Seifert, B., & Gasser, T. (1996). Finite-sample variance of local polynomials: Analysis and solutions. *Journal of the American Statistical Association, 91*, 267-275.

Vining, G. G., & Bohn, L. L. (1998). Response surfaces for the mean and variance using a nonparametric approach. *Journal of Quality Technology, 30*, 282-291.

Watson, G. (1964). Smooth regression analysis. *Sankhya A26*, 359-372.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 120-133

1538 – 9472/03/$95.00

Determining The Correct Number Of Components To Extract From A Principal Components Analysis: A Monte Carlo Study Of The Accuracy Of The Scree Plot.

Gibbs Y. Kanyongo
Department of Foundations and Leadership
Duquesne University

This article pertains to the accuracy of the of the scree plot in determining the correct number of components to retain under different conditions of sample size, component loading and variable-to-component ratio. The study employs use of Monte Carlo simulations in which the population parameters were manipulated, and data were generated, and then the scree plot applied to the generated scores.

Key words: Monte Carlo, factor analysis, principal component analysis, scree plot

Introduction

In social science research, one of the decisions that quantitative researchers make is determining the number of components to extract from a given set of data. This is achieved through several factor analytic procedures. The scree plot is one of the most common methods used for determining the number of components to extract. It is available in most statistical software such as the Statistical Software for the Social Sciences (SPSS) and Statistical Analysis Software (SAS).

Factor analysis is a term used to refer to statistical procedures used in summarizing relationships among variables in a parsimonious but accurate manner. It is a generic term that includes several types of analyses, including (a) common factor analysis, (b) principal component analysis (PCA), and (c) confirmatory factor analysis (CFA). According to Merenda, (1997) common factor analysis may be used when a primary goal of the research is to investigate how well a new set of data fits a particular well-established model. On the other hand, Stevens (2002) noted that principal components analysis is usually used to identify the factor structure or model for a set of variables. In contrast; CFA is based on a strong theoretical foundation that allows the researcher to specify an exact model in advance. In this article, principal components analysis is of primary interest.

Gibbs Y. Kanyongo is assistant professor in the Department of Foundations and Leadership at Duquesne University. 410A Canevin Hall, Pittsburgh, PA 15282. Email: kanyongog@duq.edu.

Principal component analysis

Principal component analysis develops a small set of uncorrelated components based on the scores on the variables. Tabachnick and Fidell (2001) pointed that components empirically summarize the correlations among the variables. PCA is the more appropriate method than CFA if there are no hypotheses about components prior to data collection, that is, it is used for exploratory work.

When one measures several variables, the correlation between each pair of variables can be arranged in a table of correlation coefficients between the variables. The diagonals in the matrix are all 1.0 because each variable theoretically has a perfect correlation with itself. The off-diagonal elements are the correlation coefficients between pairs of variables. The existence of clusters of large correlation coefficients between subsets of variables suggests that those variables are related and could be measuring the same underlying dimension or concept. These underlying dimensions are called components.

A component is a linear combination of variables; it is an underlying dimension of a set of items. Suppose, for instance a researcher is interested in studying the characteristics of freshmen students. Next, a large sample of freshmen are measured on a number of characteristics like personality, motivation, intellectual ability, family socio-economic status, parents' characteristics, and physical characteristics. Each of these characteristics is measured by a set of variables, some of which are correlated with one another.

An analysis might reveal correlation patterns among the variables that are thought to show the underlying processes affecting the behavior of freshmen students. Several individual variables from the personality trait may combine with some variables from motivation and intellectual ability to yield an independence component. Variables from family socio-economic status might combine with other variables from parents' characteristics to give a family component. In essence what this means is that the many variables will eventually be collapsed into a smaller number of components.

Velicer et. al., (2000) noted that a central purpose of PCA is to determine if a set of *p* observed variables can be represented more parsimoniously by a set of *m* derived variables (components) such that $m < p$. In PCA the original variables are transformed into a new set of linear combinations (principal components). Gorsuch (1983) described the main aim of component analysis as to summarize the interrelationships among the variables in a concise but accurate manner. This is often achieved by including the maximum amount of information from the original variables in as few derived components as possible to keep the solution understandable.

Stevens (2002) noted that if we have a single group of participants measured on a set of variables, then PCA partitions the total variance by first finding the linear combination of variables that accounts for the maximum amount of variance. Then the procedure finds a second linear combination, uncorrelated with the first component, such that it accounts for the next largest amount of variance, after removing the variance attributable to the first component from the system. The third principal component is constructed to be uncorrelated with the first two, and accounts for the third largest amount of variance in the system. This process continues until all possible components are constructed. The final result is a set of components that are not correlated with each other in which each derived component accounts for unique variance in the dependent variable.

Uses of principal components analysis

Principal component analysis is important in a number of situations. When several tests are administered to the same examinees, one aspect of validation may involve determining whether there are one or more clusters of tests on which examinees display similar relative performances. In such a case, PCA functions as a validation procedure. It helps evaluate how many dimensions or components are being measured by a test.

Another situation is in exploratory regression analysis when a researcher gathers a moderate to a large number of predictors to predict some dependent variable. If the number of predictors is large relative to the number of participants, PCA may be used to reduce the number of predictors. If so, then the sample size to variable ratio increases considerably and the possibility of the regression equation holding up under cross-validation is much better (Stevens 2002). Here, PCA is used as a variable reduction scheme because the number of simple correlations among the variables can be very large. It also helps in determining if there is a small number of underlying components, which might account for the main sources of variation in such a complex set of correlations. If there are 30 variables or items, 30 different components are probably not being measured. It therefore makes sense to use some variable reduction scheme that will indicate how the variables or items cluster or "hang" together.

The use of PCA on the predictors is also a way of attacking the multicollinearity problem (Stevens, 2002). Multicollinearity occurs when predictors are highly correlated with each other. This is a problem in multiple regression because the predictors account for the same variance in the dependent variable. This redundancy makes the regression model less accurate in as far as the number of predictors required to explain the

variance in the dependent variable in a parsimonious way is concerned. This is so because several predictors will have common variance in the dependent variable. The use of PCA creates new components, which are uncorrelated; the order in which they enter the regression equation makes no difference in terms of how much variance in the dependent variable they will account for.

Principal component analysis is also useful in the development of a new instrument. A researcher gathers a set of items, say 50 items designed to measure some construct like attitude toward education, sociability or anxiety. In this situation PCA is used to cluster highly correlated items into components. This helps determine empirically how many components account for most of the variance on an instrument. The original variables in this case are the items on the instrument.

Stevens (2002) pointed out several limitations (e.g., reliability consideration and robustness) of the *k* group MANOVA (Multivariate Analysis of Variance) when a large number of criterion variables are used. He suggests that when there are a large number of potential criterion variables, it is advisable to perform a PCA on them in an attempt to work with a smaller set of new criterion variables.

The scree plot

The scree plot is one of the procedures used in determining the number of factors to retain in factor analysis, and was proposed by Cattell (1966). With this procedure eigenvalues are plotted against their ordinal numbers and one examines to find where a break or a leveling of the slope of the plotted line occurs. Tabachnick and Fidell (2001) referred to the break point as the point where a line drawn through the points changes direction. The number of factors is indicated by the number of eigenvalues above the point of the break. The eigenvalues below the break indicate error variance. An eigenvalue is the amount of variance that a particular variable or component contributes to the total variance. This corresponds to the equivalent number of variables that the component represents. Kachigan, (1991) provided the following explanation: a component associated with an eigenvalue of 3.69 indicates that the component accounts for as much variance in the data collection as would 3.69 variables on average. The concept of an eigenvalue is important in determining the number of components retained in principal component analysis.

The scree plot is an available option in most statistical packages. A major weakness of this procedure is that it relies on visual interpretation of the graph. Because of this, the scree plot has been accused of being subjective. Some authors have attempted to develop a set of rules to help counter the subjectivity of the scree plot. Zoski and Jurs (1990) presented rules for the interpretation of the scree plot. Some of their rules are: (a) the minimum number of break points for drawing the scree plot should be three, (b) when more than one break point exists in the curve, the first one should be used, and (c) the slope of the curve should not approach vertical. Instead, it should have an angle of 40 degrees or less from the horizontal.

Previous studies found mixed results on the accuracy of the scree plot. Zwick and Velicer (1986) noted that “the scree plot had moderate overall reliability when the mean of two trained raters was used” (p.440). Cattell and Jaspers (1967) discovered that the scree plot displayed very good reliability. On the other hand, Crawford and Koopman (1979) reported very poor reliability of the scree plot.

Monte Carlo study

Hutchinson and Bandalos, (1997) pointed that Monte Carlo studies are commonly used to study the behavior of statistical tests and psychometric procedures in situations where the underlying assumptions of a test are violated. They use computer-assisted simulations to provide evidence for problems that cannot be solved mathematically. Robey and Barcikowski (1992) stated that in Monte Carlo simulations, the values of a statistic are observed in many samples drawn from a defined population.

Monte Carlo studies are often used to investigate the effects of assumption violations on statistical tests. Statistical tests are typically developed mathematically using algorithms based on the properties of known mathematical distributions such as the normal distribution. Hutchinson and Bandalos, (1997) further noted

that these distributions are chosen because their properties are understood and because in many cases they provide good models for variables of interest to applied researchers. Using Monte Carlo simulations in this study has the advantage that the population parameters are known and can be manipulated; that is, the internal validity of the design is strong although this will compromise the external validity of the results.

According to Brooks et al. (1999), Monte Carlo simulations perform functions empirically through the analysis of random samples from populations whose characteristics are known to the researcher. That is, Monte Carlo methods use computer assisted simulations to provide evidence for problems that cannot be solved mathematically, such as when the sampling distribution is unknown or hypothesis is not true.

Mooney, (1997) pointed that the principle behind Monte Carlo simulation is that the behavior of a statistic in a random sample can be assessed by the empirical process of actually drawing many random samples and observing this behavior. The idea is to create a pseudo-population through mathematical procedures for generating sets of numbers that resemble samples of data drawn from the population.

Mooney (1997) further noted that other difficult aspects of the Monte Carlo design are writing the computer code to simulate the desired data conditions and interpreting the estimated sampling plan, data collection, and data analysis. An important point to note is that a Monte Carlo design takes the same format as a standard research design. This was noted by Brooks et al., (1999) when they wrote "It should be noted that Monte Carlo design is not very different from more standard research design, which typically includes identification of the population, description of the sampling plan, data collection and data analysis" (p. 3).

Methodology

Sample size (n)

Sample size is the number of participants in a study. In this study, sample size is the number of cases generated in the Monte Carlo simulation. Previous Monte Carlo studies by (Velicer et al. 2000, Velicer and Fava, 1998, Guadanoli & Velicer, 1988) found sample size as one of the factors that influences the accuracy of procedures in PCA. This variable had three levels (75, 150 and 225). These values were chosen to cover both the lower and the higher ends of the range of values found in many applied research situations.

Component loading (a_{ij})

Field (2000) defined a component loading as the Pearson correlation between a component and a variable. Gorsuch, 1983 defined it as a measure of the degree of generalizability found between each variable and each component. A component loading reflects a quantitative relationship and the further the component loading is from zero, the more one can generalize from that component to the variable. Velicer and Fava, (1998), Velicer et al., (2000) found the magnitude of the component loading to be one of the factors having the greatest effect on accuracy within PCA. This condition had two levels (.50 and .80). These values were chosen to represent a moderate coefficient (.50) and a very strong coefficient (.80).

Variable-to-component ratio (p:m)

This is the number of variables per component. The number of variables per component will be measured counting the number of variables correlated with each component in the population conditions. The number of variables per component has repeatedly been found to influence the accuracy of the results, with more variables per component producing more stable results. Two levels for this condition were used (8:1 and 4:1). Because the number of variables in this study was fixed at 24, these two ratios yielded three and six variables per factor respectively.

Number of variables

This study set the number of variables a constant at 24, meaning that for the variable-to-component ratio of 4:1, there were six variables loading onto one component, and for variable-to-component ratio of 8:1, eight variables loaded onto a component (see Appendixes A to D).

Generation of population correlation matrices

A pseudo-population is an artificial population from which samples used in Monte Carlo studies are derived. In this study, the underlying population correlation matrices were generated for each possible a_{ij} and p:m combination, yielding a total of four matrices (see Appendixes E to H).

The population correlation matrices were generated in the following manner using RANCORR programme by Hong (1999):

1. The factor pattern matrix was specified based on the combination of values for p:m and a_{ij} (see Appendixes A to D).
2. After specifying the factor pattern matrix and the program is executed, a population correlation matrix was produced for each combination of conditions.
3. The program was executed four times to yield four different population correlation matrices, one correlation matrix for each combination of conditions (see Appendixes E to H).

After the population correlation matrices were generated, the Multivariate Normal Data Generator (MNDG) program (Brooks, 2002) was used to generate samples from the population correlation matrices. This program generated multivariate normally distributed data. A total of 12 cells were created based on the combination of n, p: m and a_{ij}. For each cell, 30 replications were done to give a total of 360 samples, essentially meaning that 360 scree plots were generated. Each of the samples had a pre-determined factor structure since the parameters were set by the researcher. The scree plots were then examined to see if they extracted the exact number of components as set by the researcher.

Interpretation of the scree plots

The scree plots were given to two raters with some experience in interpreting scree plots. These raters were graduate students in Educational Research and Evaluation and had taken a number of courses in Educational Statistics and Measurement.

First, the raters were asked to look at the plots independently to determine the number of components extracted. Second, they were asked to interpret the scree plots together. The raters had no prior knowledge of how many components were built into the data. The accuracy of the scree plot was measured by how many times it extracted the exact number of components.

Results

The first research question of the study is: How accurate is the scree plot in determining the correct number of components? This question was answered in two parts. First, this question was answered by considering the degree of agreement between the two raters. Table 1 is of the measure of agreement between the two raters when component loading was .80. To interpret Table 1, the value of 1 indicates a correct decision and a value of 0 indicates a wrong decision by the raters as they interpreted the scree plots. A correct decision means that the scree plot extracted the correct number of components (either three components for 8:1 ratio or six components for 4:1). Thus, from Table 1, the two raters agreed correctly 108 of the times while they agreed wrongly 52 times.

Table 1. A cross tabulation of the measure of agreement when component loading was .80 between rater 1 and rater 2.

		Rater 2		
		0	1	Total
Rater 1	0	52	11	63
	1	9	108	117
Total		61	119	180

An examination of Figures 1 and 2 show that when component loading was .80, it was relatively clear where the cut-off point was for determining the number of components to extract. Figure 1 clearly shows that six the

Figure 1. The scree plot for variable-to-component ratio of 4:1, component loading of .80

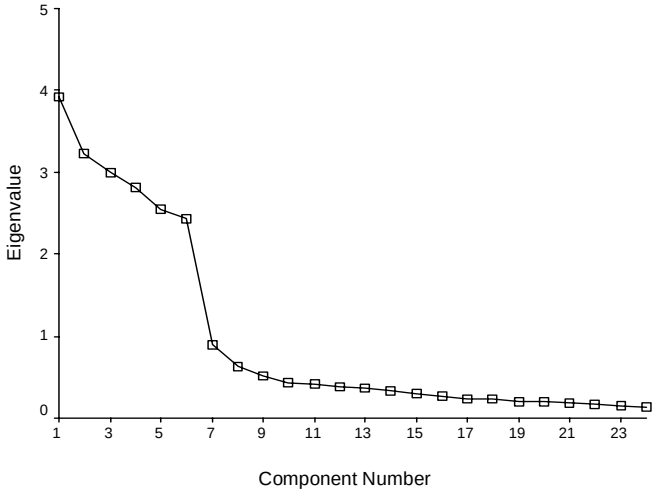
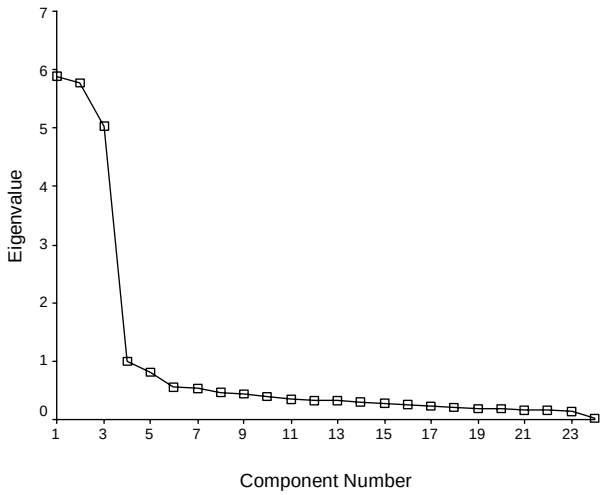


Figure 2. The scree plot for variable-to-component ratio of 8:1, component loading of .80



components were extracted and in Figure 2, three components were extracted. These two plots show why it was easy for the raters to have more agreement for component loading of .80. This was not the case when component loading was .50 as the raters had few cases of agreement and more cases of disagreement.

In Table 2, when component loading was .50, the two raters agreed correctly only 28 times and agreed wrongly 97 times. Compared to component loading of .80, the scree plot was not as accurate when component loading was .50. This finding is consistent with that of Zwick and Velicer (1986) who noted in their study that, "The raters in this study showed greater agreement at higher than at lower component

loading levels." (p. 440). Figures 1 and 2 show typical scree plots that were obtained for component loading of .50. In Figure 3, the number of components extracted was supposed to be six, but it is not clear from the plot were the cut-off point is for six components. One can see why there were a lot of disagreements between the two raters when component loading was low. In Figure 4, the plot was supposed to extract three components, but it is not quite clear even to an experienced rater, how many components to be extracted with this plot. These cases show how it is difficult to use the scree plots especially in exploratory studies when the researcher does not know the number of components that exist.

Table 2. A cross tabulation of the measure of agreement when component loading was .50 between rater 1 and rater 2.

		Rater 2		
		0	1	Total
Rater 1	0	97	50	147
	1	5	28	33
Total		102	78	180

Reports of rater reliability on the scree plot have ranged from very good (Cattell & Jaspers, 1967) to quite poor (Crawford & Koopman, 1979). This wide range and the fact that data encountered in real life situations rarely have perfect structure with high component loading makes it difficult to recommend this procedure as a stand-alone procedure for practical uses in determining the number of components. Generally, most real data have low to moderate component loading, which makes the scree plot an unreliable procedure of choice (Zwick & Velicer, 1986).

The second part of question one was to consider the percentages of time that the scree plots were accurate in determining the exact number of components, and those percentages were computed for each cell (see table 5). In Table 5, results of the two raters are presented according to variable-to-component ratio, component loading and sample size. The table shows mixed results of the interpretation of the scree plot by the two raters. However, the scree plot appeared to do well when component loading was high (.80) with a small number of variables (three). When variable-to-component ratio was 8:1 and component loading was .80, the scree plot was very accurate. The lowest performance of the scree plot in this cell was 87% for a sample size of 75. On the other hand, when variable-to-component ratio was 4:1, component loading was .80, and sample size was 225, the scree plot was only accurate 3% of the time with rater 1. With rater 2 under the same conditions, the scree plot was correct 13% of the time.

The second question was: Does the accuracy of the scree plot change when two experienced raters interpret the scree plots together? For this question, percentages were computed of how many times the two raters were correct when they interpreted the scree plots together. The results are presented in table 5 in the row Consensus row. These results show that even if two raters work together, the accuracy of the scree plot does not necessarily improve when component loading was .50. When variable-to-component ratio was 8:1 and component loading was .50, rater 2 was actually better than when the two raters worked together. This is again an example of the mixed results obtained by the scree plot which makes it unreliable. On the other hand, the accuracy of the scree plot improved when component loading was .80, and variable-to-component ratio was 4:1. When component loading was .80, and variable-to-component ratio was 8:1, having two rates work together did not change anything since the scree plot was very accurate when the two raters work independently.

The bottom line is in this study, the scree plot produced mixed results and this is mainly due to its subjectivity. Although it was 100% accurate under certain conditions, it was also terrible under other conditions. It however emerged from this study that the accuracy of the scree plot improves when the component loading is high, and the number of variables per component is few.

Figure 3. The scree plot for variable-to-component ratio of 4:1, component loading of .50

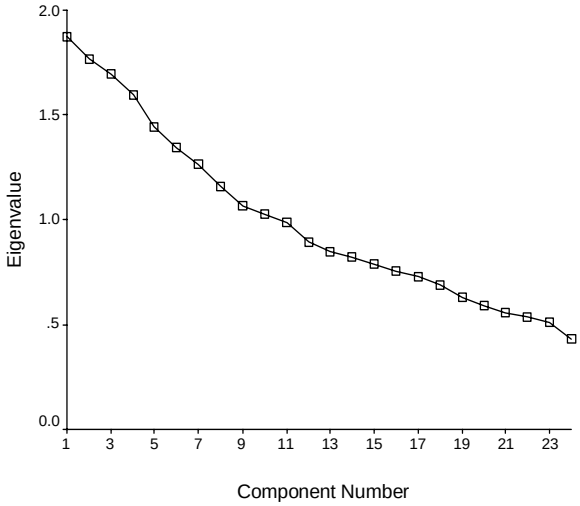


Figure 4. The scree plot for variable-to-component ratio of 8:1, component loading of .50

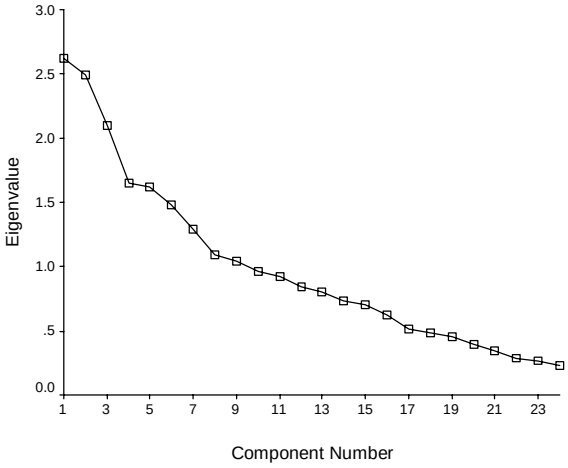


Table 5. Performance of scree plot (as a percentage) under different conditions of variable-to-component ratio, component loading and sample size.

V-C-R	**4: 1**						**8: 1**					
Comp.loading	**.50**			**.80**			**.50**			**.80**		
Sample size	75	150	225	75	150	225	75	150	225	75	150	225
Rater 1	73%	20%	10%	10%	16%	3%	33%	13%	27%	87%	100%	100%
Rater 2	67%	10%	16%	26%	23%	13%	63%	57%	77%	100%	100%	100%
Consensus	23%	20%	10%	75%	100%	100%	47%	23%	47%	100%	97%	100%

Conclusion

Generally, the findings of this study are in agreement with previous studies that found mixed results on the scree plot. The subjectivity in the interpretation of the procedure makes it such an unreliable procedure to use as a stand-alone procedure. The scree plot would probably be useful in confirmatory factor analysis to provide a quick check of the factor structure of the data. In that case the researcher already knows the structure of the data as opposed to using it in exploratory studies where the structure of the data is unknown. If used in exploratory factor analysis, the scree plot can be misleading even for experienced researcher because of its subjectivity.

Based on the findings of this study, it is recommended that the scree plot not be used as a stand-alone procedure in determining the number of components to retain. Researchers should use it with other procedures like parallel analysis or Velicer's Minimum Average Partial (MAP) and parallel analysis. In situations where the scree plot is the only procedure available, users should be very cautious in using it and they can do so in confirmatory studies but not exploratory studies.

References

Brooks, G. P. (2002). MNDG. http://oak.cats.ohiou.edu.edu/~brooksg/mndg.htm.

Brooks, G. P., Barcikowski, R. S., & Robey, R. R. (1999). *Monte Carlo simulation for perusal and practice*. A paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC Document Reproduction Service No. ED449178).

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cattell, R. B., & Jaspers, J. (1967). A general plasmode for factor analytic exercises and research. *Multivariate Behavioral Research Monographs, 3,* 1-212.

Crawford, C. B., & Koopman, P. (1973). A note on Horn's test for the number of factors in factor analysis. *Multivariate Behavioral Research, 8*, 117-125.

Field, A. (2000). *Discovering statistics using SPSS for Windows*. London: UK. Sage.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments & Computers, 31*, 727-730.

Hutchinson, S. R., & Bandalos, D. L. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research, 22*, 233-245.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York: Radius Press.

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika, 33*, 37-71.

Merenda, F. P. (1997). A Guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development, 30*, 156-164.

Mooney, C. Z. (1997). *Monte Carlo simulation* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: Sage.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments & Computers, 32*, 396-402.

Robey, R. R., & Barcikowski, R. S. (1992). Type 1 error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. Mahwah. NJ: Lawrence Erlbaum Associates.

Steiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports, 83*, 689-694.

Tabachnick , B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.) Needham Heights., MA: Pearson.

Velicer, F. W., Eaton, C. A., & Fava, J. L. (2000). Construct Explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 42 -71). Boston: Kluwer Academic Publishers.

Zoski, K. W., & Jurs, S. (1990). Priority determination in surveys: an application of the scree test. *Evaluation Review,* 14, 214-219.

Zwick, R. W., & Velicer, F. V. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Appendix:

Appendix A : Population Pattern Matrix $p:m$ = 8:1 (p = 24, m = 3 a_{ij} = .80).

	Components (m)		
p	1	2	3
1	**.80**	.00	.00
2	**.80**	.00	.00
3	**.80**	.00	.00
4	**.80**	.00	.00
5	**.80**	.00	.00
6	**.80**	.00	.00
7	**.80**	.00	.00
8	**.80**	.00	.00
9	.00	**.80**	.00
10	.00	**.80**	.00
11	.00	**.80**	.00
12	.00	**.80**	.00
13	.00	**.80**	.00
14	.00	**.80**	.00
15	.00	**.80**	.00
16	.00	**.80**	.00
17	.00	.00	**.80**
18	.00	.00	**.80**
19	.00	.00	**.80**
20	.00	.00	**.80**
21	.00	.00	**.80**
22	.00	.00	**.80**
23	.00	.00	**.80**
24	.00	.00	**.80**

Appendix B: Population Pattern Matrix p:m = 8:1 (p = 24, m = 3 aij = .50).

	Components (*m*)		
p	1	2	3
1	**.50**	.00	.00
2	**.50**	.00	.00
3	**.50**	.00	.00
4	**.50**	.00	.00
5	**.50**	.00	.00
6	**.50**	.00	.00
7	**.50**	.00	.00
8	**.50**	.00	.00
9	.00	**.50**	.00
10	.00	**.50**	.00
11	.00	**.50**	.00
12	.00	**.50**	.00
13	.00	**.50**	.00
14	.00	**.50**	.00
15	.00	**.50**	.00
16	.00	**.50**	.00
17	.00	.00	**.50**
18	.00	.00	**.50**
19	.00	.00	**.50**
20	.00	.00	**.50**
21	.00	.00	**.50**
22	.00	.00	**.50**
23	.00	.00	**.50**
24	.00	.00	**.50**

Appendix C: Population Pattern Matrix p:m = 4:1 (p = 24, m = 6, aij = .80).

	Components (*m*)					
p	1	2	3	4	5	6
1	**.80**	.00	.00	.00	.00	.00
2	**.80**	.00	.00	.00	.00	.00
3	**.80**	.00	.00	.00	.00	.00
4	**.80**	.00	.00	.00	.00	.00
5	.00	**.80**	.00	.00	.00	.00
6	.00	**.80**	.00	.00	.00	.00
7	.00	**.80**	.00	.00	.00	.00
8	.00	**.80**	.00	.00	.00	.00
9	.00	.00	**.80**	.00	.00	.00
10	.00	.00	**.80**	.00	.00	.00
11	.00	.00	**.80**	.00	.00	.00
12	.00	.00	**.80**	.00	.00	.00
13	.00	.00	.00	**.80**	.00	.00
14	.00	.00	.00	**.80**	.00	.00
15	.00	.00	.00	**.80**	.00	.00
16	.00	.00	.00	**.80**	.00	.00
17	.00	.00	.00	.00	**.80**	.00
18	.00	.00	.00	.00	**.80**	.00
19	.00	.00	.00	.00	**.80**	.00
20	.00	.00	.00	.00	**.80**	.00
21	.00	.00	.00	.00	.00	**.80**
22	.00	.00	.00	.00	.00	**.80**
23	.00	.00	.00	.00	.00	**.80**
24	.00	.00	.00	.00	.00	**.80**

Appendix D: Population Pattern Matrix p:m = 4:1 (p = 24, m = 6, aij = .50).

	Components (*m*)					
p	1	2	3	4	5	6
1	**.50**	.00	.00	.00	.00	.00
2	**.50**	.00	.00	.00	.00	.00
3	**.50**	.00	.00	.00	.00	.00
4	**.50**	.00	.00	.00	.00	.00
5	.00	**.50**	.00	.00	.00	.00
6	.00	**.50**	.00	.00	.00	.00
7	.00	**.50**	.00	.00	.00	.00
8	.00	**.50**	.00	.00	.00	.00
9	.00	.00	**.50**	.00	.00	.00
10	.00	.00	**.50**	.00	.00	.00
11	.00	.00	**.50**	.00	.00	.00
12	.00	.00	**.50**	.00	.00	.00
13	.00	.00	.00	**.50**	.00	.00
14	.00	.00	.00	**.50**	.00	.00
15	.00	.00	.00	**.50**	.00	.00
16	.00	.00	.00	**.50**	.00	.00
17	.00	.00	.00	.00	**.50**	.00
18	.00	.00	.00	.00	**.50**	.00
19	.00	.00	.00	.00	**.50**	.00
20	.00	.00	.00	.00	**.50**	.00
21	.00	.00	.00	.00	.00	**.50**
22	.00	.00	.00	.00	.00	**.50**
23	.00	.00	.00	.00	.00	**.50**
24	.00	.00	.00	.00	.00	**.50**

Appendix E. Population correlation matrix *p:m* = 8: 1 (*p* = 24, *m* = 3, a_{ij} = .80).

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
1.00																							
.620	1.00																						
.569	.674	1.00																					
.613	.697	.698	1.00																				
.562	.679	.750	.732	1.00																			
.626	.655	.708	.711	.740	1.00																		
.706	.606	.565	.579	.542	.585	1.00																	
.686	.602	.579	.568	.543	.573	.716	1.00																
-.04	-.01	.035	-.02	.018	-.02	.000	.019	1.00															
-.01	-.02	.030	.003	.041	.064	-.02	-.04	.637	1.00														
-.02	.006	.049	.046	.066	.058	-.05	-.02	.644	.632	1.00													
-.07	.012	.085	.041	.108	.054	-.07	-.07	.682	.669	.681	1.00												
-.05	.024	.058	.039	.068	.031	-.07	-.04	.653	.623	.700	.693	1.00											
-.05	.020	.102	.107	.157	.148	-.09	-.11	.630	.733	.701	.738	.674	1.00										
-.04	.002	.032	.005	.024	.011	-.05	-.01	.648	.636	.681	.654	.692	.638	1.00									
.000	-.02	-.02	-.04	-.04	-.06	.034	.046	.679	.582	.638	.642	.640	.564	.625	1.00								
.033	.018	-.03	.004	-.04	-.01	.000	.021	-.04	-.04	.027	-.05	.017	-.05	.030	-.01	1.00							
-.03	.011	.001	-.02	-.01	-.06	.011	.013	.045	-.04	-.03	.020	.000	-.05	-.02	.053	.608	1.00						
.002	.033	-.04	-.02	-.06	-.09	.015	.030	.000	-.09	-.02	-.04	.017	-.13	.012	.054	.683	.695	1.00					
.06	-.01	-.04	.001	-.04	.028	.033	.021	-.05	.015	.004	-.05	-.03	.002	-.02	-.03	.675	.588	.610	1.00				
-.04	.038	.034	.046	.060	.017	-.05	-.07	-.01	.006	-.01	.051	.024	.051	-.01	-.02	.608	.663	.648	.608	1.00			
-.03	.015	.000	-.03	-.03	-.07	-.02	.009	.015	-.04	-.01	.000	.036	-.10	.044	.024	.670	.678	.727	.604	.656	1.00		
-.02	-.02	.040	.013	,057	.074	-.02	-.04	.003	.087	.000	.045	-.01	.109	-.01	-.05	.578	.609	.540	.644	.658	.580	1.00	
-.02	.027	.021	.053	.044	.041	-.05	-.05	-.03	.021	.026	.007	.025	.061	.031	-.06	.666	.602	.613	.656	.667	.640	.649	1.00

Appendix F : Population correlation matrix *p*:*m* = 8:1 (*p*= 24, *m*= 3, a_{ij} = .50)

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
1.00																							
.271	1.00																						
.281	.303	1.00																					
.184	.297	.275	1.00																				
.271	.271	.276	.243	1.00																			
.233	.293	.287	.303	.239	1.00																		
.280	.226	.250	.213	.250	.199	1.00																	
.190	.236	.229	.296	.233	.242	.218	1.00																
-.03	-.03	-.02	.005	-.03	.000	-.03	.054	1.00															
-.03	-.04	-.04	-.02	.000	-.05	.007	.041	.276	1.00														
-.04	.065	.038	.096	.008	.072	-.09	.039	.262	.224	1.00													
-.04	-.03	-.03	.016	-.05	-.03	.025	.049	.293	.283	.234	1.00												
.018	-.05	-.02	-.06	-.02	-.03	.077	-.05	.228	.244	.149	.253	1.00											
.061	.001	.027	-.04	-.02	.000	.055	-.05	.229	.204	.191	.250	.302	1.00										
.000	.034	.040	.041	.013	.023	.037	-.03	.215	.207	.260	.245	.269	.287	1.00									
-.02	-.02	-.05	-.02	.000	-.02	-.05	.025	.265	.289	.262	.246	.219	.193	.189	1.00								
.029	.002	.017	-.02	.024	-.02	.075	-.05	-.05	-.08	-.06	-.01	.062	.057	.052	-.05	1.00							
-.09	.026	-.01	103	-.01	.052	-.09	.061	.010	.000	.110	.002	-.07	-.09	.000	.050	.190	1.00						
.083	.000	.025	-.07	-.02	.010	.031	-.08	-.01	-.04	-.05	-.02	.052	.088	.006	-.03	.283	.152	1.00					
.000	.027	.000	.016	.026	.011	-.08	.031	.013	.020	.071	-.03	-.09	-.07	-.05	.060	.185	.310	.208	1.00				
.050	-.05	-.01	-.08	-.02	-.06	.105	-.03	.011	.014	-.12	.043	.087	.084	.008	-.05	.308	.115	.321	.164	1.00			
-.04	.005	-.03	.017	.000	.026	-.09	.029	.013	.019	.062	-.02	-.07	-.06	-.05	.075	.185	.338	.204	.326	.153	1.00		
.018	-.06	-.04	-.07	.007	-.06	.040	.008	.018	.045	-.08	.021	.032	.015	-.04	.015	.262	.180	.259	.235	.325	.233	1.00	
-.09	.019	-.02	.106	.000	.025	.057	.072	.008	.009	.085	.027	.058	-.07	.018	.029	.219	.400	.138	.280	.150	.305	.196	1.00

Appendix G: Population correlation matrix $p{:}m$ = 4:1 (p = 24, m = 6, a_{ij} = .8).

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
1.00																							
.662	1.00																						
.686	.688	1.00																					
.593	.659	.606	1.00																				
-.01	-.07	-.03	-.04	1.00																			
.046	.017	.054	-.05	.647	1.00																		
.051	-.08	-.01	-.08	.713	.645	1.00																	
.020	-.09	-.04	-.04	.694	.617	.776	1.00																
-.02	-.03	-.01	-.03	.016	.012	.004	.000	1.00															
.046	.012	.029	-.05	.004	.060	.011	-.01	.658	1.00														
.030	-.03	.035	-.06	.017	.053	.038	.014	.675	.685	1.00													
.011	.098	.053	.046	-.09	.013	-.11	-.09	.591	.636	.612	1.00												
.029	-.07	.006	-.05	.074	.002	.137	.105	-.02	-.02	.033	-.07	1.00											
-.01	.027	.001	.067	-.06	-.03	-.06	-.04	-.04	-.04	-.04	.062	.600	1.00										
.017	.000	.045	-.01	-.03	.033	-.02	-.02	.012	.020	.041	.021	.632	.665	1.00									
-.04	.081	.034	.043	-.06	.025	-.17	-.14	.016	.015	-.01	.097	.510	.676	.676	1.00								
.005	.046	.057	.022	-.08	.038	-.10	-.08	.000	.022	.029	.091	-.06	.050	.073	.110	1.00							
.005	-.07	-.01	-.06	.064	.032	.063	.050	.052	.030	.059	-.08	.048	-.07	.003	-.04	.617	1.00						
-.01	.045	.003	.062	-.08	-.04	-.06	-.03	-.04	-.03	-.04	.070	-.03	.070	.004	.030	.690	.572	1.00					
.006	-.02	-.02	.028	.014	-.01	.002	.022	-.04	.003	-.02	.006	.004	.030	.000	-.02	.644	.634	.655	1.00				
-.06	.010	-.03	.061	-.04	-.03	-.10	-.05	.023	-.02	-.03	.027	-.10	.046	.016	.085	.046	-.03	.028	.007	1.00			
-.06	-.02	-.06	.048	.007	-.04	-.04	-.01	.017	-.03	-.04	-.02	-.05	.027	-.01	.024	-.02	-.01	.006	.017	.706	1.00		
.007	.000	.001	-.04	.008	.000	.054	.034	.028	-.01	.011	-.04	.035	-.05	-.02	-.05	-.05	.017	-.03	-.06	.603	.617	1.00	
-.05	.003	.006	.013	-.01	.015	-.09	-.07	.043	.005	.019	.019	-.06	.004	.033	.100	.065	.015	-.01	-.02	.703	.668	.621	1.00

Appendix H: Population matrix $p{:}\ m$ = 4:1 (p = 24, m = 6, a_{ij} = .50).

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
1.00																							
.181	1.00																						
.322	.167	1.00																					
.248	.236	.210	1.00																				
-.04	.069	-.06	.005	1.00																			
.042	-.07	.037	.037	.192	1.00																		
.057	-.04	.077	-.06	.230	.254	1.00																	
.014	-.03	-.03	.061	.243	.301	.214	1.00																
.070	-.06	.045	.005	-.03	.034	.052	.014	1.00															
.049	-.07	.101	-.04	-.06	.039	.074	-.03	.295	1.00														
.063	-.05	.046	-.01	-.03	.031	.054	.023	.314	.297	1.00													
.055	-.06	.030	.011	-.05	.052	.025	.046	.295	.287	.303	1.00												
.011	-.03	-.01	.030	.000	.044	.003	.037	.010	.032	.001	.030	1.00											
-.07	.083	-.10	.004	.076	-.06	-.04	-.02	-.05	-.06	-.06	-.06	.255	1.00										
-.02	-.01	.073	-.06	-.03	-.01	.029	-.07	-.03	.082	-.02	-.02	.237	.231	1.00									
.062	-.07	.077	.001	-.04	.039	.056	-.01	.054	.084	.043	.037	.276	.191	.268	1.00								
-.03	.044	-.01	-.04	.028	-.05	.003	-.03	-.02	-.02	.002	-.03	-.05	.016	.018	-.04	1.00							
-.05	.029	-.07	.054	.048	.000	-.07	.028	-.04	-.04	-.06	-.03	.038	.064	-.02	-.04	.224	1.00						
.066	-.05	.061	-.04	-.05	.017	.075	-.02	.060	.055	.061	.044	-.01	-.06	.004	.056	.243	.171	1.00					
.035	-.04	.003	.065	-.01	.037	-.02	.044	.042	-.01	.025	.026	.010	-.03	-.06	.030	.221	.256	.253	1.00				
.043	-.03	.102	-.08	-.05	-.01	.067	-.06	.022	.07	.042	.017	-.06	-.07	.082	.034	.036	-.10	.074	-.03	1.00			
.026	-.04	.040	.015	-.02	.025	.028	.000	.032	.08	.024	.014	.029	-.02	.023	.067	-.02	-.01	.018	.024	.254	1.00		
-.08	.093	-.15	.063	.075	-.04	-.10	.058	-.06	-.15	-.05	-.04	-.02	.090	-.10	-.10	.032	.073	-.09	.016	.134	.186	1.00	
-.05	.043	-.08	.064	.039	-.01	-.10	.048	-.05	-.10	-.05	-.03	-.01	.038	-.05	-.07	.010	.064	-.08	.025	.178	.207	.375	1.00

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 134-139

1538 – 9472/05/$95.00

Testing The Casual Relation Between Sunspots And Temperature Using Wavelets Analysis

Abdullah Almasri
Department of Statistics
Lund University, Sweden

Ghazi Shukur
Departments of Economics and Statistics
Jönköping University and Växjö University

Investigated and tested in this article are the causal nexus between sunspots and temperature by using statistical methodology and causality tests. Because this kind of relationship cannot be properly captured in the short run (daily, monthly or yearly data), the relationship is investigated in the long run using a very low frequency Wavelets-based decomposed data such as D8 (128 - 256 months). Results indicate that during the period 1854-1989, the causality nexus between these two series is as expected of one-directional form, i.e., from sunspots to temperature.

Key words: Wavelets, time scale, causality tests, sunspots, temperature

Introduction

The Sun is the energy source that powers Earth's weather and climate, and therefore it is natural to ask whether changes in the Sun could have caused past climate variations and might cause future changes. At some level the answer must be yes. Recently, concerns about human-induced global warming have focused attention on just how much climatic change the Sun could produce. Accordingly, many authors tried to investigate the relation between the sunspots and the climate change, e.g., Friis-Christensen (1997) compared observations of cloud cover and cosmic particles and concluded that variation in global cloud cover was correlated with the cosmic ray flux from 1980 to 1995. They proposed the observed variation in cloud cover seemed to be caused by the varying solar activity related cosmic ray flux and postulated that an accompanying change in the earth's albedo could explain the observed correlations between solar activity and climate. However, Jorgensen and Hansen (2000) showed that any evidence supporting that the mechanism of cosmic rays affecting the cloud cover and hence climate does not exist.

Nevertheless, most of these studies suffered from the lack of statistical methodology. In this study, well selected statistical tools are used to investigate the causal relation between the sunspots and the temperature. A vector autoregressive (VAR) model is constructed and applied, which allows for causality test, on low frequency Wavelets based decomposed data. Processing in this manner we can see the nature of the causal relation between these two variables.

Wavelet is a fairly new approach in analysing data (e.g. Daubechies, 1992) that is becoming increasingly popular for a wide range of applications (e.g. time series analyses). This subject is not really familiar in other areas such as in statistics with environmental application. The idea behind using this technique is based on the fact that the time period (time scale) of the analysis is very crucial for determining those aspects that are relatively more important, and those that are relatively less important. In time series one can envisage a cascade of time scales

Abdullah Almasri also holds an appointment in the Department of Primary Care and General Practice, University of Birmingham, UK. Email: abdullah.almasri@stat.lu.se. Ghazi Shukur holds appointments in the Departments of Economics and Statistics, Jönköping University and Växjö University, Sweden.

within which different levels of information are available. Some information is with long horizons, others with short horizons.

In this article, the discrete wavelet transform (DWT) is used in studying the relationship between the sunspots and temperature in Northern Hemisphere 1854-1989 (see Figures 1 and 2). The DWT has several appealing qualities that make it a useful method for time series, exhibiting features that vary both in time and frequency. By using the DWT, it is possible to investigate the role of time scale in sunspots and temperature relationships.

The article is organized as follows: After the introduction, the wavelets analysis is introduced. Next presented is the methodology and testing procedure used in this study. Estimated results follow, and finally, summary and the conclusion.

Methodology

The wavelet transform has been expressed by Daubechies (1992) as "a tool that cuts up data or functions into different frequency components, and then studies each component with a resolution matched to its scale." Thus, with wavelet transform, series with heterogeneous (unlike Fourier transform) or homogeneous information at each scale may be analyzed. Unlike the Fourier transform, which uses only sins and cosines as basis functions, the wavelet transform can use a variety of basis functions.

The wavelet decomposition is made with respect to the so-called Symmlets basis. Thus, a brief presentation about this decomposition methodology, which called the discrete wavelete transform (DWT), is given.

Let $\mathbf{X} = (X_0, X_2, ..., X_{T-1})'$ be a column vector containing T observations of a real-valued time series, and assume that T is an integer multiple of 2^M, where M is a positive integer. The discrete wavelet transform of level J is an orthonormal transform of $\mathbf{X}$ defined by

$$\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_J, \mathbf{s}_J)' = \mathbf{W}\mathbf{X},$$

where $\mathbf{W}$ is an orthonormal $T \times T$ real-valued matrix, i.e. $\mathbf{W}^{-1} = \mathbf{W}'$ so $\mathbf{W}'\mathbf{W} = \mathbf{W}\mathbf{W}' = \mathbf{I}_T$. $\mathbf{d}_j = \{d_{j,k}\}$, $j = 1,2,\ldots,J$, are $T/2^j \times 1$ real-valued vectors of wavelet coefficients at scale j and location k.

The real-valued vector $\mathbf{s}_J$ is made up of $T/2^J$ scaling coefficients. Thus, the first $T - T/2^J$ elements of $\mathbf{d}$ are wavelet coefficients and the last $T/2^J$ elements are scaling coefficients, where $J \le M$. Notice that the length of $\mathbf{X}$ does coincide with the length of $\mathbf{d}$ (length of $\mathbf{d}_j = 2^{M-j}$, and $\mathbf{s}_J = 2^{M-J}$).

The multiresolution analysis of the data leads to a better understanding of wavelets. The idea behind multiresolution analysis is to express $\mathbf{W}'\mathbf{d}$ as the sum of several new series, each of which is related to variations in $\mathbf{X}$ at a certain scale. Because the matrix $\mathbf{W}$ is orthonormal, the time series may be constructed from the wavelet coefficients $\mathbf{d}$ by using $\mathbf{X} = \mathbf{W}'\mathbf{d}$.

Partition the columns of $\mathbf{W}'$ commensurate with the partitioning of $\mathbf{d}$ to obtain

$$\mathbf{W}' = [\mathbf{W}_1\ \mathbf{W}_2 \ldots \mathbf{W}_J\ \mathbf{V}_J],$$

where $\mathbf{W}_j$ is a $T \times T/2^j$ matrix and $\mathbf{V}_J$ is a $T \times T/2^J$ matrix. Define the multiresolution analysis of a series by expressing $\mathbf{W}'\mathbf{d}$ as a sum of several new series, each of which is related to variations in $\mathbf{X}$ at a certain scale:

$$\mathbf{X} = \mathbf{W}'\mathbf{d} = \sum_{j=1}^{J} \mathbf{W}_j \mathbf{d}_j + \mathbf{V}_J \mathbf{s}_J = \sum_{j=1}^{J} \mathbf{D}_j + \mathbf{S}_J .$$

Figure 1: Monthly data of the sunspots.

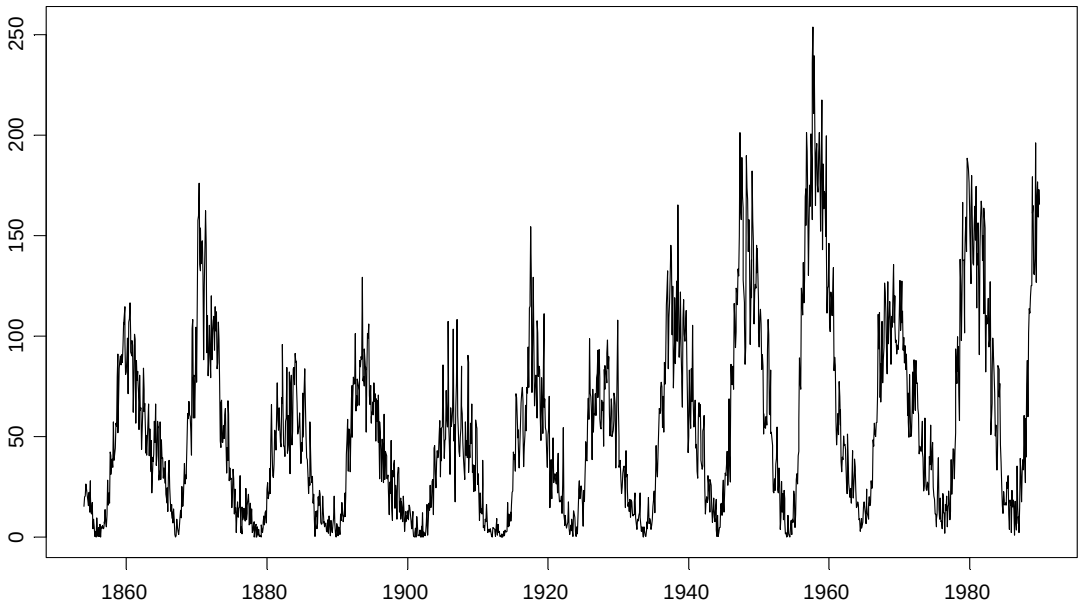
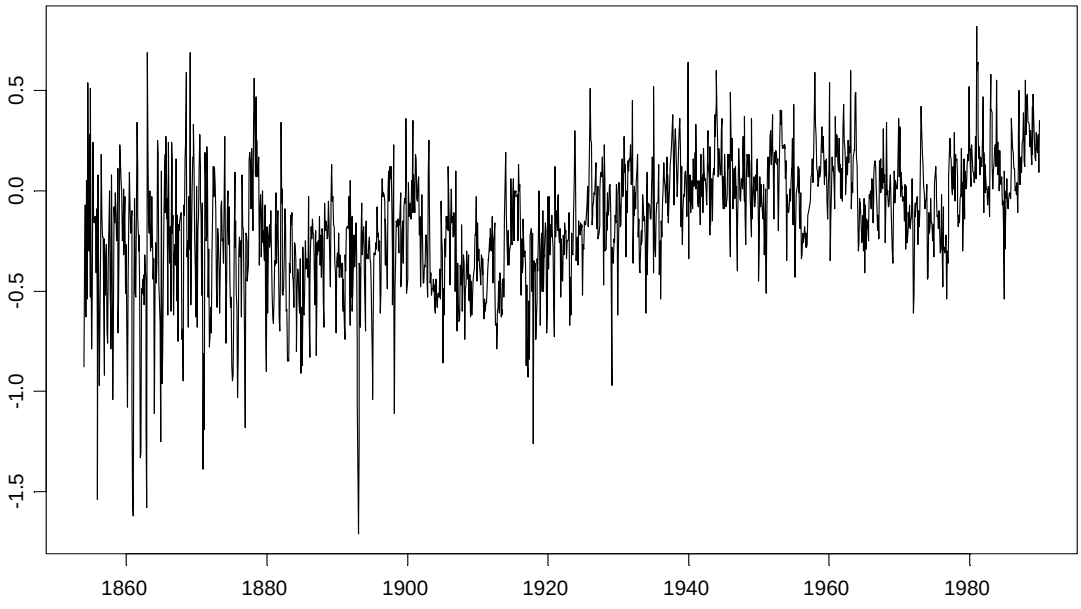


Figure 2: Monthly data of the Northern Hemisphere temperature.



Monthly temperature for the northern hemisphere for the years 1854-1989, from the data base held at the Climate Research Unit of the University of East Anglia, Norwich, England (Briffa. & Jones, 1992). The numbers consist of the temperature (degrees C) difference from the monthly average over the period 1950-1979.

The terms in the previous equation constitute a decomposition of **X** into orthogonal series components $\mathbf{D}_j$ (detail) and $\mathbf{S}_J$ (smooth) at different scales, and the length of $\mathbf{D}_j$ and $\mathbf{S}_J$ coincides with the length of **X** ($T \times 1$ vector). Because the terms at different scales represent components of **X** at different resolutions, the approximation is called a multiresolution decomposition, see Percival and Mofjeld (1997).

As mentioned earlier the wavelet decompositions in this paper will be made with respect to the Symmlets basis. This has been done by using the S-plus Wavelets package produced by StatSci of MathSoft that was written by Bruce and Gao (1996). Figure 3 shows the multiresolution analysis of order $J = 6$ based on the Symmlets of length 8.

The Causality Between the Sunspots and Temperature

Next the wavelets analysis is used in investigating the hypothesis that the sunspots may affect the temperature. This will mainly be done by using causality test. Because this kind of relationship can not properly be captured in the short run (daily, monthly or yearly data), only the relationship in the long run is investigated, either by using 10-20 years data (which is not available in this case) or a very low frequency Wavelets decomposed data like D8 (128 - 256 months). This is used in this article (see Figure 3). This will be done empirically by constructing a (VAR) model that allows for causality test in the Granger sense.

Causality is intended as in the sense of Granger (1969). That is, to know if one variable precedes the other variable or if they are contemporaneous. The Granger approach to the question whether sunspots (Sun) causes temperature (Tem) is to see how much of the current value of the second variables can be explained by past values of the first variable. (Tem) is said to be Granger-caused by (Sun) if (Sun) helps in the prediction of (Tem), or equivalently, if the coefficients of the lagged (Sun) are statistically significant in a regression of (Tem) on (Sun). Empirically, one can test for causality in Granger sense by means of the following vector autoregressive (VAR) model:

$$Tem_t = a_0 + \sum_{i=1}^{k} a_i Tem_{t-i} + \sum_{i=1}^{k} b_i Sun_{t-i} + e_{1t} \quad (1)$$

$$Sun_t = c_0 + \sum_{i=1}^{k} c_i Tem_{t-i} + \sum_{i=1}^{k} f_i Sun_{t-i} + e_{2t} \quad (2)$$

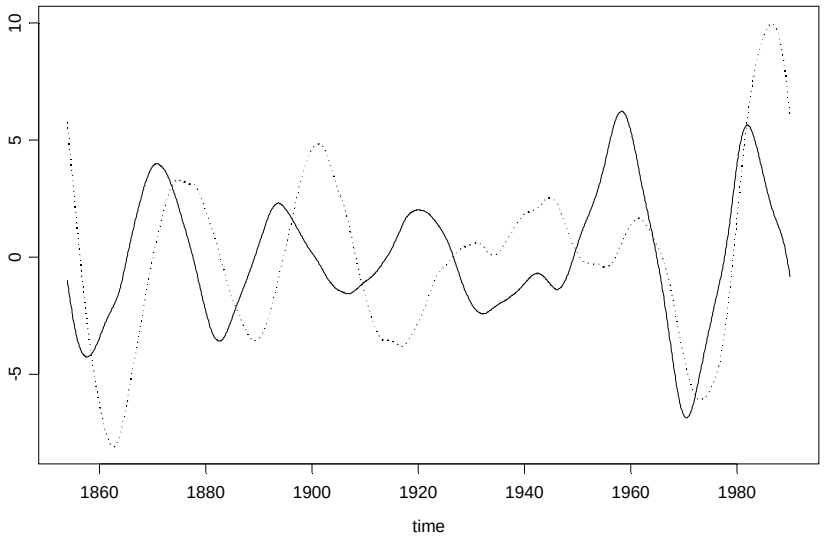
where e_{1t} and e_{2t} are error terms, which are assumed to be independent white noise with zero mean. The number of lags, *k*, will be decided by using the Schwarz (1978) information criteria, in what follows referred to as SC.

According to Granger and Newbold (1986) causality can be tested for in the following way: A joint *F*-tests is constructed for the inclusion of lagged values of (Sun) in (1) and for the lagged values of (Tem) in (2). The null hypothesis for each *F*-test is that the added coefficients are zero and therefore the lagged (Sun) does not reduce the variance of (Tem) forecasts (i.e. b_i in (1) are jointly zero for all *i*), or that lagged (Tem) does not reduce the variance of (Sun) forecasts (i.e. f_i in (2) are jointly zero for all *i*). If neither null hypothesis is rejected, the results are considered as inconclusive.

However, if both of the *F*-tests rejected the null hypothesis, the result is labeled as a feedback mechanism. A unique direction of causality can only be indicated when one of the pair of *F*-tests rejects and the other accepts the null hypothesis which should be the case in the study.

Moreover, before testing for causality, the augmented Dickey-Fuller (1979, 1981) is applied, in what follows referred to as ADF, test for deciding the integration order of each aggregate variable. When looking at the Wavelets decomposed data for sun and temperature used here, i.e. the D8 in Figure 3 below, the ADF test results indicate that each variable is integrated of the same order zero, i.e., I(0), indicating the both of the series are stationary implying that the VAR model can be estimated by standard statistical tools.

Figure 3: The solid line denotes the D8 for the sunspots data and the dotted line denotes the D8 for the temperature.



Results

According to the model selection criteria proposed by Schwarz (1978), it is found that the model that minimizes this criteria is the VAR(3). When this model is used to test for causality, the inference is drawn that only the (Sun) Granger causes the (Tem). The test results can be found in Table 1, below. This means that the causality nexus between these two series is a one-directional form, i.e., from (Sun) to (Tem). This should be fairly reasonable, since it is not logical to assume that the temperature in the earth should have any significant effect on the sunspots.

Table 1: Testing results for the Granger causality.

Null Hypothesis	P-value
Sun does not Granger Cause Tem	0.0037
Tem does not Granger Cause Sun	0.5077

Conclusion

The main purpose of this article is to model the causality relationship between sunspots and temperature. Although other studies exist for the similar purpose, they are not based on a careful statistical modeling. Moreover, these studies have sometimes shown to end up with conflicting results and inferences. Here, in this article, well selected statistical methodology for estimation and testing the causality relation between these two variables is used.

A very low frequency Wavelets based decomposed data indicates that, during the period 1854-1989, the causality nexus between these two series is the expected one-directional form, i.e., from sunspots to temperature

References

Bruce, A. G., & Gao, H.-Y. (1996). *Applied wavelet analysis through S-Plus*, New York: Springer-Verlag.

Daubechies, I. (1992). *Ten lectures on wavelets*, Volume 61 of *CBMS-NFS Regional Conference Series in Applied Mathematics*. Philadelphia: Society for Industrial and Applied Mathematics.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427-431.

Dickey, D. A., & Fuller, W. A. (1981). The likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*, 1057-1072.

Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series*, 2nd New York: Academic Press.

Granger, C. W. J. (1969). Investigating causal relations by econometric models an cross-spectral methods. *Econometrica*, *37*, 24-36.

Percival, D. B., & Mofjeld, H. O. (1997). Analysis of subtidal coastal sea level fluctuations using Wavelets. *Journal of the American Statistical Association, 92*(439), 868-880.

Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge, UK: Cambridge University Press.

Schwarz, G. (1978). Estimation the dimention of a model. *Annals of Statistics*, 6, 461-464.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 140-154

1538 – 9472/05/$95.00

Bayesian Wavelet Estimation Of Long Memory Parameter

Leming Qu
Department of Mathematics
Boise State University

A Bayesian wavelet estimation method for estimating parameters of a stationary I(d) process is represented as an useful alternative to the existing frequentist wavelet estimation methods. The effectiveness of the proposed method is demonstrated through Monte Carlo simulations. The sampling from the posterior distribution is through the Markov Chain Monte Carlo (MCMC) easily implemented in the WinBUGS software package.

Key words: Bayesian method, wavelet, discrete wavelet transform (DWT), I(d) process, long memory

Introduction

Stationary processes exhibiting long range dependence have been widely studied now since the works of Granger and Joyeux (1980) and Hosking (1981). The long range dependence has found applications in many areas, including economics, finance, geosciences, hydrology, and statistics. The estimation of the long-memory parameter of the stationary long-memory process is one of the important tasks in studying this process.

There exist parametric, non-parametric and semi-parametric methods of estimation for the long-memory parameter in literature. In the parametric method, the long-memory parameter is one of the several parameters that determine the parametric model; hence the usual classical methods such as the maximum likelihood estimation can be applied. The non-parametric method, not assuming restricted parametric form of the model, usually uses regression methods by regressing the logarithm of some sampling statistics for estimation.

Leming Qu is an Assistant Professor of Statistics, 1910 University Dr., Boise, ID 83725-1555. Email: qu@math.boisestate.edu. His research interests include Wavelets in statistics, time series, Bayesian analysis, statistical and computational inverse problems, nonparametric and semiparametric regression.

The widely and often used Geweke and Poter-Hudak (1983) estimation method belongs to non-parametric methods. The semi-parametric method makes intermediate assumptions by not specifying the covariance structure at short ranges. The article by Bardet et al. (2003) surveyed some semi-parametric estimation methods and compared their finite sample performance by Monte-Carlo simulation.

Wavelet has now been widely used in statistics, especially in time series, as a powerful mutiresolution analysis tool since 1990's. See Vidakovic (1999) for reference from the statistical perspective. The wavelet's strength rests in its ability to localize a process in both time and frequency scale simultaneously.

This article presents a Bayesian Wavelet estimation method of the long-memory parameter d and variance σ^2 of a stationary long-memory I(d) process implemented in the MATLAB computing environment and the WinBUGS software package.

Methodology

A time series $\{X_t\}$ is a fractionally integrated process, *I(d)*, if it follows:

$$(1-L)^d X_t = \varepsilon_t ,$$

where $\varepsilon_t \sim i.i.d.\ N(0, \sigma_\varepsilon^2)$ and L is the lag operator defined by $LX_t = X_{t-1}$. The parameter d is not necessarily an integer so that fractional differencing is allowed. The process $\{X_t\}$ is

stationary if $|d|< 0.5$. The fractionally differencing operator $(1-L)^d$ is defined by the general binomial expansion:

$$(1-L)^d=\sum_{k=0}^{\infty}\binom{d}{k}(-L)^k \ ,$$

where

$$\binom{d}{k}=\frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}$$

and $\Gamma(\cdot)$ is the usual Gamma function.

Denote the autocovariance function of $\{X_t\}$ as $\gamma(k)$, that is $\gamma(k)=E(X_t X_s)$ where $k=|t-s|$. The formula for $\gamma(k)$ of a stationary *I(d)* process is well-known (Beran 1994, pp. 63):

$$\gamma(0)=\sigma_\varepsilon^2\Gamma(1-2d)/\Gamma^2(1-d),$$
$$\gamma(k+1)=\gamma(k)(k+d)/(k+1-d),$$
$$k=0,1,2,\cdots$$

When $0 < d < 0.5$, the $\gamma(k)$ has a slow hyperbolic decay, hence the process $\{X_t\}$ is a long-memory process.

The fractional difference parameter d and the nuisance parameter σ^2 are usually unknown in an *I(d)* process. They need to be estimated from the observed time series X_t, $t=1,\ldots, N$.

Assume $N=2^J$ for some positive integer J in order to apply the fast algorithm of the Discrete Wavelet Transform (DWT) on $X=(X_t)_{t=1}^N$. Let $\omega=WX$ denote the DWT of X, where $\omega=(c_{j_0}^T, d_{j_0}^T, d_{j_0+1}^T, d_{J-1}^T)^T$. The j_0 is the lowest resolution level for which we use $j_0=0$ in this article. The smoothed wavelet coefficient vector $c_{j_0}=(c_{j_0,0}, c_{j_0,1},\cdots, c_{j_0,2^{j_0}-1})^T$. At the resolution level j, the detailed wavelet coefficient vector $d_j=(d_{j,0}, d_{j,1},\cdots, d_{j,2^j-1})^T$ for $j=0, 1,\ldots, J-1$.

McCoy and Walden (1996) argued heuristically that the DWT coefficients of X has the following distribution:

$$d_{j,k}\sim N(0,\sigma_j^2),$$

where $j=0,1,\cdots,J-1;\quad k=0,1,\cdots,2^j-1$, $c_{0,0}\sim N(0,\sigma_{-1}^2)$, and the $d_{j,k}$'s and $c_{0,0}$ are approximately uncorrelated due to the whitening property of the DWT. The σ_j^2, $j=-1, 0, 1,\ldots, J-1$ depend on d and σ_ε^2 as $\sigma_j^2=2^{J-j}\times 2\times 4^{-d}\sigma_\varepsilon^2\int_{2^{-(J-j+1)}}^{2^{-(J-j)}}\sin^{-2d}(\pi f)df$.

When $J-j\geq 2$, $f<2^{-2}$, then $\sin(\pi f)\approx \pi f$, so that σ_j^2 can be simplified as (see equation (2.10) of McCoy and Walden 1996)

$$\sigma_j^2=(2\pi)^{-2d}\sigma_\varepsilon^2 2^{2(J-j)d}(2-2^{2d})/(1-2d) \quad (1)$$

where $j=-1,0,\cdots,J-2$.

McCoy and Walden (1996) used these facts to estimate d and σ_ε^2 by the Maximum Likelihood Method. They demonstrated through simulation that d could be estimated as well, or better by wavelet methods than the best Fourier-based method.

Jensen (1999) derived the similar result about the distribution of the wavelet coefficients, and by the fact that $Var(d_{j,k})\propto 2^{-2jd}$, he used the Ordinary Least Squares method to estimate d. That is, by regressing *log* of the sample variance of the wavelet coefficients at resolution level j, against $\log(2^{-2j})$ for $j=2,3,\ldots, J-2$, he obtained the OLS estimate of d. The sample variance of the wavelet coefficients at resolution level j is estimated by the sample second moment of the observed wavelet coefficients at resolution level j.

Vannucci and Corradi (1999) section 5 proposed a Bayesian approach. They used independent priors and assumed Inverse Gamma distribution for σ_ε^2 and a Beta distribution for $2d$. They did not use formula (1), instead, they used a recursive algorithm to compute the variances of wavelet coefficients. The posterior inference is done through Markov chain Monte Carlo (MCMC) sampling procedure. They did

not give details of the implementation in the paper.

McCoy and Walden (1996) did not give the variance of their estimates. Jensen (1999) only estimated *d* using the OLS method, it is not clear how σ_ε^2 is estimated. In both cases, the estimated *d* can not be guaranteed in the range *(-0.5, 0.5)*.

Here, we propose a Bayesian approach to estimate *d* and σ_ε^2 in the same spirit of Vannucci and Corradi (1999) section 5. The distinction of this article from Vannucci and Corradi (1999) is that firstly, we use the explicit formula (1) for the variances of wavelet coefficients at resolution level j instead the recursive algorithm to compute these variances; secondly, the MCMC is implemented in the WinBUGS software package.

Denoting $\theta = (d, \sigma_\varepsilon^2)$, the parameters of the models for the data ω. If a prior distribution of $\pi(\cdot)$ of θ is chosen, i.e., $\theta \sim \pi(\theta)$, then by Bayesian formula, the posterior distribution of θ is

$$\pi(\theta \mid \omega) \propto f(\omega \mid \theta)\pi(\theta)$$

where $f(\omega \mid \theta)$ is the likelihood of the data ω given the parameters θ, which is the density of the multivariate normal distribution $N(0, \Sigma)$ with

$$\Sigma = diag(\sigma_{-1}^2, \Sigma_0, \Sigma_1, \cdots, \Sigma_{J-1})$$

and $$\Sigma_j = diag(\sigma_j^2, \cdots, \sigma_j^2)$$

for $j = 0,1,\cdots, J-1$ is a $2^j \times 2^j$ diagonal matrix.

The inference of θ is based on the posterior distribution $\pi(\theta \mid \omega)$. The MCMC methods are popular to draw repeated samples from the intractable $\pi(\theta \mid \omega)$. We focus on the implementation of the Gibbs sampling for estimating *d* and σ_ε^2 in the WinBUGS software. The easy programming in the WinBUGS software provides practitioners an useful and convenient tool to carry out Bayesian computation for long memory time series data analysis.

The following priors will be used. The first prior is the Jefferys' noninformative prior subject to the constraints of the range of model parameters:

$$\pi(\theta) \propto [J(\theta)]^{1/2} I_{(0,+\infty)}(\sigma_\varepsilon^2) I_{(-0.5,0.5)}(d),$$

where $I(\cdot)$ is an indicator function for the subscripted set and $J(\theta)$ is the Fisher information for θ:

$$J(\theta) = -E\left[\frac{\partial^2 \ln f(\omega \mid \theta)}{\partial\theta\partial\theta^T}\right].$$

Simple calculation shows that $J(\theta) \propto 1/\sigma_\varepsilon^2$. The second prior is the other independent priors on *d* and σ_ε^2, i.e.,

$$\pi(\theta) = \pi(d)\pi(\sigma_\varepsilon^2).$$

The prior for *d+0.5* is $Beta(\alpha, \beta)$ where $\alpha > 0$, $\beta > 0$ are the hyperparameters. This prior restricts *|d|<0.5*, thus imposing stationarity for the time series. When $\alpha = \beta = 1$, the prior is the noninformative uniform prior. When historical information or expert opinion is available, α and β can be selected to reflect this extra information, thus obtaining an informative prior. Hyper priors can also be used on α and β to reflect uncertainties on them, thus forming a hierarchical Bayesian model.

A $Gamma(\alpha_1, \alpha_2)$ prior is chosen for the precision $\tau^2 = 1/\sigma_\varepsilon^2$, where $\alpha_1 > 0, \alpha_2 > 0$ are the hyperparameters. When α_1 and α_2 are close to zero, the prior for σ_ε^2 is practically equivalent to $\pi(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$, an improper prior. The non-informative prior $\pi(\sigma_\varepsilon^2) \propto I_{(0,+\infty)}(\sigma_\varepsilon^2)$ can also be chosen.

Simulation

The MCMC sampling is carried out in the WinBUGS software package. WinBUGS is the current windows-based version of the BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling), a newly developed, user-friendly and free software package for general-purpose Bayesian computation, Lunn et al. (2000). It is developed by the MRC, Biostatistics Group, Institute of Public Health (www.mrc-bsu.cam.ac.uk/bugs), Cambridge.

In WinBUGS programming, user only needs to specify the full proper data distribution and prior distributions, WinBUGs will then use certain sophisticated sampling methods to sample the posterior distribution.

In this Monte Carlo experiment, we compare the proposed Bayesian approach with the approach in McCoy and Walden (1996) and Jensen (1999). Different values of *d, N* and different prior distributions $\pi(\theta)$ are used to determine the effectiveness of the estimation procedure. Also used were two different wavelet bases to compare the effect of this choice.

The Davis and Harte (1987) algorithm was used to generate an *I(d)* process because of its efficiency compared to other computationally intensive methods (McLeod & Hipel (1978). This algorithm generates a Gaussian time series with the specified autocovariances by discrete Fourier transform and discrete inverse Fourier transform. It is well known that Fast Fourier Transform (FFT) can be carried out in *O(N log N)* operations, so the computation is fast.

The generation of the *I(d*) process using the Davis and Harte algorithm and the DWT of the generated *I(d)* process are carried out in the MATLAB 6.5 on a Pentium III running Windows 2000. The DWT tool used is the *WAVELAB802* developed by the team from the Statistics Department of Stanford University (http://www-stat.stanford.edu/~wavelab).

The following two different wavelet basis for comparison were chosen: (a) Harr wavelet; (b) LA(8): Daubechies least asymmetric compactly supported wavelet basis with four vanish moments, see p.198 of Daubechies (1992).

The periodic boundary handling is used. The data of the discrete wavelet transformed *I(d)* process is first saved in a file in R data file format. Then WinBUGS1.4 is activated under MATLAB to run a script file that implements the proposed Bayesian estimation procedure. The estimation results from WinBUGS1.4 are then converted to the MATLAB variables for further uses.

The model parameters are estimated under the following independent priors on *d* and σ_ε^2

(a) $$d \sim Unif(-0.5, 0.5), \quad \sim Gamma(0.01, 0.01);$$

(b) $d \sim Unif(-0.5, 0.5), \sigma_\varepsilon^2 \sim Unif(0, 1000).$

The prior (a) is practically equivalent to Jefferys' noninformative prior:

$$\pi(d, \sigma_\varepsilon^2) \propto \frac{1}{\sigma_\varepsilon^2} I_{(0,+\infty)}(\sigma_\varepsilon^2) I_{(-0.5,0.5)}(d).$$

BUGS only allow the use of proper prior specification, so the non-informative or improper prior distribution can be regarded as the limit of a corresponding proper prior.

The estimation results using the proposed Bayesian approach for the simulated *I(d)* process and the method by Jensen (1999) and McCoy and Walden (1996) are found in Table 1 for Haar wavelets and Table 2 for LA(8) wavelets. For the chosen prior, it reports the estimated posterior mean, posterior standard deviation (SD). In addition, it also tabulated in the parenthesis below the value of Mean and SD the 95% credible intervals of the parameters using the 2.5% and 97.5% quantiles of the random samples.

In all cases, two independent chains of *10500* iterations each were run, keeping every tenth one, after burn-in 500, with random initial values. The posterior inference is based on the actual random samples of 2000. For the case of *N=256, d=0.1,* σ_ε^2 *=1.0* and prior (b), Figure 1 shows the trace of the random samples and the kernel estimates of the posterior densities of the parameters.

The autocorrelation function of the random samples shows very little autocorrelations for the drawn series of the random samples. The two parallel chains mix well after small steps of the initial stage. All other diagnostics for convergence indicate a good convergence behavior.

In most cases, the Bayesian wavelet estimates of d and σ_ε^2 are quite good. They are very close to the truth. The 95% credible interval given by the Bayesian wavelet approach is well centered around the true parameter and is also very tight.

The estimation results using the two different priors (a) and (b) are very similar. The estimates by Jensen's method differ most from those by the other methods. It seems that LA(8) generally gives better estimates than Haar. This is in agreement with the results of McCoy and Walden (1996) section 5.2.

Figure 1: Trace and Kernel Density Plot for d and σ_ε^2.

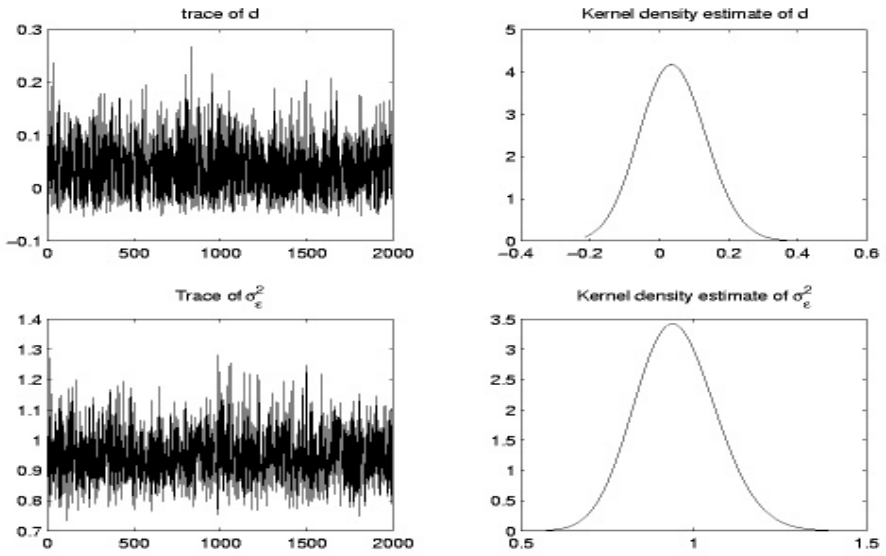


Table 1: Estimation of the simulated *I(d)* process when *N=256* Using Haar Basis.

			Prior (a)		Prior (b)	
Parameter	Jensen	MW	Mean	SD	Mean	SD
d=0.1	0.1620	0.1629	0.1686	0.0499	0.1692	0.0499
			(0.0739,	0.2711)	(0.0768,	0.2674)
$\sigma_\varepsilon^2 = 1.0$		1.0226	1.0452	0.0931	1.0485	0.0977
			(0.8801,	1.2460)	(0.8791,	1.2540)
d=0.25	0.1431	0.1858	0.1887	0.0465	0.1880	0.0462
			(0.1049,	0.2827)	0.1008,	0.2854)
$\sigma_\varepsilon^2 = 1.0$		1.0789	1.1000	0.0972	1.1068	0.1021
			(0.9331,	1.3150)	(0.9289,	1.3220)
d=0.4	0.4121	0.4384	0.4301	0.0351	0.4284	0.0364
			(0.3567,	0.4902)	(0.3489,	0.4901)
$\sigma_\varepsilon^2 = 1.0$		1.0189	1.0445	0.0934	1.0571	0.0975
			(0.8775,	1.2395)	(0.8822,	1.2640)
d=0.1	0.1227	0.0681	0.0709	0.0470	0.0719	0.0472
			(-0.0176,	0.1663)	(-0.0172	0.1679)
$\sigma_\varepsilon^2 = 2.0$		2.1482	2.1787	0.1918	2.1943	0.1948
			(1.8455,	2.5745)	(1.8570,	2.5975)
d=0.25	0.2468	0.1855	0.1858	0.0477	0.1847	0.0462
			(0.0995,	0.2858)	(0.0938,	0.2785)
$\sigma_\varepsilon^2 = 2.0$		1.9369	1.9674	0.1729	1.9791	0.1770
			(1.6570,	2.3275)	(1.6715,	2.3675)
d=0.4	0.2154	0.3096	0.3127	0.0467	0.3105	0.0476
			(0.2238,	0.4069)	(0.2165,	0.4079)
$\sigma_\varepsilon^2 = 2.0$		1.7783	1.8130	0.1540	1.8305	0.1619
			(1.5435,	2.1385)	(1.5300,	2.1665)

Table 2: Estimation of the simulated *I(d)* process when *N=256* Using LA(8) Basis.

			Prior (a)		Prior (b)	
Parameter	Jensen	MW	Mean	SD	Mean	SD
d=0.1	0.0759	0.1701	0.1757	0.0466	0.1755	0.0446
			(0.0894,	0.2734)	(0.0936,	0.2707)
$\sigma_\varepsilon^2 = 1.0$		1.0037	1.0222	0.0935	1.0270	0.0899
			(0.8529,	1.2295)	(0.8626,	1.2190)
d=0.25	0.0904	0.2611	0.2651	0.0508	0.2661	0.0502
			(0.1681,	0.3680)	(0.1705,	0.3741)
$\sigma_\varepsilon^2 = 1.0$		1.0154	1.0398	0.0916	1.0412	0.0888
			(0.8791,	1.2295)	(0.8824,	1.2255)
d=0.4	0.4906	0.4369	0.4304	0.0359	0.4295	0.0362
			(0.3548,	0.4905)	(0.3536,	0.4895)
$\sigma_\varepsilon^2 = 1.0$		1.0148	1.0413	0.0953	1.0502	0.0932
			(0.8669,	1.2370)	(0.8826,	1.2450)
d=0.1	0.0542	0.1110	0.1175	0.0529	0.1151	0.0535
			(0.0166,	0.2278)	(0.0183,	0.2298)
$\sigma_\varepsilon^2 = 2.0$		2.1233	2.1594	0.1926	2.1694	0.1894
			(1.8185,	2.5765)	(1.8235,	2.5650)
d=0.25	0.1977	0.2609	0.2608	0.0535	0.2637	0.0540
			(0.1556,	0.3697)	(0.1630,	0.3761)
$\sigma_\varepsilon^2 = 2.0$		1.8372	1.8745	0.1609	1.8849	0.1709
			(1.5870,	2.2165)	(1.5795,	2.2420)
d=0.4	0.2632	0.3111	0.3130	0.0454	0.3117	0.0463
			(0.2257,	0.4045)	(0.2236,	0.4040)
$\sigma_\varepsilon^2 = 2.0$		1.7469	1.7942	0.1635	1.7995	0.1595
			(1.5080,	2.1510)	(1.5145,	2.1225)

Figure 2: Box plots of the estimates for N=128.

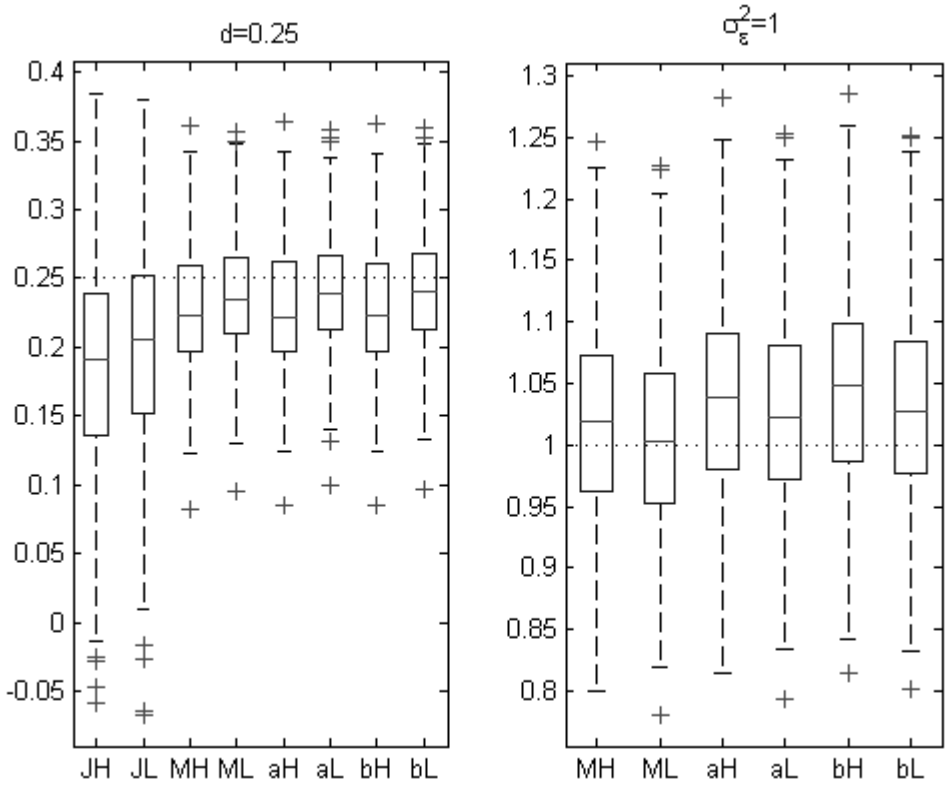
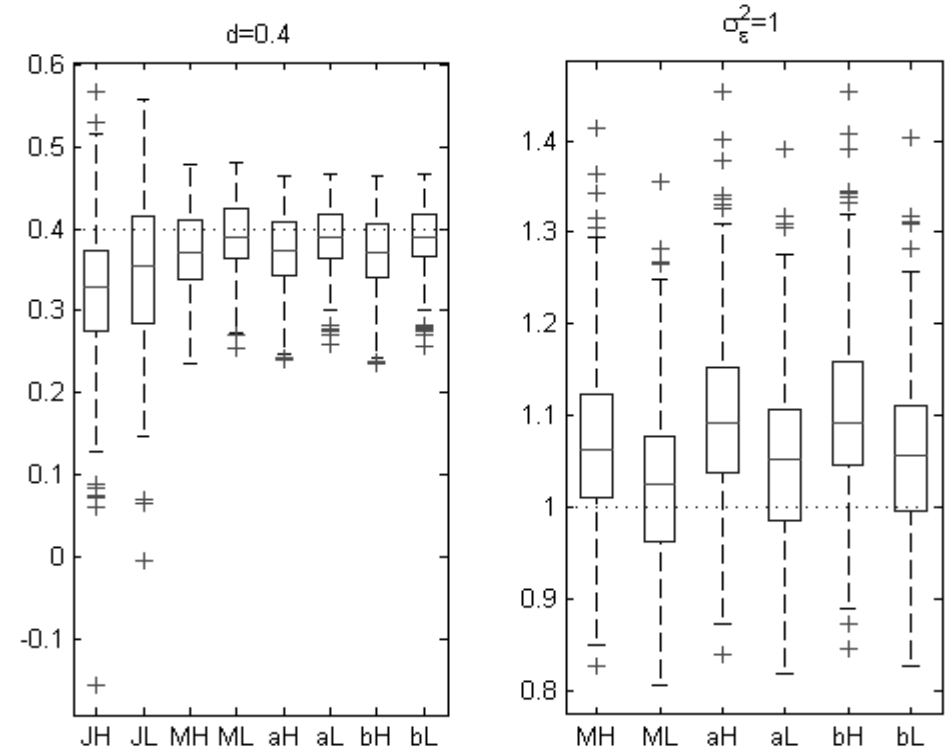


Figure 3: Box plots of the estimates for N=128.



Frequentist Comparison

Also compared were the estimates of the three methods in repeatedly simulated I(d) process. Figure 2 is the box plots of the estimates for d and σ_ε^2 respectively of 200 replicates with $N=128$, $d=0.25$ and $\sigma_\varepsilon^2=1.0$. Figure 3 is the box plots of the estimates for 200 replicates with $N=128$, $d=0.40$ and $\sigma_\varepsilon^2=1.0$. The *x*-axis labels in the box plot read as follows: `JH' denotes the case by the **J**ensen method using **H**aar; 'JL' denotes the case by the **J**ensen method using **L**A(8); and so forth. Because of the long computation time associated with the Gibbs sampling for the large number of simulated I(d) processes, we limit the burn-in to 100 iterations and the number of random samples to 500. Because only the posterior mean was calculated using the generated random samples, not much information was lost even when the slightly short chain was used.

For the estimates of d, the mean square errors of the McCoy and Walden and The Bayesian method using these two priors are very similar, and they are all smaller than the one by Jensen's OLS. LA(8) gives less biased estimates than Haar. The mean estimates for d given by the Bayesian method using LA(8) is similar to those by McCoy and Walden. In all methods, it seems the estimates for d and σ_ε^2 are a little biased in that $\hat{d}$ tends to underestimate d and $\hat{\sigma}_\varepsilon^2$ tends to overestimate σ_ε^2.

Conclusion

Bayesian wavelet estimation method for the stationary I(d) process provides an alternative to the existing frequentist wavelet estimation methods. Its effectiveness is demonstrated through Monte Carlo simulations implemented in the WinBUGS computing package.

A future effort is to extend the Bayesian wavelet method to more general fractional process such as *ARFIMA(p,d,q)*. The hypothesis testing problem for the I(d) process can also be explored via the Bayesian wavelet approach.

References

Bardet, J., Lang, G., Philippe, A., Stoev, S., & Taqqu, M. (2003). Semi-parametric estimation of the long-range dependence parameter: a survey. *Theory and Application of Long-Range Dependence*, Birkhauser.

Beran, J. (1994). *Statistics for long-memory processes*. New York: Chapman and Hall.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM

Davies, R. B., & Harte, D. S. (1987). Tests for Hurst effect. *Biometrik*a,*.74*, 95-101.

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, *4*, 221-238.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15-29.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, *68*, 165-176.

Jensen, M. J. (1999). Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter. *Journal of Forecasting*, *18*, 17-32.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

McCoy, E. J., Walden, A. T. (1996). Wavelet analysis and synthesis of stationary long-memory processes. *Journal of Computational & Graphical Statistics*, *5*, 26-56.

McLeod, B., & Hipel, K. (1978). Preservation of the rescaled adjusted range, I. A reassessment of the Hurst phenomenon. *Water Resources Research*, *14*, 491-518.

Vannucci, M., & Corradi, F. (1999). Modeling Dependence in the Wavelet Domain. In Bayesian Inference in Wavelet based Models. Lecture Notes in Statistics, Müller, P. and Vidakovic, B. (Eds). New York: Springer-Verlag.

Vidakovic, B. (1999). *Statistical modeling by wavelets*. New York: Wiley-Interscience.

WinBUGS 1.4: MRC, Biostatistics Group, Institute of Public Health, Cambridge University, Cambridge, 2003, www.mrc-bsu.cam.ac.uk/bugs

Appendix:

This appendix includes the MATLAB code for first simulating the I(d) process, then transforming it by DWT, and the WinBUGS program for the MCMC computation. In the WinBUGS programming, the symbol "~" is for the stochastic node which has the specified distribution denoted on the right side, the symbol "←" is for the deterministic node which has the specified expression denoted on the right side. All the likelihood function, the prior distributions
and initial values of the nodes without parents must be specified in the programs.

The MATLAB code:

```
function x=Generatex(J, d, sig2eps)
%Generate the I(d) process
%input:
%J: where N=2^J sample size
%d:  long memory parameter of the I(d) process, abs(d)<0.5
%sig2eps:     $\sigma_\epsilon^2$

%output:
%x: the time series

N=2^J;
c=[];
% generate the autocovariance function by the formular of covariance
% function for LRD
c(1)=sig2eps*gamma(1-2*d)/((gamma(1-d))^2);
%for i=1:N-1 c(i+1)=c(1)*gamma(i+d)*gamma(1-d)/(gamma(d)*gamma(i+1-d)); end;
for i=1:N-1 c(i+1)=c(i)*(i+d-1)/(i-d); end;

x=GlrdDH(c);


function x=GlrdDH(c);
%GlrdDH.m  Generating the stationary gaussion time seriess with specified
%         autocovariance series c
%         using Davis and Harte's method, Appendix of `Tests for Hurst Effect',
%         Biometrika, V74, No. 1 (Mar., 1987), 95-101

%c: autocovariance series

[temp, N]=size(c);   %c is a row vector

cCirculant=[];
for i=1:N-2  cCirculant(i)=c(N-i); end;

cFull=[];
cFull=[c cCirculant];
```

```
g=[];
g=fft(cFull);
%Fast Fourier Transform of cFull

Z=[];
Z=complex(normrnd(0,1,1,N), normrnd(0,1, 1,N));
Z(1)=normrnd(0,sqrt(2));   %Be careful to specify sqrt(2), if you want variance of Z(1) to be 2
Z(N)=normrnd(0,sqrt(2));

ZCirculant=[];
for i=1:N-2  ZCirculant(i)=conj(Z(N-i)); end;

ZFull=[];
ZFull=[Z ZCirculant];

X=[];
X=ifft(ZFull.*sqrt(g))*sqrt(N-1);

x=[];
x=real(X(1:N));

function [dJensen, dMW, sigMW, dBS, sigBS]=GetdHatSig2Hat(x, j0, filter)
%Wavelet estimation of Long Range Dependence parameters
%
%input:
%x: the observed I(d) process
%j0: lowest resolution level of the DWT
%filter: wavelet filter

%output:
%dJensen: estimate of d by Jensen 1999
%dMW: estimate of d by McCoy & Walden 1996
%sigMW: estimate of $\sigma_\epsilon^2$ by McCoy & Walden 1996
%dBS: estimate of d by Bayesian Wavelet Method for prior (a), (b)
%dBS.a, dBS.b
%sigBS: estimate of $\sigma_\epsilon^2$ by Bayesian Wavelet Method for prior (a), (b)
%sigBS.a, sigBS.b

N=length(x);
J=log2(N);

w=[];

w = FWT_PO(x,j0,filter)'; %w is a coulmn vector

resolution=[];        % data used in WinBUGS14
resolution(1:2^j0,1)=j0-1;
for j = j0:(J-1)
   resolution(2^j+1 : 2^(j+1),1)=j;
end;

vwj=[];
for j=j0+1:(J-1)
  vwj(j, :)=[j, mean(w(dyad(j)).^2)];
```

```
end;

tempd=[];
tempd=-[ones(J-2,1), log(2.^(2*vwj(2:J-1,1)))]\log(vwj(2:J-1,2));
dJensen=tempd(2);

OPTIONS=optimset(@fminbnd);
dMW=fminbnd(@NcllhMW, -0.5, 0.5, OPTIONS, j0, w, J);
sigMW=findSig2epsHat(dMW, j0, w, J);

n=N-2^(J-1);        %the first n data of w, approximation of variance

%function mat2bugs() converts matlab variable to BUGS data file
mat2bugs('c:\WorkDir\LRD_data.txt', 'w',w,' twopowl', 2^j0, 'n', n, 'N', N,
        'resolution', resolution, 'J', J, 'pi', pi,'K',500);

%set the current directory at MATLAB to 'C:\Program Files\WinBUGS14\'
cd 'C:\Program Files\WinBUGS14\';

%prior (a)
dos('WinBUGS14 /par BWIdSt_a.odc');
Sa=bugs2mat('C:\WorkDir\bugsIndex.txt', 'C:\WorkDir\bugs1.txt');

dBS.a=mean(Sa.d);       %the posterior mean as the estimate of d
sigBS.a=mean(Sa.sig2eps);          %the posterior mean as the estimate of sig2eps

%prior (b)
dos('WinBUGS14 /par BWIdSt_b.odc');
Sb=bugs2mat('C:\WorkDir\bugsIndex.txt', 'C:\WorkDir\bugs1.txt');

dBS.b=mean(Sb.d);       %the posterior mean as the estimate of d
sigBS.b=mean(Sb.sig2eps);          %the posterior mean as the estimate of sig2eps

cd 'C:\WorkDir';

function y=NcllhMW(d, j0, w, J);
%NcllhMW.m --- Negative Concentrated log likelihood of McCoy & Walden
%
%input:
%d: the long memory parameter, a value in (0,0.5)
%j0: Lowest Resolution Level
%w: w=Wx, x is the observed time series
%J: N=2^J sample size
%
%output:
%y: Negative Concentrated log likelihood for the given data w
```

```
  m=J-j;
  bmP(j+1)=2*4^(-d)*quad(@sinf,2^(-m-1),2^(-m),[],[],d);
  %by McCoy & Walden's formula, P37, (2.9)
  smP(j+1)=2^m*bmP(j+1);
end;

bpp1P=gamma(1-2*d)/((gamma(1-d))^2)-sum(bmP);
%B_{p+1} in McCoy & Walden's notation, p=J here
spp1P=2^J*bpp1P*(bpp1P>0);
%S_{p+1} in McCoy & Walden's notation, it should be nonnegative

if spp1P>0
  sig2epsHat=w(1)^2/spp1P;
else
   sig2epsHat=0;
end;
sumlogsmP=0;
for j = j0:(J-1)
   sig2epsHat=sig2epsHat+sum(w(2^j+1 : 2^(j+1)).^2)/smP(j+1);
   sumlogsmP=sumlogsmP+2^j*log(smP(j+1));
end;
sig2epsHat=sig2epsHat/N;
%McCoy & Walden, Page 49, formular (5.1)

y=N*log(sig2epsHat)+log(spp1P)+sumlogsmP;
%McCoy & Walden, Page 49

function sig2epsHat=findSig2epsHat(d, j0, w, J);
%find Sig2epsHat by McCoy & Walden Page 49, formular (5.1)
%
%input:
%d: the long memory parameter, a value calculated by function NcllhMW();
%j0: Lowest Resolution Level
%J: N=2^J sample size

N=2^J;

bmP=[];
smP=[];
for j=j0:(J-1)  %j is the resolution level
  m=J-j;
  bmP(j+1)=2*4^(-d)*quad(@sinf,2^(-m-1),2^(-m),[],[],d);
  %by McCoy & Walden's formula, P37, (2.9)
  smP(j+1)=2^m*bmP(j+1);
end;

bpp1P=gamma(1-2*d)/((gamma(1-d))^2)-sum(bmP);
%B_{p+1} in McCoy & Walden's notation, p=J here
spp1P=2^J*bpp1P*(bpp1P>0);
%S_{p+1} in McCoy & Walden's notation, it should be nonnegative
```

```
N=2^J;
bmP=[];
smP=[];
for j=j0:(J-1)  %j is the resolution level

if spp1P>0
   sig2epsHat=w(1)^2/spp1P;
else
    sig2epsHat=0;
end;

for j = j0:(J-1)
    sig2epsHat=sig2epsHat+sum(w(2^j+1 : 2^(j+1)).^2)/smP(j+1);
end;

sig2epsHat=sig2epsHat/N;
\end{verbatim}
```

The WinBUGS script file: BWIdSt_a.odc

```
check('C:/MyDir/LRD_model_a.odc')
data('C:/MyDir/LRD_data.txt')
compile(1)
gen.inits()
update(100)
set(d)
set(sig2eps)
update(500)
coda(*, 'C:/Documents and Settings/MyDir/bugs')
#save('C:/Documents and Settings/MyDirlog.txt')
quit()
```

The WinBUGS model file: LRD_model_a.odc

```
model {
# This takes care of the father wavelet coefficients from level L+1 to J-1
# which are detailed wavelet coefficients, $D$
for (i in twopowl+1:n) {
    tau[i]<-1/(pow(2*pi, -2*d)*sig2eps*pow(2, 2*d*(J-resolution[i])) *(2-pow(2,2*d))/(1-2*d))
    w[i] ~ dnorm (0, tau[i])
   }
```

```
#The following takes care the wavelet coefficients at the resolution level J-1.
#It uses the exact formula instead of the approximation.
for (i in 1:K) { sinf[i]<-pow(sin(pi*(0.25+i/(4*K))),-2*d)}
integration<-sum(sinf[])/(4*K)
B1<-2*pow(4,-d)*sig2eps*integration

tau1<-1/(2*B1)            #S_1=2*B_1   in McCoy & Walden 1996's notation


for (i in (n+1): N) {
     w[i] ~ dnorm (0, tau1)
  }

# This takes care of the scaling coefficients on the lowest level $j_0=L$
#  which are mother wavlelet coefficients, $C$

#  twopowl <- pow(2, L)

for (jp1 in 1:J-1) {             #jp1=j+1, m=J-j
   b[jp1]<-(2*pow(2*pi, -2*d)*sig2eps*pow(2, -(J-jp1+1)*(1-2*d)) *(1-pow(2,2*d-1))/(1-2*d))
 }
   bpp1<-sig2eps*exp(loggam(1-2*d))/pow(exp(loggam(1-d)),2)-sum(b[])-B1;
   #B_{p+1} in McCoy & Walden's notation
   spp1<-pow(2,J)*bpp1*step(bpp1)+1.0E-6;
   #S_{p+1} in McCoy & Walden's notation, this should be positive

   tau0 <- 1/spp1
   for (i in 1: twopowl) {
    w[i] ~ dnorm(0, tau0)
    }

#note: m=J-resolution[i]  in McCoy & Walden's 1996 paper

# prior (a)
  d~dunif(-0.5, 0.5)
  sig2eps<-1/ tau2
   tau2~dgamma(1.0E-2,1.0E-2)

#prior (b)
#   d~dunif(-0.5, 0.5)
#  sig2eps~dunif(0,1000)
}
```

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 155-162

1538 – 9472/05/$95.00

Model-Selection-Based Monitoring Of Structural Change

Kosei Fukuda
College of Economics
Nihon University, Japan

Monitoring structural change is performed not by hypothesis testing but by model selection using a modified Bayesian information criterion. It is found that concerning detection accuracy and detection speed, the proposed method shows better performance than the hypothesis-testing method. Two advantages of the proposed method are also discussed.

Key words: Modified Bayesian information criterion, model selection, monitoring, structural change

Introduction

Deciding whether a time series has a structural change is tremendously important for forecasters and policymakers. If the data generating process (DGP) changes in ways not anticipated, then forecasts lose accuracy. In the real world, not only historical analysis but also real-time analysis should be performed, because new data arrive steadily and the data structure changes gradually. Given a previously estimated model, the arrival of new data presents the challenge of whether yesterday's model can explain today's data. This is why real-time detection of structural change is an essential task. Such forward-looking methods are closely related to the sequential test in the statistics literature but receive little attention in econometrics except for Chu, Stinchcombe, and White (1996) and Leisch, Hornik, and Kuan (2000).

Chu et al. (1996) has proposed two tests for monitoring potential structural changes: the fluctuation and CUSUM monitoring tests. In their fluctuation test, when new observations are obtained, estimates are computed sequentially from all available data (historical and newly obtained sample) and compared to the estimate based only on the historical sample. The null hypothesis of no change is rejected if the difference between these two estimates becomes too large. One drawback of their test is however that it is less sensitive to a change occurring late in the monitoring period.

Leisch et al. (2000) proposed the generalized fluctuation test which includes the fluctuation test of Chu et al. (1996) as a special case and shown that their tests have roughly equal sensitivity to a change occurring early or late in the monitoring period. Two drawbacks of their test are however that there is no objective criterion in selecting the window sizes, and that it has low power in the case of small samples.

In this article, a model-selection-based monitoring of structural change is presented. The existence of structural change is examined, not by hypothesis testing but by model selection using a modified Bayesian information criterion proposed by Liu, Wu, and Zidek (1997). Liu et al. (1997) presented segmented linear regression model and proposed model-selection method in determining the number and location of changepoints. Their criterion has been applied to examine what happened in historical data sets while it has not been applied to examine what happens in real time.

Therefore this criterion is applied to monitor structural change. In this method, whether the observed time series contains a structural change is determined as a result of model selection from a battery of alternative models with and without structural change.

Kosei Fukuda is Associate Professor of Economics at Nihon University, Japan. He has served as an economist in the Economic Planning Agency of the Japanese government (1986-2000). Email: fukuda@eco.nihon-u.ac.jp

Another contribution of this article is the introduction of minimum length of each segment (L). Liu et al. (1997) pay little attention to this topic and make the minimum length equivalent to the number of explanatory variables. This possibly leads to over-fit problem in samples. In order to overcome this problem, $L = 10$ is set arbitrarily and practically in simulations and obtain better performance than the Liu et al. method.

The rest of the article is organized as follows. The hypothesis- testing method and the model-selection method are reviewed briefly. Next, simulation results are shown to illustrate the efficacy of the proposed method. Finally, conclusions and discussions are presented.

Methodology

Leisch et al. (2000) hypothesis-testing method considered the following regression model

$$y_i = x_i'\beta_i + \varepsilon_i, \ i = 1,...,T,T+1,..., \qquad (1)$$

where x_i is the $n \times 1$ vector of explanatory variables, and ε_i is a i.i.d. disturbance term. Suppose an economist is currently at time T and has observed historical data $(y_i, x_i')', i = 1,...,T$. He takes as given that the parameter vector β_i was constant and unknown historically. Consider testing the null hypothesis that β_i remains constant against the alternative that β_i changes at some unknown point in the future.

Leisch et al. (2000) first considered tests based on recursive estimates and show that Chu et al. (1996) fluctuation test is a special case of this class of tests. They write the Chu et al. fluctuation test as

$$\max - RE_T(\tau) = \max_{k=T+1,...,[T\tau]} \frac{k}{\hat{\sigma}_T\sqrt{T}} \left\| Q_T^{1/2}(\hat{\beta}_k - \hat{\beta}_T) \right\| \qquad (2)$$

where $\hat{\beta}_k = \left(\sum_{i=1}^{k} x_i x_i' \right)^{-1} \sum_{i=1}^{k} x_i y_i$,

$Q_T = \frac{1}{T}\sum_{i=1}^{T} x_i x_i'$, $\hat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta}_T)^2$, and $\|\bullet\|$ denotes the maximum norm. The period from time $T+1$ through $[T\tau]$, $\tau > 1$, is the expected monitoring period. For a suitable boundary function q,

$$\lim_{T\to\infty} P\left\{ \frac{k}{\hat{\sigma}_T\sqrt{T}} \left\| Q_T^{1/2}(\hat{\beta}_k - \hat{\beta}_T) \right\| < q(k/T), \ \text{for all } T+1 \le k \le [T\tau] \right\}$$

$$= P\left\{ \left\| W^0(t) \right\| < q(t), \ \text{for all } 1 \le t \le \tau \right\}, \qquad (3)$$

where W^0 is the generalized Brownian bridge on $[0, \infty]$, as shown by Chu et al. (1996), and

$$q(t) = \sqrt{t(t-1)\left[a^2 + \log\left(\frac{t}{t-1}\right)\right]}, \qquad (4)$$

where $t = k/T$. The limiting distribution of $\max - RE_T(\tau)$ is thus determined by the boundary crossing probability of W^0 on $[1, \tau]$. Choosing $a^2 = 7.78$ and $a^2 = 6.25$ gives 95% and 90% monitoring boundaries, respectively.

Leisch et al. (2000) next considered tests based on moving estimates. Define the moving OLS estimates computed from windows of a constant size $[Th]$, where $0 < h \le 1$ and $[Th] > n$, as

$$\tilde{\beta}_T(k,[Th]) = \left(\sum_{i=k-[Th]+1}^{k} x_i x_i' \right)^{-1} \sum_{i=k-[Th]+1}^{k} x_i y_i,$$

$$k = [Th],[Th]+1,... \qquad (5)$$

They propose tests on the maximum and range of the fluctuation of moving estimates:

$$\max - ME_{T,h}(\tau)$$
$$= \max_{k=T+1,\ldots,[T\tau]} \frac{[Th]}{\hat{\sigma}_T\sqrt{T}} \left\| Q_T^{1/2}(\tilde{\beta}_T(k,[Th]) - \hat{\beta}_T) \right\|, \tag{6}$$

$$range - ME_{T,h}(\tau)$$
$$= \max_{i=1,\ldots,n} \frac{[Th]}{\hat{\sigma}_T\sqrt{T}} \Bigg(\max_{k=T+1,\ldots,[T\tau]} [Q_T^{1/2}(\tilde{\beta}_T(k,[Th]) - \hat{\beta}_T)]_i$$
$$- \min_{k=T+1,\ldots,[T\tau]} [Q_T^{1/2}(\tilde{\beta}_T(k,[Th]) - \hat{\beta}_T)]_i \Bigg). \tag{7}$$

The following asymptotic results are obtained:

$$\lim_{T\to\infty} P\Bigg\{ \frac{[\mathrm{Th}]}{\hat{\sigma}_\mathrm{T}\sqrt{T}} \left\| Q_T^{1/2}(\tilde{\beta}_T(k,[Th]) - \hat{\beta}_T) \right\|$$
$$< z(h)\sqrt{2\log_+(k/T)},$$
$$\text{for all } T+1 \le k < [T\tau] \Bigg\} = [F_1(z(h),\tau)]^n; \tag{8}$$

$$\lim_{T\to\infty} P\Bigg\{ \max_{i=1,\ldots,n} \frac{[Th]}{\hat{\sigma}_T\sqrt{T}} \Bigg(\max_{k=T+1,\ldots,J} [Q_T^{1/2}(\tilde{\beta}_T(k,[Th]) - \hat{\beta}_T)]_i$$
$$- \min_{k=T+1,\ldots,J} [Q_T^{1/2}(\tilde{\beta}_T(k,[Th])$$
$$- \hat{\beta}_T)]_i \Bigg) < z(h)\sqrt{2\log_+(J/T)},$$
$$\text{for all } T+1 \le J < [T\tau] \Bigg\} = [F_2(z(h),\tau)]^n; \tag{9}$$

where $\log_+ t = 1$ if $t \le e$, $\log_+ t = \log t$ if $t > e$. In contrast with the boundary-crossing probability of (4), the $F_i(i=1,2)$ do not have analytic forms. Nevertheless, the critical values $z(h)$ can be obtained via simulations, and some typical values are shown in Leisch et al. (2000).

Liu et al. model-selection method

Liu et al. (1997) considered the following segmented linear regression model

$$y_t = x_t\beta_i + \varepsilon_t,\ t = T_{i-1}+1,\ldots,T_i,\ i = 1,\ldots,m+1, \tag{10}$$

where $T_0 = 0$ and $T_{m+1} = T$.

The indices $(T_1,\ldots,T_m)$, or the changekpoints, are explicitly treated as unknown. In addition, the following conditions are newly imposed:

$$T_i - T_{i-1} \ge L \ge n \text{ for all } i\,(i = 1,\ldots,m+1). \tag{11}$$

Changepoints too close to each other or to the beginning or end of the sample cannot be considered, as there are not enough observations to identify the subsample parameters.

In subsequent simulations, comparisons are made between $L = 1$ and $L = 10$ in the case of $n = 1$, and it is concluded that the latter shows better performance than the former.

The purpose of this method is to estimate the unknown parameter vector β_i together with the changepoints when T observations on y_t are available. Their estimation method is based on the least-squares principle. The estimates of the regressive parameters and the changepoints are all obtained by minimizing the sum of squared residuals

$$S_T(T_1,\ldots,T_m) = \sum_{i=1}^{m+1} \sum_{t=T_{i-1}+1}^{T_i} (y_t - x_t\beta_i)^2. \tag{12}$$

Liu et al. (1997) estimated m, the number of changepoints, and $T_1,\ldots,T_m$, by minimizing the modified Schwarz's criterion (Schwarz 1978)

$$\text{LWZ} = T\ln S_T(\hat{T}_1,\ldots,\hat{T}_m)/(T-q) + qc_0(\ln(T))^{2+\delta_0} \tag{13}$$

where $q = n(m+1)+m$, and c_0 and δ_0 are some constants such as $c_0 > 0$ and $\delta_0 > 0$. Liu et al. (1997) recommended

using $\delta_0 = 0.1$ and $c_0 = 0.299$; here small simulations are implemented to examine how the structural change detection is affected by changing these two parameter values in the next Section. This criterion is an extended version of Yao (1988) such as

$$\text{YAO} = T \ln S_T(\hat{T}_1,...,\hat{T}_m)/T + q\ln(T). \quad (14)$$

So, LWZ and YAO differ in the severity of their penalty for overspecification. In general, in model selection, a relatively large penalty term would be preferable for easily identified models. A large penalty will greatly reduce the probability of overestimation while not unduly risking underestimation. Because the optimal penalty is model dependent, however, no globally optimal pair of (c_0, δ_0) can be recommended.

In subsequent simulations, some alternative pairs of (c_0, δ_0) are considered and compared in selecting structural change models. In the model-selection method using the LWZ criterion in the case of possibly one structural change, for example, the following procedure is carried out. First, the OLS estimation for no structural change model ($m = 0$ in equ. 10) is performed, and the LWZ value is stored. Next the OLS estimations for one structural change models obtained by changing the changepoint on the condition of (11) are carried out, and the LWZ values are stored. Finally, the best model is selected using the minimum LWZ procedure from alternative models with and without structural change.

Results

Historical analysis of the structural change using the Liu et al. criterion.

Liu et al. recommended setting the parameters in their information criterion as $\delta_0 = 0.1$ and $c_0 = 0.299$, but they have not shown the efficacy of these parameter values via simulations in which several alternative pairs of (c_0, δ_0) are considered. Such simulations are implemented. The following two DGPs are considered:

DGP 1: $y_t = 2 + e_t$, $t = 1,...,T$,

DGP 2: $y_t = 2 + e_t$ if $t \le T/2$, $y_t = 2.8 + e_t$ if $t > T/2$,

where e_t is generated from i.i.d. $N(0,1)$. Considered are historical samples of sizes $T = 50$, 100, 200, 400, $L = 1$, 10, $c_0 = 0.01$, 0.05, 0.1, 0.3, 0.5, and $\delta_0 = 0.01$, 0.05, 0.1, 0.2. The number of replications is 1,000.

Table 1 shows frequency counts of selecting structural change models using the Liu et al. information criterion. First consider comparing the performances between two pairs of $(c_0 = 0.1, \delta_0 = 0.05)$ and $(c_0 = 0.299, \delta_0 = 0.1)$.

The former significantly outperforms the latter, particularly in the structural-change cases of $T = 50$ and 100. The pair of $(c_0 = 0.299, \delta_0 = 0.1)$ imposes too heavy penalty to select structural change models correctly. Next consider comparing between $L = 1$ and $L = 10$. The latter outperforms the former, particularly in small samples of $T = 50$ and 100. In the case of $L = 1$, it happens to occur that a structural change is incorrectly detected in the beginning or end of the sample.

Monitoring structural change via the Leisch et al. simulations

In Leisch et al. (2000), the DGP for examining empirical size is the same as DGP 1. They show the performances of $\max - RE$, $\max - ME$, and $range - ME$ tests and consider moving window sizes $h = 0.25$, 0.5, 1, and $\tau = 10$ for the expected monitoring period $[T\tau]$.

However, the DGP for examining empirical power is not the same as DGP2. The mean changes from 2.0 to 2.8 at $1.1T$ or $3T$. Similarly to Leisch et al. (2000), only the results for the 10% significance level are reported. All experiments were repeated 1,000 times.

Table 1

Frequency counts of selecting structural change models

L	T	c_0	δ_0 for DGP1				δ_0 for DGP2			
			0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
1	50	0.01	0.590	0.585	0.583	0.576	0.968	0.968	0.968	0.968
1	50	0.05	0.379	0.360	0.347	0.305	0.919	0.913	0.908	0.893
1	50	0.1	0.192	0.172	0.155	0.128	0.826	0.817	0.798	0.746
1	50	0.3	0.024	0.018	0.012	0.008	0.330	0.301	0.266	0.200
1	50	0.5	0.001	0.000	0.000	0.000	0.080	0.068	0.055	0.028
1	100	0.01	0.653	0.647	0.639	0.623	0.999	0.999	0.999	0.999
1	100	0.05	0.360	0.348	0.327	0.269	0.994	0.994	0.992	0.987
1	100	0.1	0.147	0.137	0.112	0.070	0.959	0.951	0.946	0.925
1	100	0.3	0.003	0.002	0.001	0.001	0.591	0.530	0.456	0.349
1	100	0.5	0.001	0.000	0.000	0.000	0.183	0.142	0.112	0.056
1	200	0.01	0.697	0.692	0.677	0.656	1.000	1.000	1.000	1.000
1	200	0.05	0.297	0.275	0.244	0.186	1.000	1.000	1.000	1.000
1	200	0.1	0.077	0.064	0.047	0.026	0.999	0.999	0.999	0.997
1	200	0.3	0.000	0.000	0.000	0.000	0.906	0.881	0.843	0.707
1	200	0.5	0.000	0.000	0.000	0.000	0.514	0.430	0.331	0.202
1	400	0.01	0.710	0.703	0.684	0.660	1.000	1.000	1.000	1.000
1	400	0.05	0.271	0.240	0.202	0.136	1.000	1.000	1.000	1.000
1	400	0.1	0.047	0.034	0.026	0.014	1.000	1.000	1.000	1.000
1	400	0.3	0.000	0.000	0.000	0.000	0.999	0.999	0.998	0.992
1	400	0.5	0.000	0.000	0.000	0.000	0.952	0.924	0.868	0.706
10	50	0.01	0.368	0.366	0.362	0.352	0.933	0.932	0.931	0.930
10	50	0.05	0.212	0.206	0.200	0.176	0.869	0.863	0.851	0.840
10	50	0.1	0.111	0.101	0.091	0.075	0.763	0.745	0.726	0.694
10	50	0.3	0.009	0.008	0.008	0.003	0.344	0.315	0.279	0.212
10	50	0.5	0.001	0.001	0.001	0.001	0.101	0.080	0.059	0.028
10	100	0.01	0.461	0.457	0.449	0.434	0.993	0.993	0.992	0.992
10	100	0.05	0.230	0.217	0.200	0.164	0.980	0.977	0.974	0.969
10	100	0.1	0.091	0.078	0.068	0.051	0.948	0.943	0.930	0.907
10	100	0.3	0.001	0.000	0.000	0.000	0.552	0.504	0.446	0.342
10	100	0.5	0.000	0.000	0.000	0.000	0.157	0.132	0.094	0.039
10	200	0.01	0.550	0.539	0.531	0.512	1.000	1.000	1.000	1.000
10	200	0.05	0.226	0.210	0.191	0.147	1.000	1.000	0.999	0.999
10	200	0.1	0.071	0.058	0.046	0.028	0.999	0.999	0.999	0.997
10	200	0.3	0.001	0.001	0.000	0.000	0.900	0.875	0.840	0.722
10	200	0.5	0.000	0.000	0.000	0.000	0.514	0.423	0.320	0.176
10	400	0.01	0.596	0.587	0.580	0.552	1.000	1.000	1.000	1.000
10	400	0.05	0.176	0.155	0.124	0.094	1.000	1.000	1.000	1.000
10	400	0.1	0.033	0.026	0.019	0.006	1.000	1.000	1.000	1.000
10	400	0.3	0.000	0.000	0.000	0.000	0.999	0.999	0.997	0.981
10	400	0.5	0.000	0.000	0.000	0.000	0.942	0.918	0.869	0.696

One fundamental difference between the Leisch et al. method and the proposed method is whether the changepoint is estimated. In the Leisch et al. method, the changepoint estimation cannot be performed. In order to do so, another step is needed. As in Chu et al. (1996), for example, it is possible to define the changepoint by the point at which the maximum of the LR statistics is obtained for the period from the starting point to the first hitting point. In contrast, the proposed method presents not only the first hitting point but also the changepoint simultaneously. This is because in the proposed method, from a battery of alternative models obtained by changing the changepoint on the condition of (11), including no structural change model, the best model is selected in each monitoring point. Therefore, the proposed method is very computer intensive.

Table 2 shows frequency counts of selecting structural change models. In the LWZ criterion used are a pair of $(c_0 = 0.1, \delta_0 = 0.05)$, considering the results of the preceding simulation results. In the cases of no structural change, the YAO criterion ($L = 1$ and $L = 10$) and the LWZ criterion ($L = 1$) show poor performance. In contrast, the performance of the LWZ criterion ($L = 10$) is comparable to other hypothesis-testing methods. In addition, it is shown that the more samples are obtained, the better performances are realized, because larger penalty ($\ln(T)^{2.05}$) is imposed in the LWZ criterion than in the YAO criterion ($\ln(T)$).

In the cases of structural change, the proposed method using the LWZ criterion ($L = 10$) outperforms other hypothesis-testing methods, particularly in the late change case. The max – *ME*, and *range* – *ME* tests with small window sizes of $h = 1/4$ and $h = 1/2$ shows poor performances in small samples.

More interesting features are shown in Table 3. Concerning the mean of detection delay, the proposed method using the LWZ criterion ($L = 10$) significantly outperforms other hypothesis-testing methods. One fundamental drawback of the Leisch et al. method is that it remains unknown how small h should be. The smaller h is used, the quicker detection is obtained, but the lower power is also realized.

Conclusion

In this article, a model-selection-based monitoring of structural change was presented. The existence of structural change was examined not by hypothesis testing but by model selection using a modified Bayesian information criterion proposed by Liu, Wu, and Zidek (1997). It was found that concerning detection accuracy and detection speed, the proposed method shows better performance than the hypothesis-testing method of Leisch, Hornik, and Kuan (2000).

This model-selection-based method has two advantages in comparison to the hypothesis-testing method. First, by the introduction of a modified Bayesian information criterion, the subjective judgment required in the hypothesis-testing procedure for determining the levels of significance is completely eliminated, and a semiautomatic execution becomes possible. Second, the model-selection-based method frees time series analysts from complex works of hypothesis testing. In order to provide better data description, different alternative models should usually be considered by changing the number of structural changes. In the conventional framework of hypothesis testing, however, different alternative models lead to different test statistics (Bai & Perron, 1998). In the model-selection method, any model change can be made very simply and the performance of the new model is evaluated consistently using the information criterion.

References

Bai, J., Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica, 66*, 47-78.

Chu, C-S. J., Stinchcombe, M., White, H. (1996). Monitoring structural change. *Econometrica, 64*, 1045-1065.

Leisch, F., Hornik, K., Kuan, C-M., (2000). Monitoring structural changes with the generalized fluctuation test. *Econometric Theory, 16*, 835-854.

Table 2. Frequency counts of selecting structural change models

T	CP	YAO		LWZ		max-RE	max-ME			range-ME		
		L=1	L=10	L=1	L=10		h=1/4	h=1/2	h=1	h=1/4	h=1/2	h=1
25		0.838	0.401	0.424	0.146	0.088	0.091	0.104	0.121	0.058	0.065	0.081
50		0.822	0.472	0.245	0.104	0.078	0.090	0.108	0.105	0.051	0.064	0.049
100		0.852	0.538	0.138	0.067	0.073	0.109	0.109	0.109	0.065	0.055	0.061
200		0.840	0.582	0.054	0.031	0.084	0.090	0.105	0.108	0.054	0.053	0.045
300		0.868	0.590	0.020	0.012	0.087	0.090	0.098	0.103	0.060	0.055	0.042
25	28	0.986	0.961	0.941	0.890	0.931	0.685	0.832	0.925	0.108	0.277	0.660
50	55	1.000	0.999	0.994	0.989	0.996	0.948	0.992	1.000	0.206	0.640	0.950
100	110	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.607	0.966	0.999
200	220	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.955	0.999	1.000
300	330	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000
25	75	0.999	0.996	0.994	0.994	0.691	0.445	0.660	0.823	0.034	0.100	0.417
50	150	1.000	1.000	1.000	1.000	0.953	0.828	0.966	0.992	0.072	0.386	0.858
100	300	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.320	0.847	0.999
200	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.824	0.998	1.000
300	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.979	1.000	1.000

Note: CP denotes change point.

Table 3. The mean and standard deviation of detection delay

T	Change point	YAO		LWZ		max-RE
		L=1	L=10	L=1	L=10	
25	28	16(23)	25(26)	17(24)	25(25)	28(36)
50	55	15(18)	21(21)	20(26)	27(30)	25(27)
100	110	15(11)	19(11)	22(18)	25(19)	24(16)
200	220	16(11)	19(10)	24(16)	26(15)	27(14)
300	330	16(11)	19(10)	27(15)	27(14)	30(15)
25	75	15(13)	19(13)	21(20)	25(20)	69(45)
50	150	15(11)	19(11)	23(17)	25(17)	104(69)
100	300	16(10)	19(9)	26(14)	27(13)	127(75)
200	600	17(10)	19(10)	30(15)	31(14)	147(73)
300	900	18(10)	20(9)	34(15)	34(15)	165(80)

T	Change point	max-ME			range-ME		
		h=1/4	h=1/2	h=1	h=1/4	h=1/2	h=1
25	28	22(22)	23(25)	24(19)	32(10)	30(14)	33(21)
50	55	30(32)	26(26)	32(19)	49(15)	44(26)	46(29)
100	110	25(16)	30(11)	44(13)	73(43)	60(47)	58(14)
200	220	30(10)	42(13)	62(16)	75(63)	66(20)	82(16)
300	330	37(11)	53(15)	76(19)	71(50)	80(15)	102(18)
25	75	26(28)	27(29)	29(28)	37(8)	35(13)	40(22)
50	150	39(47)	34(38)	37(29)	48(16)	55(32)	61(45)
100	300	30(28)	33(19)	47(18)	74(43)	77(67)	69(33)
200	600	31(12)	45(15)	64(25)	98(114)	74(24)	91(17)
300	900	38(12)	55(18)	78(30)	84(86)	87(17)	111(19)

Note: The number in each parenthesis indicates standard deviation.

Liu, J., Wu, S., Zidek, J. V. (1997). On segmented multivariate regressions. *Statistica Sinica*, *7*, 497-525.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters*, *6*, 181-189.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 163-171

1538 – 9472/05/$95.00

On The Power Function Of Bayesian Tests With Application To Design Of Clinical Trials: The Fixed-Sample Case

Lyle Broemeling
Department of Biostatistics and Applied Mathematics
University of Texas MD Anderson Cancer Center

Dongfeng Wu
Department of Mathematics and Statistics
Mississippi State University

Using a Bayesian approach to clinical trial design is becoming more common. For example, at the MD Anderson Cancer Center, Bayesian techniques are routinely employed in the design and analysis of Phase I and II trials. It is important that the operating characteristics of these procedures be determined as part of the process when establishing a stopping rule for a clinical trial. This study determines the power function for some common fixed-sample procedures in hypothesis testing, namely the one and two-sample tests involving the binomial and normal distributions. Also considered is a Bayesian test for multi-response (response and toxicity) in a Phase II trial, where the power function is determined.

Key words: Bayesian; power analysis; sample size; clinical trial

Introduction

The Bayesian approach to testing hypotheses is becoming more common. For example, in a recent review volume, see Crowley (2001), many contributions where Bayesian considerations play a prominent role in the design and analysis of clinical trials. Also, in an earlier Bayesian review (Berry & Stangl, 1996), methods are explained and demonstrated for a wide variety of studies in the health sciences, including the design and analysis of Phase I and II studies.

At our institution, the Bayesian approach is often used to design such studies. See Berry (1985,1987,1988), Berry and Fristed (1985), Berry and Stangl (1996), Thall and Russell (1998), Thall, Estey, and Sung (1999), Thall, Lee, and Tseng (1999), Thall and Chang (1999),and Thall et al. (1998), for some recent references where Bayesian ideas have been the primary consideration in designing Phase I and II studies. Of related interest in the design of a trial is the estimation of sample size based on Bayesian principles, where Smeeton and Adcock (1997) provided a review of formal decision-theoretic ideas in choosing the sample size.

Typically, the statistician along with the investigator will use information from previous related studies to formulate the null and alternative hypotheses and to determine what prior information is to be used for the Bayesian analysis. With this information, the Bayesian design parameters that determine the critical region of the test are given, the power function calculated, and lastly the sample size determined as part of the design process. In this study, only fixed-sample size procedures are used.

First, one-sample binomial and normal tests will be considered, then two-sample tests for binomial and normal populations, and lastly a test for multinomial parameters of a multi-response Phase II will be considered. For each test, the null and alternative hypotheses will be formulated and the power function determined. Each case will be illustrated with an example, where the power function is calculated for several values of the Bayesian design parameters.

Lyle Broemeling is Research Professor in the Department of Biostatistics and Applied Mathematics at the University of Texas MD Anderson Cancer Center. Email: lbroemel@mdanderson.org. Dongfeng Wu got her PhD from the University of California, Santa Barbara. She is an assistant professor at Mississippi State University.

Methodology

For the design of a typical Phase II trial, the investigator and statistician use prior information on previous related studies to develop a test of hypotheses. If the main endpoint is response to therapy, the test can be formulated as a sample from a binomial population, thus if Bayesian methods are to be employed, prior information for a Beta prior must be determined. However, if the response is continuous, the design can be based on a one-sample normal population. Information from previous related studies and from the investigator's experience will be used to determine the null and alternative hypotheses, as well as the other design parameters that determine the critical region of the test.

The critical region of a Bayesian test is given by the event that the posterior probability of the alternative hypothesis will exceed some threshold value. Once a threshold value is used, the power function of the test can be calculated. The power function of the test is determined by the sample size, the null and alternative hypotheses, and the above-mentioned threshold value.

Results

Binomial population

Consider a random sample from a Bernoulli population with parameters n and θ, where n is the number of patients and θ is the probability of a response. Let X be the number of responses among n patients, and suppose the null hypotheses is H: $\theta \le \theta_0$ versus the alternative A: $\theta > \theta_0$. From previous related studies and the experience of the investigators, the prior information for θ is determined to be Beta(a,b), thus the posterioir distribution of θ is Beta (x+a, n-x+b), where x is the observed number of responses among n patients. The null hypothesis is rejected in favor of the alternative when

$$\Pr[\theta > \theta_0 / x, n] > \gamma, \qquad (1)$$

where γ is usually some large value as .90, .95, or .99. The above equation determines the critical region of the test, thus the power function of the test is

$$g(\theta) = \Pr_{X/\theta} \{\Pr[\theta > \theta_0 / x, n] > \gamma\}, \qquad (2)$$

where the outer probability is with respect to the conditional distribution of X given θ. The power (2) at a given value of θ is interpreted as a simulation as follows:

(a) select (n, θ), and set S=0,

(b) generate a X~Binomial(n, θ),

(c) generate a θ~Beta(x+a, n-x+b),

(d) if Pr $[\theta > \theta_0 / x, n] > \gamma$, let the counter S =S+1, otherwise let S=S,

(e) repeat (b)-(d) M times, where M is 'large',

and

(f) select another θ and repeat (b)-(d).

The power of the test is thus S/M and can be used to determine a sample size by adjusting the threshold γ, the probability of a Type I error g(θ_0), and the desired power at a particular value of the alternative. The approach taken is fixing the Type I error at α and finding n so that the power is some predetermined value at some value of θ deemed to be important by the design team. This will involve adjusting the critical region by varying the value of the threshold γ. An example of this method is provided in the next section. The above hypotheses are one-sided, however it is easy to adjust the above testing procedure for a sharp null hypothesis.

Normal Population

Let N(θ, τ^{-1}) denote a normal population with mean θ and precision τ, where both are unknown and suppose we want to test the null hypothesis H: $\theta = \theta_0$ versus A: $\theta \neq \theta_0$, based on a random sample X of size n

with sample mean $\bar{X}$ and variance S^2. Using a non-informative prior distribution for θ and τ, the Bayesian test is to reject the null in favor of the alternative if the posterior probability P of the alternative hypothesis satisfies

$$P > \gamma, \text{ where} \quad (3)$$

$$P = D_2/D \quad (4)$$

and, $D = D_1 + D_2$.

It can be shown that

$$D_1 = \{\pi\Gamma(n/2)\}2^{n/2}\} / \{(2\pi)^{n/2} [\, n(\theta_0 - \bar{X})^2 + (n-1)\, S^2]^{n/2}\} \quad (5)$$

and

$$D_2 = \{(1-\pi)\Gamma((n-1)/2)\, 2^{(n-1)/2}\} / \{n^{1/2}(2\pi)^{(n-1)/2}[(n-1)\, S^2]^{(n-1)/2}\} \quad (6)$$

where π is the prior probability of the null hypothesis.

The power function of the test is

$$g(\theta,\tau) = \Pr_{X/\theta,\tau}[\, P > \gamma / \text{ n}, \bar{X}, S^2], \quad \theta \in R \text{ and } \tau > 0 \quad (7)$$

where P is given by (3) and the outer probability is with respect to the conditional distribution of X given θ and τ.

The above test is for a two-sided alternative, but the testing procedure is easily revised for one-sided hypotheses. This will be used to find the sample size in an example to be considered in a following section.

In the case when the null and alternative hypotheses are H: $\theta \le \theta_0$ and A: $\theta > \theta_0$ and the prior distribution for the parameters is $f(\theta,\tau) \propto 1/\tau$, where H is rejected in favor of A whenever

$$\Pr[\theta > \theta_0 / \text{n}, \bar{X}, S^2] > \gamma,$$

it can be shown that the power (size) of the test at θ_0 is 1-γ. Thus in this sense, the Bayesian and classical t-test are equivalent.

Two binomial populations

Comparing two binomial populations is a common problem in statistics and involves the null hypothesis H: $\theta_1 = \theta_2$ versus the alternative A: $\theta_1 \neq \theta_2$, where θ_1 and θ_2 are parameters from two Bernoulli populations. Assuming uniform priors for these two populations, it can be shown that the Bayesian test is to reject H in favor of A if the posterior probability P of the alternative hypothesis satisfies

$$P > \gamma, \text{ where} \quad (8)$$

$$P = D_2/D, \quad (9)$$

and $D = D_1 + D_2$. It can be shown that

$$D_1 = \{\pi \text{BC}(n_1 : x_1)\text{BC}(n_2 : x_2) \Gamma(x_1 + x_2 + 1)\Gamma(n_1 + n_2 - x_1 - x_2)\} \div \Gamma(n_1 + n_2 + 2),$$

where BC(n,x) is the binomial coefficient "x from n". Also, $D_2 = (1-\pi)(n_1+1)^{-1}(n_2+1)^{-1}$, where π is the prior probability of the null hypothesis. X_1 and X_2 are the number of responses from the two binomial populations with parameters (n_1, θ_1) and (n_2, θ_2) respectively. The alternative hypothesis is two-sided, however the testing procedure is easily revised for one-sided hypotheses.

In order to choose sample sizes n_1 and n_2, one must calculate the power function

$$g(\theta_1,\theta_2) = \Pr_{x_1,x_2/\theta_1,\theta_2}[P > \gamma / x_1, x_2, n_1, n_2], (\theta_1,\theta_2) \in (0,1)\,\text{x}\,(0,1) \quad (10)$$

where P is given by (9) and the outer probability is with respect to the conditional distribution of X_1 and X_2, given θ_1 and θ_2. As given above, (10) can be evaluated by a simulation procedure similar to that described in 3.1.

Two normal populations

Consider two normal populations with means θ_1 and θ_2 and precisions τ_1 and τ_2 respectively, and suppose the null and alternative hypotheses are H: $\theta_1 \leq \theta_2$ and A: $\theta_1 > \theta_2$ respectively. Assuming a non-informative prior for the parameters, namely $f(\theta_1,\theta_2,\tau_1,\tau_2) = 1/\tau_1\tau_2$, one can show that the posterior distribution of the two means is such that θ_1 and θ_2 are independent and θ_i /data ~ t(n_i-1, $\bar{X}_i$ n_i/S_i^2), where n_i is the sample size and $\bar{X}_i$ and S_i^2 are the sample mean and variance respectively.

That is, the posterior distribution of θ_i is a t distribution with n_i-1 degrees of freedom, mean $\bar{X}_i$, and precision n_i/S_i^2. It is known that $(\theta_i - \bar{X}_i)(n_i/S_i^2)^{1/2}$ has a Student's t-distribution with n_i-1 degrees of freedom.

Therefore the null hypothesis is rejected if

$$\Pr[\theta_1 > \theta_2/\text{data}] > \gamma . \quad (11)$$

Multinomial Populations

Consider a multinomial population with k categories and corresponding probabilities θ_i, i= 1,2,…,k, where $\sum_{i=1}^{i=k} \theta_i = 1$ and $0 < \theta_i < 1$ for i=1,2,…,k. Suppose there are n patients and that n_i belong to the i-th category.

The multinomial model is quite relevant to the Phase II trial where the k categories represent various responses to therapy. Let $\theta = (\theta_1,\theta_2,...,\theta_k)$, then if a uniform prior distribution is appropriate, the posterior distribution is

$$f(\theta/\text{ data}) \propto \prod_{i=1}^{i=k} \theta_i^{n_i}, \quad \sum_{i=1}^{i=k} \theta_i = 1, \text{ and } 0 < \theta_i < 1 \text{ for i=1,2,…,k.} \quad (12)$$

and the distribution is Dirichlet $(n_1+1, n_2+1,..., n_k+1)$.

A typical hypothesis testing problem, see [14], is given by the null hypothesis (k=4), where

$$\text{H: } \theta_1 + \theta_2 \leq k_{12} \text{ or } \theta_1 + \theta_3 \geq k_{13}$$

versus the alternative

$$\text{A: } \theta_1 + \theta_2 > k_{12} \text{ and } \theta_1 + \theta_3 < k_{13} .$$

The null hypothesis states that the response rate $\theta_1 + \theta_2$ is less than some historical value or that the toxicity rate $\theta_1 + \theta_3$ is greater than some historical value k_{13}. The null hypothesis is rejected if the response rate is larger than the historical or the toxicity rate is too low compared to the historical.

$$\Pr[A/\text{data}] > \gamma \quad (13)$$

where γ is some threshold value. This determines the critical region of the test, thus the power function is

$$g(\theta) = \Pr_{n/\theta}\{ \Pr[A/\text{ data}] > \gamma \}, \quad (14)$$

where the outer probability is with respect to the conditional distribution of

$$n = (n_1, n_2,..., n_k) \text{ given } \theta .$$

The power function will be illustrated for the multinomial test of hypothesis with a Phase I trial, where response to therapy and toxicity are considered in designing the trial.

Examples

The above problems in hypothesis testing are illustrated by computing the power function of some Bayesian tests that might be used in the design of a Phase II trial.

One-Sample Binomial

No prior information

Consider a typical Phase II trial, where the historical rate for toxicity was determined as .20. The trial is to be stopped if this rate exceeds the historical value. See Berry (1993) for a good account of Bayesian stopping rules in clinical trials. Toxicity rates are carefully defined in the study protocol and are based on the NCI list of toxicities. The null and alternative hypotheses are given as

$$\text{H: } \theta \le .20 \text{ and A: } \theta > .20, \qquad (15)$$

where θ is the probability of toxicity. The null hypothesis is rejected if the posterior probability of the alternative hypothesis is greater than the threshold value γ.

The power curve for the following scenarios will be computed (see Equation 2), with sample sizes n = 125, 205, and 500, threshold values γ = .90, .95, .99, M=1000, and null value θ_0 = .20.

It is seen that the power of the test at $\theta = .30$ and γ = .95, is .841, .958, and .999 for n = 125, 205, and 500, respectively.

Note that for a given N and γ, the power increases with θ and for given N and θ, the power decreases with γ, and for given γ and θ, the power of course increases with N.

The Bayesian test behaves in a reasonable way. For the conventional type I error of .05, a sample size of N=125 would be sufficient to detect the difference .3 versus .2 with a power of .841. It is interesting to note that the usual binomial test, with alpha = .05 and power .841, requires a sample of size 129 for the same alternative value of θ. For the same alpha and power, one would expect the Bayesian (with a uniform prior for θ) and the binomial tests to behave in the same way in regard to sample size.

With prior information

Suppose the same problem is considered as above, but prior information is available with 50 patients, 10 of whom have experienced toxicity. The null and alternative hypotheses are as above, however the null is rejected whenever

$$\Pr[\theta > \phi / \text{x, n}] > \gamma, \qquad (16)$$

where θ is independent of $\phi \sim$ Beta(10,40). This can be considered as a one-sample problem where a future study is to be compared to a historical control.

As above, compute the power function (see Table 2) of this Bayesian test with the same sample sizes and threshold values in Table 1. The power of the test is .758, .865, and .982 for θ = .4 for N= 125, 205, and 500, respectively. This illustrates how important is prior information in testing hypotheses. If the hypothesis is rejected with the critical region

$$\Pr[\theta > .2 / \text{x, n}] > \gamma, \qquad (17)$$

the power (Table 1) will be larger than the corresponding power (Table 2) determined by the critical region (16), because of the additional posterior variability introduced by the historical information contained in ϕ. Thus, larger sample sizes are required with (16) to achieve the same power as with the test given by (17).

Table 1. Power function for H versus A, N=125,205,500.

θ \ γ	.90	.95	.99
0	0,0,0	0,0,0	0,0,0
.1	0,0,0	0,0,0	0,0,0
.2	.107,.099,.08	.047,.051,.05	.013,.013,.008
.3	.897,.97,1	.841,.958,.999	.615,.82,.996
.4	1,1,1	1,1,1	.996,1,1
.5	1,1,1	1,1,1	1,1,1
.6	1,1,1	1,1,1	1,1,1
.7	1,1,1	1,1,1	1,1,1
.8	1,1,1	1,1,1	1,1,1
.9	1,1,1	1,1,1	1,1,1
1.0	1,1,1	1,1,1	1,1,1

Table 2. Power function for H versus A, N=125,205,500.

θ \ γ	.90	.95	.99
0	0,0,0	0,0,0	0,0,0
.1	0,0,0	0,0,0	0,0,0
.2	.016,.001,.000	.002,.000,.000	.000,.000,.000
.3	.629,.712,.850	.362,.374,.437	.004,.026,.011
.4	.996,.999,1	.973,.998,1	.758,.865,.982
.5	1,1,1	1,1,1	.999,1,1
.6	1,1,1	1,1,1	1,1,1
.7	1,1,1	1,1,1	1,1,1
.8	1,1,1	1,1,1	1,1,1
.9	1,1,1	1,1,1	1,1,1
1.0	1,1,1	1,1,1	1,1,1

Two Binomial Populations

The case of two binomial populations was introduced in section 4.2, where equation (10) gives the power function for testing H: $\theta_1 = \theta_2$ versus the alternative A: $\theta_1 \neq \theta_2$.

In this example, let $n_1 = 20 = n_2$ be the sample sizes of the two groups and suppose the prior probability of the null hypotheses is π = .5. The power at each point (θ_1, θ_2) is calculated via simulation, using equation (10) with γ = .90. Table 3 lists the power function for this test.

When the power is calculated with the usual two-sample, two-tailed, binomial test with alpha = .013, sample sizes $n_1 = 20 = n_2$, and (θ_1, θ_2) = (.3, .9), the power is .922, which is almost equivalent to the above Bayesian test. This is to be expected, because we are using a uniform prior density for both Bernoulli parameters. It is not too uncommon to have two binomial populations in a Phase II setting, where θ_1 and θ_2 are response rates to therapy.

Table 3. Power for Bayesian Binomial Test.

θ_1 \ θ_2	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
.1	.004	.032	.135	.360	.621	.842	.958	.992	1	1
.2	.031	.011	.028	.106	.281	.536	.744	.913	.997	1
.3	.171	.028	.006	.029	.107	.252	.487	.767	.961	1
.4	.368	.098	.025	.013	.028	.075	.244	.542	.847	.999
.5	.619	.289	.100	.022	.007	.017	.108	.291	.640	.981
.6	.827	.527	.237	.086	.035	.005	.027	.116	.357	.882
.7	.950	.775	.464	.254	.113	.037	.013	.049	.171	.587
.8	.996	.928	.768	.491	.316	.132	.028	.010	.040	.205
.9	1	.996	.946	.840	.647	.359	.156	.037	.006	.014
1	1	1	1	1	.984	.873	.567	.200	.017	.000

A Phase II trial with toxicity and response rates

With Phase II trials, response to therapy is usually taken to be the main endpoint, however in reality one is also interested in the toxicity rate, thus it is reasonable to consider both when designing the study. Most Phase II trials are conducted not only to estimate the response rate, but to learn more about the toxicity. In such a situation, the patients can be classified by both endpoints as follows:

Table 4. Number of and Probability of Patients by Response and Toxicity.

Response \ Toxicity	Yes	No
Yes	(n_1, θ_1)	(n_2, θ_2)
No	(n_3, θ_3)	(n_4, θ_4)

Let the response rate be $\theta_r = \theta_1 + \theta_2$ and the rate of toxicity be $\theta_t = \theta_1 + \theta_3$, where θ_1 is the probability a patient will experience toxicity and respond to therapy, and n_1 is the number of patients who fall into that category. Following Petroni and Conoway (2001), let the null hypothesis be

$$\text{H: } \theta_r \le \theta_{r0} \text{ or } \theta_t \ge \theta_{t0}$$

and the alternative hypothesis be

$$\text{A: } \theta_r > \theta_{r0} \text{ and } \theta_t < \theta_{t0},$$

where θ_{r0} and θ_{t0} are given and estimated by the historical rates in previous trials.

Table 5. Power of Bayesian Multinomial Test.

	θ_t	.2	.3	.4	.5
θ_r					
.2		.000	.000	.000	.000
.3		.000	.000	.000	.000
.4		.070	.002	.000	.000
.5		.600	.114	.000	.000
.6		.794	.154	.000	.000
.7		.818	.158	.000	.000
.8		.822	.084	.000	.000

In this example, let θ_{r0} = .40 and θ_{t0} = .30. That is, the alternative hypothesis is that the response rate exceeds .40 and the toxicity rate is less than .30, and the null is rejected in favor of the alternative if the latter has a posterior probability in excess of γ. Table 5 gives the power for n=100 patients and threshold γ = .90.

From above, the power of the test is .818 when (θ_r, θ_t) = (.7, .2), and the test behaves in a reasonable way. When the parameter values are such that the response rate is in excess of .40 and the toxicity rate is less than or equal to .30, the power is higher, relative to those parameter values when the null hypothesis is true.

Conclusion

We have provided a way to assess the sampling properties of some Bayesian tests of hypotheses used in the design and analysis of Phase II clinical trials.

The one-sample binomial scenario is the most common in a Phase II trial, where the response to therapy is typically binary. We think it is important to know the power function of a critical region that is determined by Bayesian considerations, just as it is with any other test.

The Bayesian approach has one major advantage and that is prior information, and when this is used in the design of the trial, the power of the test will be larger then if prior information had not been used.

We have confined this investigation to the fixed-sample case, but will seek to expand the results to the more realistic situation where Bayesian sequential stopping rules will be used to design Phase II studies.

References

Berry D. A. (1985). Interim analysis in clinical trials: Classical versus Bayesian approaches. *Statistics in Medicine*, *4*, 521-526.

Berry D. A. (1987). Interim analysis in clinical trials: the role of the likelihood principal. *The American Statistician*, *41*, 117-122.

Berry D. A. (1988). Interim analysis in clinical research. *Cancer Investigations*, *5*, 469-477.

Berry D. A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine*, *12*, 1377-1404.

Berry D. A., & Fristed, B. (1985). *Bandit problems. Sequential allocation of experiments*. New York: Chapman-Hall.

Berry D. A., & Stangl D. K., (1996). Bayesian methods in health-related research,. *Bayesian Biostatistics*. (In D. A. Berry, & D. Stangl, eds.) New York: Marcel Dekker Inc., p. 3 – 66.

Crowley J. (2001). *Handbook of statistics in clinical oncology*. New York: Marcel Dekker Inc.

Petroni G. R., & Conoway M. R. (2001). Designs based on toxicity and response. (In J. Crowley, ed.) *Handbook of Statistics in Clinical Oncology*. New York: Marcel-Dekker Inc., p. 105 – 118.

Smeeton N. C., & Adcock C. J. (1997, special issue). (eds.) Sample size determination. *The Statistician*, *4*.

Thall P. F, Simon R., Ellenberg S. S., & Shrager R. (1988). Optimal two-stage designs for clinical trials with binary responses. *Statistics in Medicine*, *71*, 571-579.

Thall P. F., & Russell K. E. (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, *54*, 251-264.

Thall P. F., Estey E. H., & Sung H. G. (1999). A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. *Investigational New Drugs*, *17*, 155-167.

Thall P. F., Lee J. J., & Tseng C. H. (1999). Accrual strategies for Phase I trials with delayed patient outcome. *Statistics in Medicine*, *18*, 1155-1169.

Thall P. F., & Cheng S. C. (1999) Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics*, *55*, 746-753.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 172-186

1538 – 9472/05/$95.00

Bayesian Reliability Modeling Using Monte Carlo Integration

Vincent A. R. Camara Chris P. Tsokos
Department of Mathematics
University of South Florida

The aim of this article is to introduce the concept of Monte Carlo Integration in Bayesian estimation and Bayesian reliability analysis. Using the subject concept, approximate estimates of parameters and reliability functions are obtained for the three-parameter Weibull and the gamma failure models. Four different loss functions are used: square error, Higgins-Tsokos, Harris, and a logarithmic loss function proposed in this article. Relative efficiency is used to compare results obtained under the above mentioned loss functions.

Key words: Estimation, loss functions, Monte Carlo Integration, Monte Carlo Simulation, reliability functions, relative efficiency.

Introduction

In this article, the concept of Monte Carlo Integration (Berger, 1985) is used to obtain approximate estimates of the Bayes rule that is ultimately used to derive estimates of the reliability function. Monte Carlo Integration is used to first obtain approximate Bayesian estimates of the parameter inherent in the failure model, and using this estimate directly, obtain approximate Bayesian estimates of the reliability function. Secondly, the subject concept is used to directly obtain Bayesian estimates of the reliability function.

In the present modeling effort, the three-parameter Weibull and the gamma failure models are considered, that are respectively defined as follows:

Vincent A. R. Camara earned a Ph.D. in Mathematics/Statistics. His research interests include the theory and applications of Bayesian and empirical Bayes analyses with emphasis on the computational aspect of modeling. Chris P. Tsokos is a Distinguished Professor of Mathematics and Statistics. His research interests are in statistical analysis and modeling, operations research, reliability analysis-ordinary and Bayesian, time series analysis.

$$f(x;a,b,c)=\frac{c}{b}(x-a)^{c-1}e^{-\frac{(x-a)^c}{b}},\quad x\geq a;\ b,c>0 \tag{1}$$

where a, b and c are respectively the location, scale and shape parameters;

and

$$g(x;\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}, \tag{2}$$

where α and β are respectively the shape and scale parameters.

For these two failure models, consider the scale parameters b and β to behave as random variables that follow the lognormal distribution which is given by

$$\pi(\theta)=\frac{1}{\theta\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left[\frac{Ln(\theta)-\mu}{\sigma}\right]^2}, \tag{3}.$$

For each of the above underlying failure models, approximate Bayesian estimates will be obtained for the subject parameter and the reliability function with the squared error, the Higgins-Tsokos, the Harris, and a proposed logarithmic loss functions. The loss functions

along with a statement of their key characteristics are given below.

Square error loss function

The popular square error loss function places a small weight on estimates near the true value and proportionately more weight on extreme deviation from the true value of the parameter. Its popularity is due to its analytical tractability in Bayesian reliability modeling. The squared error loss is defined as follows:

$$L_{SE}(\hat{R}, R)=\left(\hat{R}-R\right)^2 \qquad (4)$$

Higgins-Tsokos loss function

The Higgins-Tsokos loss function places a heavy penalty on extreme over-or underestimation. That is, it places an exponential weight on extreme errors. The Higgins-Tsokos loss function is defined as follows:

$$L_{HT}(\hat{R}, R)= \frac{f_1 e^{f_2(\hat{R}-R)}+f_2 e^{-f_1(\hat{R}-R)}}{f_1+f_2}-1,$$

$$f_1, f_2 \succ 0.$$

Harris loss function

The Harris loss function is defined as follows:

$$L_H(\hat{R}, R)=\left|\frac{1}{1-\hat{R}}-\frac{1}{1-R}\right|^k, k \succ 0. \qquad (6)$$

To our knowledge, the properties of the Harris loss function have not been fully investigated. However it is based on the premises that if the system is 0.99 reliable then on the average it should fail one time in 100, whereas if the reliability is 0.999 it should fail one time in 1000. Thus, it is ten times as good.

Logarithmic loss function

The logarithmic loss function characterizes the strength of the loss logarithmically, and offers useful analytical tractability. This loss function is defined as:

$$L_{Ln}(\hat{R}, R)=\left|Ln\left(\frac{\hat{R}}{R}\right)\right|^l, l \succ 0. \qquad (7)$$

It places a small weight on estimates whose ratios to the true value are close to one, and proportionately more weight on estimates whose ratios to the true value are significantly different from one. $R(t)$ and $\hat{R}(t)$ represent respectively the true reliability function and its estimate.

Methodology

Considering the fact that the reliability of a system at a given time t is the probability that the system fails at a time greater or equal to t, the reliability function corresponding to the three-parameter Weibull failure model is given by

$$R(t) = e^{-\frac{(t-a)^c}{b}}, \qquad (8)$$

and for the gamma failure model

$$R(t)=1-\frac{\gamma(\alpha, \frac{t}{\beta})}{\Gamma(\alpha)}, t>0, \alpha, >0 \ . \qquad (9)$$

where $\gamma(l_1, l_2)$ denotes the incomplete gamma function. When α is an integer, equation (9) becomes

$$R(t) = \left(\sum_{i=0}^{\alpha-1}\frac{1}{i!}\left(\frac{t}{\beta}\right)^i\right)e^{-\frac{t}{\beta}},$$

and in particular when $\alpha = 1$

$$R(t) = e^{-\frac{t}{\beta}}, \quad t>0 \ .$$

Consider the situation where there are *m* independent random variables $X_1, X_2, ..., X_m$ with the same probability density function $dF(x|\theta)$, and each of them having *n* realizations, that is,

$X_1 : x_{11}, x_{21}, ..., x_{n1};$
$X_2 : x_{12}, x_{22}, ..., x_{n2};$ ……. ;
$X_m : x_{1m}, x_{2m}, ..., x_{nm}$

The minimum variance unbiased estimate, MVUE, $\hat{\theta}_j$ of the parameter θ_j is obtained from the *n* realizations $x_{1j}, x_{2j}, \ldots x_{nj}$, where j = *1,...m.*

Repeating this independent procedure *k* times, a sequence of MVUE is obtained for the θ_j's, that is, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$. Using the $\hat{\theta}_j$'s and their common probability density function, approximate Bayesian reliability estimates are obtained.

Let $L(x;\theta)$, $g(\theta)$, $\pi(\theta)$ and $h(\theta)$ represent respectively the likelihood function, a function of θ, a prior distribution of θ and a probability density function of θ called the importance function. Using the strong law of large numbers, [7], write

$$\int_{\Theta} g(\theta)L(x;\theta)\pi(\theta)d\theta = E^{h}\left[\frac{g(\theta)L(x;\theta)\pi(\theta)}{h(\theta)}\right]$$

$$= \lim_{m\to\infty} \sum_{i=1}^{m} \frac{g(\hat{\theta}_i)L(x;\hat{\theta}_i)\pi(\hat{\theta}_i)}{h(\hat{\theta}_i)} \qquad (10).$$

Note that E^h represents the expectation with respect to the probability density function *h*, and $g(\theta)$ is any function of θ which assures convergence of the integral; also, $h(\theta)$ mimics the posterior density function.

For the special case where $g(\theta) = 1$, equation (10) yields

$$\int_{\Theta} L(x;\theta)\pi(\theta)d\theta = \lim_{m\to\infty} \sum_{i=1}^{m} \frac{L(x;\hat{\theta}_i)\pi(\hat{\theta}_i)}{h(\hat{\theta}_i)} \qquad (11)$$

Equations (10) and (11) imply that the posterior expected value of $g(\theta)$ is given by

$$E(g(\theta)\,|\,x) = \frac{\int_{\Theta} g(\theta)L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta} L(x;\theta)\pi(\theta)d\theta}$$

$$= \lim_{m\to\infty} \frac{\sum_{i=1}^{m} \frac{g(\hat{\theta}_i)L(x;\hat{\theta}_i)\pi(\hat{\theta}_i)}{h(\hat{\theta}_i)}}{\sum_{i=1}^{m} \frac{L(x;\hat{\theta}_i)\pi(\hat{\theta}_i)}{h(\hat{\theta}_i)}} \qquad (12)$$

This approach is used to obtain approximate Bayesian estimates of $g(\theta)$, for the different loss functions under study. Approximate Bayesian estimates of the parameter θ and the reliability are then obtained by replacing $g(\theta)$ by θ and *R(t)* respectively in the derived expressions corresponding to the approximate Bayesian estimates of $g(\theta)$.

The Bayesian estimates used to obtain approximate Bayesian estimates of the function $g(\theta)$ are the following when the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are used:

$$\hat{g}(\theta)_{B(SE)} = \frac{\int_{\Theta} g(\theta)L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta} L(x;\theta)\pi(\theta)d\theta}$$

$$\hat{g}(\theta)_{B(HT)} = \frac{1}{f_1 + f_2} Ln\left(\frac{\int_{\Theta} e^{f_1 g(\theta)} L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta} e^{-f_2 g(\theta)} L(x;\theta)\pi(\theta)d\theta}\right),$$

$f_1, f_2 \succ 0.$

$$\hat{g}(\theta)_{B(H)} = \frac{\int_{\Theta} \frac{g(\theta)}{1-g(\theta)} L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta} \frac{1}{1-g(\theta)} L(x;\theta)\pi(\theta)d\theta}$$

$$\hat{g(\theta)}_{B(Ln)} = e^{\frac{\int_{\Theta} Ln(g(\theta))L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta} L(x;\theta)\pi(\theta)d\theta}} \quad (13).$$

Using equation (12) and the above Bayesian decision rules, approximate Bayesian estimates of $g(\theta)$ corresponding respectively to the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are respectively given by the following expressions when *m* replicates are considered.

$$\hat{g(\theta)}_{E(SE)} = \frac{\sum_{i=1}^{m} \frac{\hat{g(\theta i)}L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}{\sum_{i=1}^{m} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}} \quad (14)$$

$$\hat{g(\theta)}_{E(HT)} = \frac{1}{f_1 + f_2} Ln\left(\frac{\sum_{i=1}^{m} e^{f_1 \hat{g(\theta i)}} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}{\sum_{i=1}^{m} e^{-f_2 \hat{g(\theta i)}} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}} \right) \quad (15)$$

$f_1, f_2 \succ 0.$

$$\hat{g(\theta)}_{E(H)} = \frac{\sum_{i=1}^{m} \frac{\hat{g(\theta i)}}{1-\hat{g(\theta i)}} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}{\sum_{i=1}^{m} \frac{1}{1-\hat{g(\theta i)}} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}, \hat{g(\theta i)} \neq 1 \quad , \quad (16)$$

and

$$\hat{g(\theta)}_{E(Ln)} = e^{\frac{\sum_{i=1}^{m} \frac{Ln(g(\hat{\theta i}))L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}{\sum_{i=1}^{m} \frac{L(x;\hat{\theta i})\pi(\hat{\theta i})}{\hat{h(\theta i)}}}} \quad . \quad (17)$$

First, use the above general functional forms of the Bayesian estimates of $g(\theta)$ to obtain approximate Bayesian estimates of the random parameter inherent in the underlying failure model. Furthermore, these estimates are used to obtain approximate Bayesian reliability estimates. Second, use the above functional forms to directly obtain approximate Bayesian estimates of the reliability function.

Three-parameter Weibull underlying failure model

In this case the parameter θ, discussed above, will correspond to the scale parameter *b*. The location and shape parameters *a* and *c* are considered fixed. The likelihood function corresponding to *n* independent random variables following the three-parameter Weibull failure model is given by

$$L_1(\underset{\sim}{x}, a, c; b) = e^{-\frac{1}{b}S_n - nLn(b)} \bullet e^{(c-1)\sum_{i=1}^{n} Ln(x_i - a) + nLn(c)} \quad (18)$$

where $S_n = \sum_{i=1}^{n} (x_i - a)^c$.

Furthermore, it can be shown that S_n is a sufficient statistic for the parameter *b*, and a minimum variance unbiased estimator of *b* is given by

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - a)^c}{n} \quad .$$

The probability density function of $Y=(X-a)^c$, where X follows the Weibull probability density function, is

$$p(y\mid b)=\frac{c}{b}\left(y^{\frac{1}{c}}\right)^{c-1} e^{-\frac{1}{b}y\left(\frac{1}{c}y^{\frac{1}{c}-1}\right)}$$

$$= \frac{1}{b}e^{-\frac{1}{b}y}, y>0, b>0 . \qquad (19)$$

The moment generating function of Y is given by

$$E(e^{\mu y})=\frac{1}{b}\int_0^{\infty} e^{-y(\frac{1}{b}-\mu)}dy$$

$$= (1-\mu b)^{-1} \qquad (20)$$

Using equation (20) and the fact that the X_i's are independent, the moment generating function of the minimum variance unbiased estimator of the parameter b is

$$E(e^{\mu \hat{b}})=\prod_{i=1}^{n} E\left(e^{\frac{\mu}{n}(x_i-a)^c}\right).$$

$$= \left(1-\mu\frac{b}{n}\right)^{-n} \qquad (21)$$

Equation (21) corresponds to the moment generating function of the gamma distribution $G(n,\frac{b}{n})$. Thus, the conditional probability density function of the MVUE of b is given by

$$h_1(\hat{b},a,c\mid b)=\frac{n^n}{\Gamma(n)b^n}\left(\hat{b}\right)^{n-1} e^{-\frac{n}{b}\hat{b}}, \hat{b}>0, b>0 . \qquad (22)$$

Approximate Bayesian estimates for the scale parameter b and the reliability function $R(t)$ are obtained, with the use of equations (18) and (22), by replacing respectively $g(b)$ by b and $R(t)$ in equations (14), (15), (16) and (17). The $\hat{b}_i$'s that are minimum variance unbiased estimates of the scale parameter b will play the role of the $\hat{\theta}_i$'s.

Considering the lognormal prior, equations (14), (15), (16) and (17) yield the following approximate Bayesian estimates of the scale parameter b corresponding respectively to the squared error, the Higgins-Tsokos, the Harris and our proposed lognormal loss functions, after replacing b_i by $\hat{b}_i$ in the expression of $h_1(\hat{b}_i)$:

$$\hat{b}_{E(SE)} = \frac{\sum_{j=1}^{m} \hat{b}_j e^{-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}, \qquad (23)$$

$$\hat{b}_{E(HT)} = \frac{1}{f_1+f_2}Ln\left(\frac{\sum_{j=1}^{m} e^{f_1\hat{b}_j-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-f_2\hat{b}_j-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}\right), \quad f_1, f_2 \succ 0 \qquad (24)$$

$$\hat{b}_{E(H)} = \frac{\sum_{j=1}^{m} \frac{\hat{b}_j}{1-\hat{b}_j} e^{-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} \frac{1}{1-\hat{b}_j} e^{-\frac{S_n}{\hat{b}_j}-nLn(\hat{b}_j)+(c-1)\sum_{i=1}^{n}Ln(x_i-a)-\frac{1}{2}\left(\frac{Ln(\hat{b}_j)-\mu}{\sigma}\right)^2}}, \quad \hat{b}_{j\neq 1} \qquad (25)$$

and

$$\overset{\Lambda}{b}_{E(Ln)} = e^{\frac{\sum_{j=1}^{m} Ln(\overset{\Lambda}{b}_j) e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}} \tag{26}$$

The approximate Bayesian estimates of the reliability corresponding to the first method are therefore given by

$$\overset{\approx}{R}_{Eb}(t, a, c | \overset{\Lambda}{b}_E) = e^{-\frac{1}{\overset{\Lambda}{b}_E}(t-a)^c} \qquad t > a \ , \tag{27}$$

where $\overset{\Lambda}{b}_E$ stands respectively for the above approximate Bayesian estimates of the scale parameter *b*.

Approximate Bayesian reliability estimates corresponding to the second method are also derived by replacing $g(\theta)$ by *R(t)* in equations (14), (15), (16) and (17). The obtained estimates corresponding respectively to the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are respectively given by the following expressions, after replacing b_i by $\overset{\Lambda}{b}_i$ in the expression of $h_1(\overset{\Lambda}{b}_i)$:

$$\overset{\Lambda}{R}_{E(SE)}(t) = \frac{\sum_{j=1}^{m} e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j} - \frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}, \tag{28}$$

$$\overset{\Lambda}{R}_{E(HT)}(t) = \frac{1}{f_1 + f_2} Ln \left(\frac{\sum_{j=1}^{m} e^{f_1 e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j}} - \frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-f_2 e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j}} - \frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}} \right),$$

$$f_1, f_2 \succ 0 \ , \tag{29}$$

$$\overset{\Lambda}{R}_{E(H)}(t) = \frac{\sum_{j=1}^{m} \frac{e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j}}}{1 - e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j}}} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} \frac{1}{1 - e^{-\frac{(t-a)^c}{\overset{\Lambda}{b}_j}}} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}, \tag{30}$$

and

$$\overset{\Lambda}{R}_{E(Ln)}(t) = e^{\frac{\sum_{j=1}^{m} -\frac{(t-a)^c}{\overset{\Lambda}{b}_j} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n}{\overset{\Lambda}{b}_j} - nLn(\overset{\Lambda}{b}_j) + (c-1)\sum_{i=1}^{n} Ln(x_i - a) - \frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{b}_j) - \mu}{\sigma}\right)^2}}} \ . \tag{31}$$

Gamma underlying failure model

The likelihood function corresponding to *n* independent random variables following the two-parameter gamma underlying failure model can be written under the following form:

$$L_2(x, \alpha; \beta) = e^{-\frac{1}{\beta} S_n' - n\alpha Ln(\beta)} e^{(\alpha - 1)\sum_{i=1}^{n} Ln(x_i) - nLn(\Gamma(\alpha))} \ , \tag{32}$$

where $S_n' = \sum_{i=1}^{n} x_i$.

Note that $S_n^{'}$ is a sufficient statistic for the scale parameter β. Furthermore,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i}{n\alpha}$$

is a minimum variance unbiased estimator of β, and its moment generating function is given by

$$E(e^{\mu\hat{\beta}}) = \prod_{i=1}^{n} E(e^{\mu\frac{x_i}{n\alpha}}) = \left(1-\mu\frac{\beta}{n\alpha}\right)^{-n\alpha} \tag{33}$$

which is the moment generating function of the gamma distribution $G(n\alpha, \frac{\beta}{n\alpha})$. Thus, the conditional density function of the MVUE of β is given by

$$h_2(\hat{\beta},\alpha \mid \beta) = \frac{(n\alpha)^{n\alpha}}{\Gamma(n\alpha)\beta^{n\alpha}}\left(\hat{\beta}\right)^{n\alpha-1} e^{-\frac{n\alpha}{\beta}\hat{\beta}}, \hat{\beta} > 0 \tag{34}$$

Approximate Bayesian estimates for the scale parameter β and the reliability function *R(t)* are obtained, with the use of equations (32) and (34) by replacing respectively $g(\theta)$ by β and $R(t)$ in equations (14), (15), (16) and (17). The $\hat{\beta}_i$'s that are the minimum variance unbiased estimates of the scale parameter β will play the role of the $\hat{\theta}_i$'s.

Considering the lognormal prior, equation (14), (15), (16) and (17) yield the following approximate Bayesian estimates of the scale parameter β corresponding respectively to the squared error, the Higgins-Tsokos, the Harris and the proposed lognormal loss functions, after replacing β_i by $\hat{\beta}_i$ in the expression of $h_2(\hat{\beta}_i)$:

$$\hat{\beta}_{E(SE)} = \frac{\sum_{j=1}^{m} \hat{\beta}_j e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}, \tag{35}$$

$$\hat{\beta}_{E(HT)} = \frac{1}{f_1 + f_2} Ln \left(\frac{\sum_{j=1}^{m} e^{f_1\hat{\beta}_j - \frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-f_2\hat{\beta}_j - \frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}} \right), \quad f_1, f_2 \succ 0 \tag{36}$$

$$\hat{\beta}_{E(H)} = \frac{\sum_{j=1}^{m} \frac{\hat{\beta}_j}{1-\hat{\beta}_j} e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} \frac{1}{1-\hat{\beta}_j} e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}, \quad \hat{\beta}_j \neq 1 \tag{37}$$

and

$$\hat{\beta}_{E(Ln)} = e^{\frac{\sum_{j=1}^{m} Ln(\hat{\beta}_j) e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n^{'}}{\hat{\beta}_j} - n\alpha Ln(\hat{\beta}_j) + (\alpha-1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{2}\left(\frac{Ln(\hat{\beta}_j)-\mu}{\sigma}\right)^2}}} \tag{38}$$

Approximate Bayesian estimates of the reliability corresponding to the first method are therefore given by

$$\overset{\approx}{R}_{E\beta}(t,\alpha \mid \overset{\Lambda}{\beta}_E) = 1-\frac{\gamma\left(\alpha,\frac{t}{\overset{\Lambda}{\beta}_E}\right)}{\Gamma(\alpha)} \quad , \quad t \succ 0 \tag{39}$$

where $\overset{\Lambda}{\beta}_E$ is the approximate Bayesian estimate of the scale parameter β.

The approximate Bayesian reliability estimates corresponding to the second method are obtained by replacing $g(\theta)$ by $R(t)$ in equations(14), (15), (16) and (17). The obtained estimates corresponding respectively to the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are given by the following expressions, after replacing β_i by $\overset{\Lambda}{\beta}_i$ in the expression of $h_2(\overset{\Lambda}{\beta}_i)$:

$$\overset{\Lambda}{R}_{E(SE)}(t) = \frac{\sum_{j=1}^{m}\left(1-\frac{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}})}{\Gamma(\alpha)}\right) e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}} \tag{40}$$

$$\overset{\Lambda}{R}_{E(HT)}(t) = \frac{1}{f_1+f_2} Ln \left(\frac{\sum_{j=1}^{m} e^{f_1\left(1-\frac{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})}{\Gamma(\alpha)}\right)-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m} e^{-f_2\left(1-\frac{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})}{\Gamma(\alpha)}\right)-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}} \right),$$

$$f_1,\ f_2 \succ 0\ , \tag{41}$$

$$\overset{\Lambda}{R}_{E(H)}(t) = \frac{\sum_{j=1}^{m}\frac{\left(\Gamma(\alpha)-\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})\right)}{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})} e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}{\sum_{j=1}^{m}\frac{\Gamma(\alpha)}{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})} e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}, \tag{42}$$

and

$$\overset{\Lambda}{R}_{E(Ln)}(t) = \frac{\sum_{j=1}^{m} Ln\left(1-\frac{\gamma(\alpha,\frac{t}{\overset{\Lambda}{\beta}_j})}{\Gamma(\alpha)}\right) e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}{e^{\sum_{j=1}^{m} e^{-\frac{S_n'}{\overset{\Lambda}{\beta}_j}-n\alpha Ln(\overset{\Lambda}{\beta}_j)+(\alpha-1)\sum_{i=1}^{n}Ln(x_i)-\frac{1}{2}\left(\frac{Ln(\overset{\Lambda}{\beta}_j)-\mu}{\sigma}\right)^2}}} \tag{43}$$

Relative Efficiency with Respect to the Squared Error Loss

To compare our results, the criterion of integrated mean square error, IMSE, of the approximate Bayesian reliability estimate $\tilde{R}_E(t)$ is used. That is,

$$IMSE(\tilde{R}_E(t)) = \int_0^\infty \left(\tilde{R}_E(t) - R(t)\right)^2 dt \tag{44}$$

Define the relative efficiency as the ratio of the IMSE of the approximate Bayesian reliability estimates using a challenging loss function to that of the popular squared error loss. The relative efficiencies of the Higgins-Tsokos, the Harris and the proposed logarithmic loss are respectively defined as follows:

$$Eff(HT) = \frac{IMSE(\tilde{R}_{E(HT)}(t))}{IMSE(\tilde{R}_{E(SE)}(t))}$$

$$= \frac{\int_0^\infty \left(\tilde{R}_{E(HT)}(t) - R(t)\right)^2 dt}{\int_0^\infty \left(\tilde{R}_{E(SE)}(t) - R(t)\right)^2 dt},$$

$$Eff(H) = \frac{IMSE(\tilde{R}_{E(H)}(t))}{IMSE(\tilde{R}_{E(SE)}(t))}$$

$$= \frac{\int_0^\infty \left(\tilde{R}_{E(H)}(t) - R(t)\right)^2 dt}{\int_0^\infty \left(\tilde{R}_{E(SE)}(t) - R(t)\right)^2 dt}$$

and

$$Eff(Ln) = \frac{IMSE(\tilde{R}_{E(Ln)}(t))}{IMSE(\tilde{R}_{E(SE)}(t))}$$

$$= \frac{\int_0^\infty \left(\tilde{R}_{E(Ln)}(t) - R(t)\right)^2 dt}{\int_0^\infty \left(\tilde{R}_{E(SE)}(t) - R(t)\right)^2 dt}.$$

If the relative efficiency is smaller than one, the Bayesian estimate corresponding to the squared error loss is less efficient. The squared error will be more efficient if the relative efficiency is greater than one. If the relative efficiency is approximately equal to one, the Bayesian reliability estimates are equally efficient.

Numerical Simulations

In the numerical simulations, Bayesian and approximate Bayesian estimates of the scale parameter β for the gamma failure model and the lognormal prior will be compared, when the squared error loss is used and the shape parameter α is considered fixed. Second, the new approach will be implemented, and approximate Bayesian reliability estimates will be obtained for the three-parameter Weibull and the gamma failure model under the squared error, the Higgins-Tsokos (with $f_1 = 1, f_2 = 1$), the Harris, and the logarithmic loss functions, respectively.

Comparison between Bayesian estimates and approximate Bayesian estimates of the scale parameter β

Using the square error loss function, the gamma underlying failure model and the lognormal prior, Table 1 gives estimates of the scale parameter β when the shape parameter α is fixed and equal to one.

Table 1.

Lognormal prior	True value of β	Bayesian estimate of β	Approximate Bayesian estimates of β	Number of replicates m
$\mu = 1, \sigma = 0.5$	1	1.1688	0.9795	1
			0.9883	2
			1.0796	3
			1.0625	4
			1.0385	5
			1.0899	6
			1.0779	7
$\mu = 4, \sigma = 9$	1	1.0561	0.9795	1
			0.9880	2
			1.0351	3
			0.9943	4
			0.9665	5
			0.9945	6
			1.0017	7
$\mu = 3, \sigma = 0.8$	2	2.2808	1.9591	1
			1.9766	2
			2.1555	3
			2.1162	4
			2.0658	5
			2.1679	6
			2.1467	7
$\mu = 8, \sigma = 12$	2	2.0376	1.9591	1
			1.9761	2
			2.0704	3
			1.9886	4
			1.9331	5
			1.9892	6
			2.0034	7

The above results show that the obtained approximate Bayesian estimates of the parameter β are as good if not better than the corresponding Bayesian estimates, because they are in general closer to the true state of nature.

Approximate Bayesian Reliability Estimates of the Three-parameter Weibull and the Gamma Failure Models for the different Loss Functions

Using Monte Carlo simulation, information has been respectively generated from the three-parameter Weibull *W(a=1, b=1, c=2) and* the two-parameter gamma $G(\alpha = 1, \beta = 1)$. For each of the above underlying failure models, three different samples are generated of thirty failure times, and three minimum variance unbiased estimates of the scale parameter are obtained.

Three-parameter Weibull W(a=1,b=1,c=2)

A typical sample of thirty failure times that are randomly generated from *W(a=1,b=1,c=2)* is given below:

1.9772260	2.6416950	2.1241180
1.5575370	2.7714080	1.7158910
1.3109790	2.2144780	2.2674890
2.2136030	1.3422820	1.4691720
1.3017910	1.7534080	1.9712720
1.6897900	1.9609470	2.9533880
1.5448060	1.4516050	1.1704900
1.9409150	2.5030900	1.4788690
2.1088060	1.7306430	1.8829980
1.8939380	1.8181710	2.7016010

The obtained minimum variance unbiased estimates of the scale parameter *b* are given below

$$\hat{b}_1 = 1.1408084120$$
$$\hat{b}_2 = 1.0091278197$$
$$\hat{b}_3 = 0.9991267092$$

These minimum variance unbiased estimates will be used along with likelihood function and the lognormal prior $f(b; \mu = 0.34, \sigma = 0.115)$ to obtain approximate Bayesian reliability estimates.

Let $\tilde{R}_{Eb(SE)}(t), \hat{R}_{Eb(SE)}(t), \tilde{R}_{Eb(HT)}(t), \hat{R}_{Eb(HT)}(t), \tilde{R}_{Eb(H)}(t), \hat{R}_{Eb(H)}(t), \tilde{R}_{Eb(Ln)}(t)$ and $\hat{R}_{Eb(Ln)}(t)$,

represent, respectively, the approximate Bayesian reliability estimates obtained with the approximate Bayesian reliability estimates of the scale parameter *b*, and the ones obtained by direct computation, when the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are used. These estimates are given below in Table 2. Table 3 gives the approximate Bayesian reliability estimates obtained directly using equations (28), (29), (30) and (31).

Gamma failure model $G(\alpha = 1, \beta = 1)$

A typical sample of thirty failure times that are randomly generated from $G(\alpha = 1, \beta = 1)$ is given below.

0.95497	0.09670	0.09107
2.69516	1.47495	0.56762
1.26364	1.60653	0.94337
0.54999	0.64000	0.62536
1.44922	0.78403	1.08172
0.31084	1.47283	0.47580
3.13788	0.11715	0.92341
0.51249	0.22012	3.81572
0.57911	0.50421	0.14532
0.77497	1.07792	1.08156

The obtained minimum variance unbiased estimates of the scale parameter β are given below.

$$\hat{\beta}_1 = 1.009127916$$
$$\hat{\beta}_2 = 1.140808468$$
$$\hat{\beta}_3 = 0.9991268436$$

Table 2.

	$R(t)$	$\tilde{\tilde{R}}_{Eb(SE)}(t)$	$\tilde{\tilde{R}}_{Eb(HT)}(t)$	$\tilde{\tilde{R}}_{Eb(H)}(t)$	$\tilde{\tilde{R}}_{Eb(Ln)}(t)$
Approximation	$e^{-(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$	$e^{-\frac{1}{0.9758}(t-1)^2}$	$e^{-\frac{1}{1.1242}(t-1)^2}$
IMSE	0	2.381010^{-4}	3.676410^{-3}	1.482010^{-4}	3.630110^{-3}
Relative efficiency with respect to $\tilde{\tilde{R}}_{Eb(SE)}(t)$	0	1.0	15.44	0.62	15.25

The above approximate Bayesian estimates yield good estimates of the true reliability function.

Table 3.

Time t	$\hat{R}(t)$	$\hat{R}_{Eb(SE)}(t)$	$\hat{R}_{Eb(HT)}(t)$	$\hat{R}_{Eb(H)}(t)$	$\hat{R}_{Eb(Ln)}(t)$
1.00001	1.0000	1.0000	1.0000	1.0000	1.0000
1.25	0.9394	0.9459	0.9459	0.9459	0.9459
1.50	0.7788	0.8005	0.8005	0.8008	0.8005
1.75	0.5698	0.6062	0.6062	0.6066	0.6061
2.00	0.3679	0.4108	0.4108	0.4112	0.4105
2.25	0.2096	0.2492	0.2492	0.2495	0.2488
2.50	0.1054	0.1354	0.1354	0.1355	0.1349
2.75	0.0468	0.0659	0.0659	0.0659	0.0655
3.00	0.0183	0.0287	0.0287	0.0287	0.0284
3.25	0.0063	0.0112	0.0112	0.0112	0.0110
3.50	0.0019	0.0039	0.0039	0.0039	0.0038
3.75	0.0005	0.0012	0.0012	0.0012	0.0012
4.00	0.0001	0.0003	0.0003	0.0003	0.0003

These minimum variance unbiased estimates will be used along with the likelihood function and the lognormal prior $f(x;\mu=0.0137,\sigma=0.1054)$ to obtain approximate Bayesian reliability estimates.

Let $\tilde{R}_{E\beta(SE)}(t)$, $\hat{R}_{E\beta(SE)}(t)$, $\tilde{R}_{E\beta(HT)}(t)$, $\hat{R}_{E\beta(HT)}(t)$, $\tilde{R}_{E\beta(H)}(t)$, $\hat{R}_{E\beta(H)}(t)$, $\tilde{R}_{E\beta(Ln)}(t)$ and $\hat{R}_{E\beta(Ln)}(t)$ represent respectively the approximate Bayesian reliability estimates obtained with the approximate Bayesian estimate of β, and the ones obtained by direct computation, when the squared error, the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are used. These estimates are given in Table 5 and Table 6.

For computational convenience, the results presented in Table 3 are used to obtain approximate estimates of the analytical forms of the various approximate Bayesian reliability expressions under study. The results are given in Table 4. Table 6 gives the approximate Bayesian reliability estimates obtained directly by using equations (40), (41), (42) and (43).

For computational convenience, the results presented in Table 6 are used to obtain approximate estimates of the analytical forms of the various approximate Bayesian reliability expressions under study. The results are given in Table 7.

Table 4.

	$R(t)$	$\hat{R}_{Eb(SE)}(t)$	$\hat{R}_{Eb(HT)}(t)$	$\hat{R}_{Eb(H)}(t)$	$\hat{R}_{Eb(Ln)}(t)$
Approximation	$e^{-(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$	$e^{-\frac{1}{1.1251}(t-1)^2}$
IMSE	0	2.381310^{-3}	2.381310^{-3}	2.381310^{-3}	2.381310^{-3}
Relative efficiency with respect to $\hat{R}_{Eb(SE)}(t)$	0	1	1	1	1

Table 5.

	$R(t)$	$\tilde{R}_{E\beta(SE)}(t)$	$\tilde{R}_{E\beta(HT)}(t)$	$\tilde{R}_{E\beta(H)}(t)$	$\tilde{R}_{E\beta(Ln)}(t)$
Approximation	e^{-t}	$e^{-\frac{t}{1.0311}}$	$e^{-\frac{t}{1.1250}}$	$e^{-\frac{t}{0.9758}}$	$e^{-\frac{t}{1.1242}}$
IMSE	0.0	$2.381008 10^{-4}$	$3.676471 10^{-3}$	$1.482034 10^{-4}$	$3.630931 10^{-3}$
Relative efficiency with respect to $\tilde{R}_{E\beta(SE)}(t)$	0.0	1.0	15.44	0.62	15.25

Table 6.

Time t	$\hat{R}(t)$	$\hat{R}_{E\beta(SE)}(t)$	$\hat{R}_{E\beta(HT)}(t)$	$\hat{R}_{E\beta(H)}(t)$	$\hat{R}_{E\beta(Ln)}$
10^{-100}	1.0000	1.0000	1.0000	1.0000	1.0000
1.00	0.3679	0.3786	0.4108	0.4112	0.4105
2.00	0.1353	0.1437	0.1690	0.1692	0.1685
3.00	0.9498	0.0547	0.0696	0.0697	0.0692
4.00	0.0183	0.0209	0.0287	0.0287	0.0284
5.00	0.0067	0.0080	0.0118	0.0118	0.0117
6.00	0.0025	0.0031	0.0049	0.0049	0.0048
7.00	0.0009	0.0012	0.0020	0.0020	0.0020
8.00	0.0003	0.0005	0.0008	0.0008	0.0008
9.00	0.0001	0.0002	0.0003	0.0003	0.0003
10.00	0.0000	0.0001	0.0001	0.0001	0.0001

Table 7.

	$\hat{R}_{E\beta(SE)}(t)$	$\hat{R}_{E\beta(HT)}(t)$	$\hat{R}_{E\beta(H)}(t)$	$\hat{R}_{E\beta(Ln)}(t)$
Approximation	$e^{-\frac{t}{1.0311}}$	$e^{-\frac{t}{1.1250}}$	$e^{-\frac{t}{1.1250}}$	$e^{-\frac{t}{1.1242}}$
IMSE	2.38100810^{-4}	3.67647110^{-3}	3.67647110^{-3}	3.63093110^{-3}
Relative efficiency with respect to $\hat{R}_{E\beta(SE)}(t)$	1.0	15.44	15.44	15.25

The above approximate Bayesian estimates yield good estimates of the true reliability function.

Conclusion

Using the concept of Monte Carlo Integration, approximate Bayesian estimates of the scale parameter *b* were analytically obtained for the three-parameter Weibull failure model under different loss functions. Using these estimates, approximate Bayesian estimates of the reliability function may be obtained. Furthermore, the concept of Monte Carlo Integration may be used to directly approximate estimates of the Bayesian reliability function.

Second, similar results were obtained for the gamma failure model. Finally, numerical simulations of the analytical formulations indicate:

(1) Approximate Bayesian reliability estimates are in general good estimates of the true reliability function.

(2) When the number of replicates *m* increases, the approximate Bayesian reliability estimates obtained directly converge for each loss function to their corresponding Bayesian reliability estimates.

(3) Approximate Bayesian reliability estimates corresponding to the squared loss function do not always yield the best approximations to the true reliability function. In fact the Higgins-Tsokos, the Harris and the proposed logarithmic loss functions are sometimes equally efficient if not better.

References

Bennett, G. K. (1970). Smooth empirical bayes estimation with application to the Weibull distribution. *NASA Technical Memorandum*, X-58048.

Hogg, R. V., & Craig, A. T. (1965). *Introduction to mathematical statistics*. New York: The Macmillan Co.

Lemon, G. H., & Krutchkoff, R. G. (1969). An empirical Bayes smoothing technique. *Biometrika*, *56*, 361-365.

Maritz, J. S. (1967). Smooth empirical bayes estimation for one-paramater discrete distributions. *Biometrika*, *54*, 417-429.

Tate, R. F. (1959). Unbiased estimation: functions of location and scale parameters. *Annals of Math and Statistics*, *30*, 341-366.

Robbins, H. (1955). The empirical bayes approach to statistical decision problems. *Annals of Math and Statistics*, *35*, 1-20.

Berger J. O. (1985) Statistical decision theory and Bayesian analysis, (2nd Ed.). New York: Springer.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 187-213

1538 – 9472/05/$95.00

Right-tailed Testing of Variance for Non-Normal Distributions

Michael C. Long
Florida State Department of Health

Ping Sa
Mathematics and Statistics
University of North Florida

A new test of variance for non-normal distribution with fewer restrictions than the current tests is proposed. Simulation study shows that the new test controls the Type I error rate well, and has power performance comparable to the competitors. In addition, it can be used without restrictions.

Key words: Edgeworth expansion, Type I error rate, power performance

Introduction

Testing the variance is crucial for many real world applications. Frequently, companies are interested in controlling the variation of their products and services because a large variation in a product or service indicates poor quality. Therefore, a desired maximum variance is frequently established for some measurable characteristic of the products of a company.

In the past, most of the research in statistics concentrated on the mean, and the variance has drawn less attention. This article is about testing the hypothesis that the variance is equal to a hypothesized value σ_o^2 versus the alternative that the variance is larger than the hypothesized value. This statistical test will be referred to as a right-tailed test in further discussion.

The chi-square test is the most commonly used procedure to test a single variance of a population. Once a random sample of size n is taken, the individual values X_i, the sample mean $\overline{X}$, the sample variance S^2, and specified (σ_o^2) are used to compute the chi-squared test statistic $\chi^2 = (n-1)S^2/\sigma_0^2$, which is distributed $\chi^2_{(n-1)}$ under H_0. The χ^2 statistic is used for hypothesis tests concerning σ^2 when a normal population is assumed. It is well known that the chi-square test statistic is not robust against departures from normality such as when skewness and kurtosis are present. This can lead to rejecting H_0 much more frequently than indicated by the nominal alpha level, where alpha is the probability of rejecting H_0 when H_0 is true.

Practical alternatives to the χ^2 test are needed for testing the variance of non-normal distributions. There are nonparametric methods such as bootstrap and jackknife (see Efron & Tibshirani, 1993). The bootstrap requires extensive computer calculations and some programming ability by the practitioner making the method infeasible for some people. Although the jackknife method is easier to implement, it is a linear approximation to the bootstrap method and can give poor results when the statistic estimate is nonlinear.

Another alternative is presented in Kendall (1994) and Lee and Sa (1998). The robust chi-square statistic χ_r^2 which has the form $(n-1)\hat{d}S^2/\sigma^2$ and is chi-square distributed with $(n-1)\hat{d}$ degrees of freedom,

Michael C. Long (MA, University of North Florida) is a Research Associate and Statistician in the Department of Health, State of Florida. Email: longstats@cs.com. Ping Sa (Ph. D. University of South Carolina) is a Professor of Mathematics and Statistics at the University of North Florida. Her recent scholarly activities have involved research in multiple comparisons and quality control. Email: psa@unf.edu.

where $\hat{d} = \left(1 + \frac{\hat{\eta}}{2}\right)^{-1}$ and $\hat{\eta}$ is the sample kurtosis coefficient. The critical value for test rejection is $\chi^2_{\nu,\alpha}$ where ν is the smallest integer, which is greater than or equal to (n-1) $\hat{d}$. Because $\hat{d}$ is a function of the sample kurtosis coefficient $\hat{\eta}$ alone, this could create performance problems for χ^2_r test with skewed distributions.

Lee and Sa (1996) derived a new method for a right-tailed variance test of symmetric heavy-tailed distributions using an Edgeworth expansion (see Bickel & Doksum, 1977), and an inversion type of Edgeworth expansion provided by Hall (1983),

$$P\left((\hat{\theta} - \theta)/\sigma(\hat{\theta}) \le x + \beta_1(x^2 - 1)/6\right) = \Phi(x) + o(1/\sqrt{n}), \quad (1)$$

where $\hat{\theta}$ is any statistic, and θ, $\sigma(\hat{\theta})$ and β_1 are the mean, standard deviation and coefficient of skewness of $\hat{\theta}$, respectively. $\Phi(x)$ is the standard normal distribution function.

They considered the variable S^2/σ^2, and the variable admitted the inversion of the Edgeworth expansion above as follows:

$$P\left(\frac{\frac{S^2}{\sigma^2} - 1}{\sqrt{\frac{K_4}{n\sigma^4} + \frac{2}{n-1}}} \le x + \beta_1(x^2 - 1)/6\right) = \Phi(x) + o(1/\sqrt{n}), \quad (2)$$

where $K_4 = E(X-\mu)^4 - 3(E(X-\mu)^2)^2$ and $\beta_1 = \frac{E(S^2 - \sigma^2)^3}{(E(S^2 - \sigma^2)^2)^{3/2}}$, the coefficient of skewness of S^2, provided all the referred moments exist. The population coefficient of skewness equals $K_3/\sqrt{(\sigma^2)^3} = 0$ under symmetric and heavy-tailed assumptions, and the population coefficient of kurtosis equals $K_4/\sigma^4 > 0$, where K_i is the i^{th} cumulant (see Kendall & Stuart, 1969). This yielded a decision rule:

Reject $H_0 : \sigma^2 = \sigma_0^2$ versus $H_a : \sigma^2 > \sigma_0^2$ if

$$Z > z_\alpha + \hat{\beta}_1(z_\alpha^2 - 1)/6, \quad (3)$$

where z_α is the upper α percentage point of the standard normal distribution,

$$Z = \frac{\frac{S^2}{\sigma_0^2} - 1}{\sqrt{\frac{k_4}{n\sigma_0^4} + \frac{2}{n-1}}}, \quad \text{and}$$

$$\hat{\beta}_1 = \frac{\frac{1}{n^2}\left[\frac{3n}{2}k_4 S^2 - \frac{1}{2}k_6 + \frac{8n^2}{(n-1)^2}(S^2)^3\right]}{\sqrt{\left(\frac{k_4}{n} + \frac{2(S^2)^2}{n-1}\right)^3}},$$

where k_i is the i^{th} sample cumulant.

They approximated their decision rule even further using a Taylor series expansion of $f^{-1}(Z)$ at $-a$ where $a = \hat{\beta}_1/6$. The new test became:

Reject H_0 if

$$Z_1 = Z - a(Z^2 - 1) + 2a^2(Z^3 - Z) > z_\alpha. \quad (4)$$

After a simulation study, their study found their test provided a "controlled Type I error rate as well as good power performance when sample size is moderate or large" (p. 51).

Lee and Sa (1998) performed another study on a right-tailed test of variance for skewed distributions. A method similar to the previously proposed study was employed with the primary difference being in the estimated coefficient of skewness, $\hat{\beta}_1$. The population coefficient of skewness, $K_3/\sqrt{(\sigma^2)^3}$, was assumed zero in the heavy-tailed distribution study and estimated for the skewed distribution study. Their study performed a preliminary

simulation study for the best form of Z and found

$$Z = \frac{\frac{S^2}{\sigma_0^2} - 1}{\sqrt{\frac{k_4}{nS^2{\sigma_0}^2} + \frac{2}{n-1}}}$$

to be the Z with controllable Type I error rates as well as good power performance.

Hence, the motivation for this study is to develop an improved method for right-tailed tests of variance for non-normal distributions. A test is desired which works for both skewed and heavy-tailed distributions and also has fewer restrictions from assumptions. This test should work well for multiple sample sizes and significance levels. The test proposed uses a general Edgeworth expansion to adjust for the non-normality of the distribution and considers the variable S^2 that admits an inversion of the general Edgeworth expansion.

A detailed explanation of the new method is provided in the next section. In the "Simulation Study" Section, the simulation study is introduced for determining whether the previously proposed tests or the new test has the best true level of significance or power. The results of the simulation are discussed in the section of Simulation Results. Conclusions of the study are rendered at the end.

Methodology

Let $\hat{\theta}$ be an estimate of an unknown quantity θ_o. If $\sqrt{n}\left(\hat{\theta}-\theta_o\right)$ is asymptotically normally distributed with zero mean and variance σ^2, the distribution function of $\sqrt{n}\left(\hat{\theta}-\theta_o\right)$ may be expanded as a power series in $\sqrt{n}$ (see Hall, 1983),

$$P\left\{\frac{\sqrt{n}\left(\hat{\theta}-\theta_o\right)}{\sigma} \le x\right\} =$$

$$\Phi(x) + n^{-\frac{1}{2}} p_1(x)\phi(x) + \cdots + n^{-\frac{j}{2}} p_j(x)\phi(x) + \cdots, \quad (5)$$

where $\phi(x) = (2\pi)^{-1/2} e^{-\frac{x^2}{2}}$ is the Standard Normal density function and $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$ is the Standard Normal distribution function. The functions p_j are polynomials with coefficients depending on cumulants of $\hat{\theta}-\theta_o$.

From Hall (1992), the Edgeworth expansion for the sample variance is

$$P\left\{\frac{\sqrt{n}\left(S^2-\sigma^2\right)}{\tau} \le x\right\}$$

$$= \Phi(x) + n^{-\frac{1}{2}} p_1(x)\phi(x) + \cdots + n^{-\frac{j}{2}} p_j(x)\phi(x) + \cdots, \quad (6)$$

where

$$p_1 = -\left(B_1 + B_2\frac{x^2-1}{6}\right),\ B_1 = -(\nu_4 - 1)^{-1/2},$$

$$B_2 = (\nu_4 - 1)^{-3/2}\left(\nu_6 - 3\nu_4 - 6\nu_3^2 + 2\right),$$

$$\nu_j = E\{(X-\mu)/\sigma\}^j,$$

and $\tau = \sqrt{E(X-\mu)^4 - \sigma^4}$.

The variable S^2 admits the inversion of the Edgeworth expansion as follows:

$$P\left\{\frac{\sqrt{n}\left(S^2-\sigma^2\right)}{\tau} \le x + n^{\frac{-1}{2}}\left(B_1 + B_2\frac{x^2-1}{6}\right)\right\}.$$

$$= \Phi(x) + o(n^{-1/2}) \quad (7)$$

To test $H_o : \sigma^2 = \sigma_o^2$ versus $H_a : \sigma^2 > \sigma_o^2$, one can adapt the inversion

formula of the Edgeworth expansion, and the result is an intuitive decision rule as follows:

$$\text{Reject } H_o \text{ if } Z > z_\alpha + n^{\frac{-1}{2}}\left(\hat{B}_1 + \hat{B}_2\frac{z_\alpha^2 - 1}{6}\right), \quad (8)$$

where z_α is the upper α percentage point of the standard normal distribution,

$$Z = \frac{S^2 - \sigma_0^2}{\tau/\sqrt{n}},\ \hat{B}_1 = -\left(\frac{S^4}{k_4 + 2S^4}\right)^{1/2},$$

$$\hat{B}_2 = -\left(\frac{k_6 + 12k_4S^2 + 4k_3^2 + 8(S^2)^3}{(k_4 + 2S^4)^{3/2}}\right).$$

Simulation Study

Details for the simulation study are provided in this section. The study is used to compare Type I error rates and the associated power performance of the different right-tail tests for variance.

Distributions Examined

Distributions were chosen to achieve a range of skewness (0.58 to 9.49) or kurtosis (-1.00 to 75.1) for comparing the test procedures. The skewed distributions considered in the study included Weibull with scale parameter $\lambda = 1.0$ and shape parameters = 0.5, 0.8, 2.0 (see Kendall, 1994), Lognormal($\mu = 0, \sigma = 1$), (see Evans, Hastings, & Peacock, 2000), Gamma with scale parameter 1.0 and shape parameters = 0.15,1.2,4.0 (see Evans, Hastings, & Peacock, 2000), 10 Inverse Guassian distributions with $\mu = 1.0$, scale parameters $\lambda =$ 0.1 to 25.0 with skewness ranging from 0.6 to 9.49 (see Chhikara & Folks, 1989 and Evans, Hastings, & Peacock, 2000), Exponential with $\mu = 1.0$ and $\lambda =$ 1.0 (see Evans, Hastings, & Peacock, 2000), Chi-square with ν degrees of freedom (ν = 1, 2, 3, 4, 8, 12, 16, 24), and a polynomial function of the standard normal distribution Barnes2 (see Fleishman 1978).

The heavy-tailed distributions considered included Student's T (ν = 5,6,8,16,32,40), 10 JTB(α, τ) distributions with ($\mu = 0, \sigma = 1$) and various α , τ values including Laplace(α =2.0, τ =1.0) , (see Johnson, Tietjen, & Beckman, 1980), and special designed distributions which are polynomial functions of the standard normal distribution: Barnes1 and Barnes3 having kurtosis 6.0 and 75.1 respectively (see Fleishman 1978). All the heavy-tailed distributions are symmetric with the exception of Barnes3. Barnes3 has skewness of .374 which is negligible in comparison to the kurtosis of 75.1. Therefore, Barnes3 was considered very close to symmetric.

Simulation Description

Simulations were run using Fortran 90 for Windows on an emachines etower 400i PC computer. All the Type I error and power comparisons for the test procedures used a simulation size of 100,000 in order to reduce experimental noise. Fortran 90 IMSL library was used to generate random numbers from these distributions: Weibull, Lognormal, Gamma, Exponential, Chi-square, Normal and Student's T. In addition, the Inverse Gaussian, JTB, Barnes1, Barnes2, and Barnes3 random variates were created with Fortran 90 program subroutines using the IMSL library's random number generator for normal, gamma, and uniform in various parts of the program.

The following tests were compared in the simulation study:

1) $\chi^2 = (n-1)S^2/\sigma_0^2$; the decision rule is Reject H_0 if $\chi^2 > \chi^2_{n-1,\alpha}$.

2) $\chi_r^2 = (n-1)\hat{d}S^2/\sigma_0^2$; the decision rule is Reject H_0 if $\chi_r^2 > \chi^2_{\nu,\alpha}$, where ν is the smallest integer that is greater than or equal to (n-1) $\hat{d}$.

3) Zs = $\dfrac{\dfrac{S^2}{\sigma_0^2}-1}{\sqrt{\dfrac{k_4}{nS^2\sigma_0^2}+\dfrac{2}{n-1}}}$ from Lee and Sa (1998); the decision rule is Reject H_0 if $Zs - a(Zs^2-1) + 2a^2(Zs^3 - Zs) > z_\alpha$.

4) Zh = $\dfrac{\dfrac{S^2}{\sigma_0^2}-1}{\sqrt{\dfrac{k_4}{n\sigma_0{}^4}+\dfrac{2}{n-1}}}$ from Lee and Sa (1996); the decision rule is Reject H_0 if $Zh - a(Zh^2-1) + 2a^2(Zh^3 - Zh) > z_\alpha$.

5) The proposed test is Z = $\dfrac{S^2-\sigma_0^2}{\tau/\sqrt{n}}$, where $\tau/\sqrt{n}$ can be estimated by different forms to create different test statistics; the decision rule is

Reject H_0 if Z > $z_\alpha + n^{\frac{-1}{2}}\left(\hat{B}_1 + \hat{B}_2\dfrac{z_\alpha^2-1}{6}\right)$.

Six different test statistics were investigated:

$$\text{Zn} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{k_4}{n}+\dfrac{2S^2\sigma_0^2}{n-1}}},$$

$$\text{Z2} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{k_4}{n}+\dfrac{2\sigma_0^4}{n-1}}},$$

$$\text{Z3} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{k_4}{n}+\dfrac{2S^4}{n-1}}},$$

$$\text{Z4} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{(n-1)k_4}{n(n+1)}+\dfrac{2S^4}{n+1}}},$$

$$\text{Z5} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{k_4\sigma_0^4}{nS^4}+\dfrac{2\sigma_0^4}{n-1}}},$$

and

$$\text{Z6} = \frac{S^2-\sigma_0^2}{\sqrt{\dfrac{k_4\sigma_0^2}{nS^2}+\dfrac{2\sigma_0^4}{n-1}}}.$$

The equation $\dfrac{(n-1)k_4}{n(n+1)}+\dfrac{2S^4}{n+1}$ in Z4 is the unbiased estimator for $V(S^2) = \tau^2/n$. Sample sizes of 20 and 40 were investigated for Type I error rates along with the nominal alpha levels 0.01, 0.02, 0.05, and 0.10 for each sample size. Furthermore, any test that used z_α also used $(z_\alpha + t_{n-1,\alpha})/2$ and $t_{n-1,\alpha}$ separately with each sample size and nominal level for further flexibility in determining the best test. For each sample size and nominal level, 100,000 simulations were generated from each distribution. All the tests investigated were applied to each sample. The proportion of samples rejected from the 100,000 was then recorded based on the sample size, nominal level, and test procedure.

The steps for conducting the simulation were as follows:

1. Generate a sample of size *n* from one parent distribution under H_0.

2. Calculate: $\overline{X}$, S^2, k_3, k_4, k_6, $\hat{\beta}_1$, $\hat{B}_1$, $\hat{B}_2$.

3. Calculate all the test statistics: χ^2, χ_r^2, Zs, Zh, Zn, Z2, Z3, Z4, Z5, and Z6.

4. Find the critical value for each test considered.

5. Determine for each test whether rejection is warranted for the current sample and if so, increment the respective counter.

6. Repeat 1 through 5 for the remaining 99,999 samples.

7. Calculate the proportion of 100,000 rejected.

A power study was performed using five skewed distributions and five heavy-tailed distributions with varying degrees of skewness and kurtosis respectively. For each distribution considered, sample sizes of 20 and 40 were examined with nominal levels of 0.10 and 0.01, and k = 1,2,3,4,5,6, where k is a constant such that the $\sqrt{k}$ is multiplied to each variate.

The traditional power studies were performed by multiplying the distribution observations by $\sqrt{k}$ to create a new set of observations yielding a variance k times larger than the H_0 value. Steps 1 through 6 above would then be implemented for the desired values of k, sample sizes, and significance levels. The power would then be the proportion of 100,000 rejected for the referenced value of k, sample size, and significance level.

This method has been criticized by many researchers since tests with high Type I error rates frequently have high power also. Tests with high Type I error rates usually have fixed lower critical points relative to other tests and therefore reject more easily when the true variance is increased. Hence, these tests tend to have higher power.

Some researchers are using a method to correct this problem. With k =1, the critical point for each test under investigation is adjusted till the proportion rejected out of 100,000 is the same as the desired nominal level. The concept is that the tests can be compared better for power afterward since all the tests have critical points adjusted to approximately the same Type I error rate. Once this is accomplished, steps 1 through 7 above are performed for each k under consideration to get a better power comparison between the different tests at that level of k.

The traditional power study and the new power study were used to provide a complete picture of the power performance by each test.

Results

Type I Error Comparison

Comparisons of Type I error rates for skewed and heavy-tailed distributions were made for sample size 40 and 20 with levels of significance 0.10, 0.05, 0.02, and 0.01. However, the results are very similar between the two higher levels of significance (0.10 and 0.05) and the same situation holds for the two lower levels of significance. Therefore, only 0.05 and 0.01 levels are reported here and they are summarized into Tables 1 through 4. Also, it can be observed that the Type I error performances are quite similar for the skewed distributions with similar coefficient of skewness or for the heavy-tailed distributions with similar coefficient of kurtosis. Therefore, only 11 out of the original 27 skewed distributions and 10 out of the 18 heavy-tailed distributions studied are reported in these tables. For the complete simulation results, please see Long and Sa (2003).

Comparisons were made between the tests χ^2, χ_r^2 (first and second number in the first column), and Zs, Zh, Z2, and Z6 with z_α, $(z_\alpha + t_{n-1,\alpha})/2$, and $t_{n-1,\alpha}$ as the first, second, and third number in the respective column. The tests Zn, Z3, Z4, and Z5 were left out of the table since they were either unstable over different distributions or had highly inflated Type I error rates. From Tables 1 through 4, the following points can be observed:

The traditional χ^2 test is more inflated than the other tests for all the distributions, sample sizes and significance levels.

The χ_r^2 test does not maintain the Type I error rates well for the skewed distribution cases. The Type I error rates can be more than 300% inflated than the desired level of significance in some of the distributions. This is especially true for the distributions with a higher coefficient of skewness. However, the χ_r^2 test performs much better in the heavy-tailed

distributions. Although there are still some inflated cases, they are not severe. These results are understandable since the χ_r^2 test only adjusts for the kurtosis of the sampled distribution and not the skewness.

The Z2 test's Type I error rates reported in Tables 1 and 2 were extremely conservative for most of the skewed distributions. It becomes even more conservative when the coefficient of skewness gets larger. In fact, the Z2 test is so conservative it is rarely inflated for any of the skewed or heavy-tailed distribution cases.

Similar to the Z2 test, test Zh performs quite conservatively in all the skewed distributions as well. However, it performs differently under heavy-tailed distributions. The Type I error rates become closer to the nominal level except for one distribution, and there are even a few inflated cases. The exception in the heavy-tailed distributions is the Barnes3. In this case, test Zh is extremely conservative for all the nominal levels.

Under the skewed distribution, the Zs test performs well for the sample size 40 and the nominal level 0.05. However, the Type I error rates become more or less uncontrollable when either the alpha level gets small or the sample size is reduced. These results confirmed the recommendations of Lee and Sa (1998) that Zs is more suitable for moderate to large sample sizes and alpha levels not too small. Although Zs was specifically designed for the skewed distributions, it actually works reasonably well for the heavy-tailed distributions as long as the sample size and/or the alpha level are not too small.

Generally speaking, the proposed test Z6 controls Type I error rates the best in both the skewed distribution cases and the heavy-tailed distribution cases. Only under some skewed distributions with both small alpha and small sample size were there a few inflated Type I error rates. However, the rates of inflation are at much more acceptable level than some others.

Power Comparison Results

One of the objectives of the study is to find one test for non-normal distributions with an improved Type I error rate and power over earlier tests. It was suspected that tests with very conservative Type I error rates might have lower power than other tests since it is harder to reject with these tests. Because tests Zh and Z2 were extremely conservative for the skewed distributions, exploratory power simulations were run on a couple of mildly skewed distributions with Zs, Zh, Z2, and Z6 to further decrease the potential tests. The preliminary power comparisons confirmed our suspicion. Both Zh and Z2 have extremely low power even when k is as large as 6.0. Therefore, Z2 will not be looked at further since Z6 is the better performer of the new tests. Also, the Zh test's power is unacceptable, but it will still be compared for the heavy-tailed distributions since that is what it was originally designed for. The results of the preliminary power study are reported in Long and Sa (2003).

Tables 5 and 6 provide the partial results from the new type of power comparisons, and Tables 7 and 8 consist of some results from the traditional type of power study. Based on the complete power study in Long and Sa (2003), the following expected similarities can be found for the power performance of the tests between the skewed and heavy-tailed distributions regardless of the type of power study. When the sample size decreases from 40 to 20, the power decreases. As the k in $k \cdot \sigma_0^2$ increases, the power increases. When the significance level decreases from 0.10 to 0.01, the power decreases more than the decrease experienced with the sample size decrease. As the skewness of the skewed distribution decreases, the power increases. As the kurtosis of the heavy-tailed distribution decreases, the power increases overall with a slight decrease from the T(5) distribution to the Laplace distribution.

The primary difference overall between the skewed and heavy-tailed distributions is that the power is better for the heavy-tailed distributions when comparing the same sample size, significance level, and k. In fact, the power increases more quickly over the levels of k for the heavy-tailed distributions versus the skewed distributions, with a more noticeable difference at the higher levels of kurtosis and skewness respectively.

Some specific observations are summarized as follows:

It can be observed that the χ^2 test performed worst overall with its power lower than the other tests' power based on the new power study. There are several cases where the χ^2 test's power is lower than the other tests' powers by 10% or more. As can be expected, the χ^2 test has very good power performance under the traditional power study, which provides the true rejection power under the specific alternative hypothesis. However, since the test had uncontrollable and unstable Type I error rates, this test should not be used with confidence.

The χ_r^2 test has a better power performance than the χ^2 test in the new power study, and it performs as well as the χ^2 test in the traditional power study. But similar to the χ^2, the test is not recommended due to the unstable Type I error rates.

Differences between the power performances of the Z6 and Zs tests are very minor, and they are slightly better than the χ_r^2 test in the new power study. More than 50% of the cases studied have differences in power within 2% between any two tests. In the traditional power study, the Z6 and Zs tests are not as powerful as either the χ^2 test or the χ_r^2 test for the skewed distributions studied. However, they perform quite well also. On the heavy-tailed distributions studied, the Z6 and Zs tests have very good power performance which is constantly as high as the power of the χ_r^2 test, and sometimes almost as high as the power of the χ^2 test. To further differentiate the two in the traditional power study, the Z6 test performed better than Zs when $\alpha = 0.10$ and worse when $\alpha = 0.01$.

The Zh test is studied only for the heavy-tailed distributions. With the adjusted critical values on the new power study, Zh has the most power among the five tests. However, as far as the true rejection power is concerned, it has the lowest power in almost all of the cases studied.

More Comparisons of Type I Error Rates Between Zs and Z6

After reviewing the results from the Type I error rate comparison study and the power study, the tests Zs and Z6 are the best. Therefore, the two tests were examined for a Type I error rate comparison study of sample size 30. Looking at the skewed distributions and heavy-tailed distributions in Table 9, both tests held the Type I error rates well at $\alpha = 0.10$ and $\alpha = 0.05$. For the skewed distributions, the Zs test's Type I error rates were much more inflated overall for the lower alpha levels of 0.02 and 0.01. In fact, the number of inflated cases for Zs compared to Z6 was more than double. Breadth of the inflation was also larger with the Zs test having 22% of the cases greater than a 50% inflation rate (i.e. 50% higher than the desired nominal level), while the Z6 test had none. Similar results can be observed for the heavy-tailed distributions as well. Clearly, the Z6 test controls Type I error rates better than the Zs test for sample sizes of 30 also.

Although most of the Type I error rates for the Z6 test are stable, there was some inflation. However, the inflation is still within a reasonable amount of the nominal level. It should be noted that the Z6 test's Type I error rates for alpha 0.01 are in control if $t_{n-1,\alpha}$ is used in the critical values. Therefore, if the practitioner is very concerned with Type I error, it is recommended that the Z6 test with $t_{n-1,\alpha}$ should be used for small alphas. In addition, since the method involves higher moments such as k_6 and has (n-5) in the denominator of k_6, it is recommended that sample sizes of 30 or more be used. Even so, the simulation study found the Type I error rates for the Z6 test to be reasonable for sample sizes of 20.

Table 1. Comparison of Type I Error Rates when n=40, Skewed Distributions

Distribution (skewness)	α=0.01 χ^2, χ_r^2	Zs	Zh	Z2	Z6
IG (1.0,0.1)	.1616	.0259	.0004	.0001	.0121
(9.49)	.0429	.0250	.0003	.0000	.0110
		.0237	.0003	.0000	.0100
Weibull(1.0,0.5)	.1522	.0198	.0012	.0001	.0090
(6.62)	.0349	.0188	.0011	.0001	.0082
		.0177	.0010	.0001	.0074
LN(0,1)	.1325	.0156	.0012	.0001	.0073
(6.18)	.0274	.0148	.0011	.0001	.0065
		.0141	.0009	.0000	.0057
IG(1.0,0.25)	.1671	.0192	.0014	.0002	.0093
(6.00)	.0349	.0179	.0013	.0001	.0082
		.0168	.0011	.0001	.0074
Gamma(1.0,0.15)	.1704	.0166	.0025	.0003	.0092
(5.16)	.0322	.0154	.0024	.0003	.0081
		.0144	.0022	.0003	.0073
IG(1.0,0.5)	.1538	.0135	.0032	.0005	.0077
(4.24)	.0271	.0126	.0029	.0004	.0069
		.0117	.0028	.0004	.0061
Chi(1)	.1282	.0113	.0073	.0019	.0094
(2.83)	.0194	.0102	.0069	.0017	.0085
		.0094	.0065	.0015	.0077
Exp(1.0)	.0949	.0119	.0115	.0045	.0116
(2.00)	.0159	.0110	.0109	.0041	.0104
		.0100	.0103	.0037	.0097
Chi(2)	.0922	.0114	.0114	.0045	.0109
(2.00)	.0150	.0103	.0107	.0041	.0099
		.0095	.0100	.0038	.0091
Barnes2	.0716	.0141	.0154	.0079	.0150
(1.75)	.0127	.0127	.0146	.0072	.0137
		.0116	.0138	.0065	.0124
IG(1.0,25.0)	.0217	.0102	.0113	.0089	.0107
(0.60)	.0092	.0090	.0104	.0078	.0095
		.0081	.0093	.0067	.0084

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 1 (continued). Comparison of Type I Error Rates when n=40, Skewed Distributions

Distribution	α=0.05				
(skewness)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
IG (1.0,0.1)	.1859	.0532	.0015	.0007	.0448
(9.49)	.0761	.0520	.0015	.0007	.0433
		.0509	.0014	.0006	.0419
Weibull(1.0,0.5)	.1899	.0467	.0037	.0017	.0402
(6.62)	.0683	.0454	.0035	.0016	.0387
		.0442	.0033	.0015	.0372
LN(0,1)	.1701	.0415	.0043	.0022	.0362
(6.18)	.0610	.0404	.0040	.0021	.0347
		.0392	.0039	.0019	.0331
IG(1.0,0.25)	.1992	.0479	.0446	.0022	.0417
(6.00)	.0719	.0467	.0437	.0019	.0401
		.0454	.0418	.0017	.0385
Gamma(1.0,0.15)	.2148	.0486	.0078	.0043	.0430
(5.16)	.0743	.0469	.0075	.0039	.0412
		.0454	.0072	.0035	.0397
IG(1.0,0.5)	.1994	.0442	.0094	.0050	.0395
(4.24)	.0672	.0423	.0090	.0046	.0378
		.0408	.0087	.0043	.0360
Chi(1)	.1906	.0439	.0203	.0136	.0431
(2.83)	.0622	.0421	.0197	.0130	.0416
		.0406	.0191	.0124	.0397
Exp(1.0)	.1583	.0441	.0299	.0229	.0460
(2.00)	.0559	.0424	.0289	.0218	.0442
		.0408	.0279	.0209	.0425
Chi(2)	.1557	.0430	.0293	.0226	.0453
(2.00)	.0545	.0414	.0285	.0214	.0434
		.0399	.0278	.0204	.0415
Barnes2	.1414	.0485	.0388	.0340	.0531
(1.75)	.0549	.0466	.0376	.0324	.0511
		.0451	.0364	.0309	.0493
IG(1.0,25.0)	.0732	.0429	.0407	.0429	.0498
(0.60)	.0442	.0413	.0390	.0410	.0477
		.0397	.0376	.0389	.0454

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 2. Comparison of Type I Error Rates when n=20, Skewed Distributions

Distribution	α=0.01				
(skewness)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
IG (1.0,0.1)	.1215	.0342	.0003	.0003	.0149
(9.49)	.0443	.0321	.0003	.0003	.0122
		.0302	.0003	.0002	.0104
Weibull(1.0,0.5)	.1227	.0294	.0012	.0012	.0139
(6.62)	.0386	.0270	.0011	.0011	.0115
		.0249	.0009	.0009	.0098
LN(0,1)	.1082	.0246	.0013	.0014	.0119
(6.18)	.0316	.0226	.0012	.0012	.0100
		.0209	.0010	.0011	.0083
IG(1.0,0.25)	.1295	.0307	.0015	.0015	.0142
(6.00)	.0406	.0281	.0014	.0014	.0120
		.0258	.0013	.0012	.0098
Gamma(1.0,0.15)	.1408	.0296	.0024	.0025	.0152
(5.16)	.0396	.0269	.0021	.0021	.0128
		.0243	.0019	.0018	.0108
IG(1.0,0.5)	.1272	.0258	.0029	.0030	.0141
(4.24)	.0336	.0231	.0024	.0026	.0119
		.0208	.0022	.0023	.0102
Chi(1)	.1096	.0228	.0067	.0079	.0185
(2.83)	.0265	.0201	.0059	.0070	.0161
		.0176	.0051	.0061	.0139
Exp(1.0)	.0810	.0203	.0092	.0107	.0191
(2.00)	.0202	.0175	.0079	.0093	.0165
		.0153	.0067	.0080	.0144
Chi(2)	.0825	.0206	.0095	.0111	.0196
(2.00)	.0205	.0180	.0083	.0097	.0168
		.0156	.0071	.0082	.0145
Barnes2	.0680	.0228	.0127	.0159	.0238
(1.75)	.0192	.0198	.0112	.0137	.0206
	.0171	.0097	.0119	.0180	
IG(1.0,25.0)	.0213	.0134	.0105	.0098	.0120
(0.60)	.0095	.0113	.0087	.0079	.0095
		.0095	.0072	.0064	.0076

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 2 (continued). Comparison of Type I Error Rates when n=20, Skewed Distributions

Distribution	α=0.05				
(skewness)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
IG (1.0,0.1)	.1451	.0566	.0014	.0015	.0459
(9.49)	.0736	.0547	.0013	.0014	.0430
		.0530	.0011	.0012	.0399
Weibull(1.0,0.5)	.1538	.0534	.0033	.0039	.0444
(6.62)	.0706	.0514	.0031	.0035	.0412
		.0493	.0028	.0031	.0385
LN(0,1)	.1377	.0471	.0482	.0057	.0397
(6.18)	.0603	.0451	.0435	.0051	.0369
		.0431	.0406	.0046	.0343
IG(1.0,0.25)	.1652	.0579	.0046	.0053	.0473
(6.00)	.0760	.0552	.0041	.0047	.0437
		.0528	.0038	.0043	.0407
Gamma(1.0,0.15)	.1805	.0604	.0073	.0079	.0505
(5.16)	.0575	.0568	.0069	.0072	.0471
		.0549	.0064	.0064	.0438
IG(1.0,0.5)	.1686	.0560	.0089	.0104	.0484
(4.24)	.0725	.0535	.0083	.0095	.0446
		.0509	.0077	.0087	.0416
Chi(1)	.1635	.0545	.0176	.0215	.0523
(2.83)	.0669	.0515	.0165	.0200	.0486
		.0484	.0155	.0186	.0455
Exp(1.0)	.1394	.0529	.0260	.0313	.0544
(2.00)	.0604	.0496	.0241	.0291	.0506
		.0468	.0226	.0272	.0473
Chi(2)	.1406	.0543	.0264	.0317	.0565
(2.00)	.0605	.0511	.0245	.0293	.0524
		.0482	.0229	.0273	.0489
Barnes2	.1307	.0560	.0342	.0416	.0617
(1.75)	.0587	.0530	.0321	.0389	.0577
	.0499	.0302	.0364	.0542	
IG(1.0,25.0)	.0687	.0449	.0377	.0433	.0507
(0.60)	.0437	.0419	.0349	.0398	.0464
		.0388	.0322	.0365	.0424

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 3. Comparison of Type I Error Rates when n=40, Heavy-tailed Distributions

Distribution	α=0.01				
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
Barnes3	.1269	.0167	.0001	.0000	.0060
(75.1)	.0280	.0158	.0001	.0000	.0052
		.0151	.0001	.0000	.0047
T(5)	.0629	.0075	.0084	.0027	.0058
(6.00)	.0111	.0066	.0079	.0024	.0050
		.0059	.0074	.0021	.0045
Barnes1	.1081	.0118	.0126	.0021	.0089
(6.00)	.0188	.0105	.0119	.0019	.0078
		.0093	.0111	.0017	.0068
T(6)	.0526	.0085	.0108	.0044	.0075
(3.00)	.0103	.0076	.0100	.0040	.0067
		.0067	.0092	.0034	.0059
Laplace(2.0,1.0)	.0608	.0099	.0138	.0043	.0092
(3.00)	.0124	.0089	.0130	.0038	.0081
		.0080	.0120	.0034	.0072
JTB(4.0,1.0)	.0246	.0103	.0127	.0082	.0106
(0.78)	.0098	.0092	.0118	.0074	.0095
		.0084	.0109	.0067	.0084
T(16)	.0198	.0103	.0118	.0088	.0104
(0.50)	.0095	.0092	.0107	.0079	.0092
		.0083	.0098	.0070	.0083
JTB(1.25,0.5)	.0134	.0102	.0112	.0097	.0108
(0.24)	.0089	.0091	.0101	.0086	.0095
		.0081	.0090	.0075	.0083
T(32)	.0139	.0091	.0100	.0084	.0093
(0.21)	.0083	.0084	.0093	.0075	.0083
		.0076	.0085	.0067	.0074
JTB(2.0,0.5)	.0061	.0064	.0068	.0060	.0061
(-0.30)	.0055	.0056	.0059	.0051	.0052
		.0049	.0052	.0043	.0044

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 3 (continued). Comparison of Type I Error Rates when n=40, Heavy-tailed Distributions

Distribution	α=0.05				
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
Barnes3	.1554	.0390	.0011	.0003	.0315
(75.1)	.0590	.0380	.0011	.0002	.0302
		.0371	.0010	.0002	.0290
T(5)	.1184	.0362	.0262	.0198	.0369
(6.00)	.0456	.0348	.0254	.0188	.0352
		.0332	.0247	.0178	.0335
Barnes1	.1786	.0492	.0327	.0201	.0484
(6.00)	.0655	.0472	.0317	.0190	.0462
		.0453	.0308	.0179	.0444
T(6)	.1054	.0376	.0310	.0257	.0400
(3.00)	.0449	.0360	.0300	.0243	.0381
		.0345	.0290	.0231	.0363
Laplace(2.0,1.0)	.1263	.0417	.0359	.0268	.0449
(3.00)	.0500	.0400	.0349	.0254	.0431
		.0385	.0338	.0241	.0413
JTB(4.0,1.0)	.0770	.0447	.0428	.0429	.0506
(0.78)	.0464	.0429	.0410	.0409	.0487
		.0414	.0396	.0391	.0466
T(16)	.0683	.0436	.0419	.0438	.0498
(0.50)	.0448	.0419	.0402	.0420	.0479
		.0402	.0388	.0401	.0457
JTB(1.25,0.5)	.0577	.0445	.0431	.0481	.0515
(0.24)	.0441	.0428	.0414	.0459	.0493
		.0411	.0400	.0442	.0474
T(32)	.0591	.0444	.0434	.0471	.0510
(0.21)	.0444	.0425	.0419	.0448	.0489
		.0407	.0402	.0430	.0467
JTB(2.0,0.5)	.0381	.0344	.0355	.0396	.0405
(-0.30)	.0348	.0327	.0338	.0377	.0385
		.0312	.0323	.0359	.0366

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 4. Comparison of Type I Error Rates when n=20, Heavy-tailed Distributions

Distribution	α=0.01				
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
Barnes3	.0964	.0241	.0001	.0001	.0076
(75.1)	.0290	.0221	.0001	.0001	.0062
		.0207	.0001	.0001	.0049
T(5)	.0543	.0151	.0072	.0056	.0100
(6.00)	.0147	.0125	.0060	.0046	.0082
		.0107	.0052	.0037	.0063
Barnes1	.0590	.0205	.0084	.0059	.0136
(6.00)	.0225	.0178	.0072	.0048	.0111
		.0153	.0062	.0039	.0092
T(6)	.0461	.0146	.0088	.0070	.0110
(3.00)	.0131	.0122	.0075	.0055	.0088
		.0104	.0062	.0044	.0070
Laplace(2.0,1.0)	.0053	.0165	.0105	.0083	.0139
(3.00)	.0153	.0138	.0089	.0068	.0113
		.0117	.0077	.0055	.0092
JTB(4.0,1.0)	.0238	.0143	.0115	.0100	.0126
(0.78)	.0107	.0118	.0096	.0079	.0098
		.0098	.0081	.0061	.0076
T(16)	.0184	.0128	.0104	.0092	.0108
(0.50)	.0093	.0106	.0086	.0073	.0084
		.0089	.0072	.0058	.0066
JTB(1.25,0.5)	.0138	.0138	.0120	.0104	.0115
(0.24)	.0094	.0114	.0099	.0079	.0087
		.0096	.0081	.0062	.0069
T(32)	.0134	.0121	.0103	.0087	.0101
(0.21)	.0079	.0099	.0084	.0066	.0076
		.0079	.0066	.0050	.0056
JTB(2.0,0.5)	.0059	.0091	.0075	.0054	.0057
(-0.30)	.0051	.0076	.0059	.0038	.0040
		.0061	.0046	.0026	.0028

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 4 (continued). Comparison of Type I Error Rates when n=20, Heavy-tailed Distributions

Distribution	α=0.05				
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z2	Z6
Barnes3	.1184	.0430	.0009	.0007	.0319
(75.1)	.0544	.0414	.0008	.0005	.0294
		.0397	.0007	.0005	.0268
T(5)	.1034	.0439	.0233	.0249	.0440
(6.00)	.0489	.0409	.0215	.0225	.0398
		.0383	.0199	.0206	.0362
Barnes1	.1509	.0570	.0244	.0243	.0544
(6.00)	.0674	.0537	.0225	.0220	.0496
		.0502	.0206	.0201	.0456
T(6)	.0968	.0449	.0283	.0228	.0469
(3.00)	.0482	.0417	.0260	.0279	.0428
		.0388	.0240	.0254	.0395
Laplace(2.0,1.0)	.1166	.0493	.0303	.0324	.0516
(3.00)	.0537	.0458	.0281	.0298	.0475
		.0427	.0261	.0271	.0439
JTB(4.0,1.0)	.0742	.0468	.0386	.0436	.0520
(0.78)	.0463	.0434	.0361	.0400	.0479
		.0404	.0335	.0367	.0443
T(16)	.0658	.0440	.0381	.0430	.0494
(0.50)	.0429	.0408	.0350	.0391	.0454
		.0377	.0324	.0355	.0415
JTB(1.25,0.5)	.0587	.0457	.0417	.0483	.0529
(0.24)	.0434	.0420	.0387	.0439	.0483
		.0391	.0357	.0401	.0441
T(32)	.0583	.0447	.0406	.0462	.0512
(0.21)	.0430	.0415	.0375	.0421	.0468
		.0382	.0344	.0382	.0423
JTB(2.0,0.5)	.0387	.0359	.0350	.0394	.0410
(-0.30)	.0338	.0325	.0320	.0350	.0364
		.0298	.0291	.0313	.0325

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, Z2, and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs, Zh, Z2, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 5. New Power Comparisons for Skewed Distribution Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$, significance level 0.100, n = 40

Distribution	$k = 1.0$			$k = 2.0$			$k = 3.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.101	.102	.101	.280	.315	.315	.439	.499	.501
(6.62)	.098	.099	.098	.303	.309	.308	.485	.494	.493
Gamma(1.0,0.15)	.099	.098	.100	.318	.339	.344	.490	.523	.528
(5.16)	.100	.098	.101	.340	.340	.345	.523	.524	.528
IG(1.0,0.6)	.100	.099	.099	.382	.439	.441	.612	.695	.698
(3.87)	.100	.099	.102	.432	.437	.447	.685	.694	.703
Chi(2)	.098	.098	.098	.634	.703	.704	.903	.940	.940
(2.00)	.098	.098	.100	.697	.703	.708	.937	.940	.941

n = 40 (continued)

Distribution	$k = 4.0$			$k = 5.0$			$k = 6.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.563	.634	.636	.623	.729	.731	.725	.797	.799
(6.62)	.619	.629	.629	.715	.725	.725	.784	.793	.794
Gamma(1.0,0.15)	.611	.648	.653	.697	.731	.736	.763	.794	.798
(5.16)	.648	.649	.654	.731	.732	.737	.793	.794	.799
IG(1.0,0.6)	.762	.837	.839	.852	.912	.914	.906	.950	.951
(3.87)	.828	.836	.842	.906	.912	.916	.946	.950	.952
Chi(2)	.975	.987	.988	.993	.997	.997	.998	.999	.999
(2.00)	.987	.987	.988	.997	.997	.997	.999	.999	.999

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 5 (continued). New Power Comparisons for Skewed Distribution Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$, significance level 0.100

n = 20

Distribution	$k = 1.0$			$k = 2.0$			$k = 3.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.100	.101	.102	.231	.253	.255	.343	.382	.385
(6.62)	.101	.100	.101	.248	.251	.254	.374	.380	.384
Gamma(1.0,0.15)	.100	.101	.100	.254	.266	.265	.375	.394	.393
(5.16)	.100	.101	.100	.263	.267	.266	.389	.395	.394
IG(1.0,0.6)	.099	.098	.101	.295	.331	.340	.459	.519	.531
(3.87)	.098	.098	.100	.325	.332	.337	.511	.520	.528
Chi(2)	.099	.102	.102	.469	.525	.527	.729	.786	.788
(2.00)	.099	.100	.098	.514	.521	.519	.777	.783	.781

n = 20 (continued)

Distribution	$k = 4.0$			$k = 5.0$			$k = 6.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.432	.481	.484	.502	.557	.560	.570	.627	.631
(6.62)	.471	.478	.483	.546	.554	.559	.616	.625	.629
Gamma(1.0,0.15)	.465	.488	.487	.532	.557	.556	.585	.611	.610
(5.16)	.483	.490	.488	.551	.558	.557	.606	.612	.610
IG(1.0,0.6)	.586	.657	.667	.676	.748	.757	.742	.811	.818
(3.87)	.648	.658	.665	.739	.748	.755	.802	.811	.816
Chi(2)	.862	.903	.904	.925	.952	.952	.959	.975	.976
(2.00)	.898	.901	.900	.949	.951	.950	.974	.975	.975

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 6. New Power Comparisons for Heavy-tail Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$ and significance level 0.100

n = 40

Distribution	$k = 1.0$				$k = 2.0$				$k = 3.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.101	.102	.100	.099	.266	.413	.460	.418	.457	.904	.934	.913
(75.1)	.099	.098	.098	.097	.381	.405	.457	.416	.874	.898	.933	.912
T(5)	.099	.099	.099	.099	.775	.841	.853	.844	.978	.989	.991	.990
(6.00)	.101	.100	.101	.101	.840	.842	.856	.846	.989	.990	.991	.990
Laplace(2,1)	.102	.101	.101	.101	.766	.801	.797	.801	.968	.978	.976	.979
(3.00)	.101	.102	.102	.101	.798	.801	.821	.801	.978	.979	.980	.979
T(8)	.097	.099	.099	.102	.845	.902	.903	.905	.995	.997	.997	.997
(1.50)	.099	.101	.098	.102	.901	.904	.903	.905	.996	.997	.997	.997

n = 40 (continued)

Distribution	$k = 4.0$				$k = 5.0$				$k = 6.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.737	.998	.999	.999	.963	1.00	1.00	1.00	.997	1.00	1.00	1.00
(75.1)	.997	.998	.999	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
T(5)	1.00	.999	.999	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(6.00)	.999	.999	.999	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Laplace(2,1)	.995	.997	.996	.997	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(3.00)	.997	.997	.998	.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
T(8)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(1.50)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs, Zh, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 6 (continued). New Power Comparisons for Heavy-tail Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$ and significance level 0.100

n = 20

Distribution	$k = 1.0$				$k = 2.0$				$k = 3.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.100	.099	.099	.100	.217	.302	.323	.314	.355	.733	.778	.763
(75.1)	.101	.101	.099	.098	.290	.306	.331	.309	.714	.739	.774	.755
T(5)	.102	.102	.101	.101	.584	.646	.662	.648	.868	.907	.914	.908
(6.00)	.100	.102	.101	.102	.637	.648	.662	.652	.900	.908	.914	.907
Laplace(2,1)	.099	.099	.101	.098	.565	.601	.613	.598	.834	.861	.863	.859
(3.00)	.099	.102	.101	.099	.560	.608	.604	.598	.860	.864	.862	.858
T(8)	.102	.100	.100	.100	.691	.714	.715	.714	.931	.940	.938	.940
(1.50)	.101	.102	.098	.098	.714	.716	.714	.711	.940	.941	.936	.939

n = 20 (continued)

Distribution	$k = 4.0$				$k = 5.0$				$k = 6.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.656	.958	.967	.966	.899	.993	.996	.995	.975	.999	.999	.999
(75.1)	.854	.960	.973	.964	.992	.993	.996	.994	.998	.999	.999	.999
T(5)	.960	.973	.976	.974	.986	.992	.992	.992	.995	.997	.997	.997
(6.00)	.972	.974	.976	.975	.992	.992	.992	.993	.997	.997	.997	.997
Laplace(2,1)	.936	.950	.951	.949	.973	.980	.978	.980	.988	.992	.990	.991
(3.00)	.950	.951	.950	.949	.980	.981	.986	.980	.992	.992	.992	.991
T(8)	.984	.986	.984	.986	.996	.997	.996	.997	.999	.999	.999	.999
(1.50)	.986	.986	.984	.986	.997	.997	.996	.997	.999	.999	.999	.999

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs, Zh, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 7. Traditional Power Comparisons for Skewed Distribution Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$, significance level 0.100, n = 40

Distribution	$k = 1.0$			$k = 2.0$			$k = 3.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.207	.078	.078	.464	.270	.272	.638	.448	.452
(6.62)	.100	.077	.077	.307	.267	.269	.488	.446	.448
Gamma(1.0,0.15)	.245	.088	.089	.529	.318	.322	.694	.500	.503
(5.16)	.114	.087	.087	.361	.315	.318	.542	.497	.500
IG(1.0,0.6)	.229	.081	.083	.600	.403	.409	.805	.666	.674
(3.87)	.104	.079	.081	.440	.399	.406	.696	.663	.670
Chi(2)	.201	.085	.092	.789	.680	.695	.959	.930	.935
(2.00)	.096	.083	.090	.698	.676	.692	.936	.929	.934

n = 40 (continued)

Distribution	$k = 4.0$			$k = 5.0$			$k = 6.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.749	.585	.589	.822	.687	.691	.870	.762	.766
(6.62)	.622	.582	.586	.717	.684	.688	.788	.762	.763
Gamma(1.0,0.15)	.786	.628	.631	.846	.715	.718	.883	.776	.779
(5.16)	.664	.626	.628	.746	.713	.715	.802	.774	.776
IG(1.0,0.6)	.902	.818	.823	.948	.899	.903	.971	.942	.944
(3.87)	.837	.816	.821	.910	.898	.901	.949	.941	.943
Chi(2)	.992	.986	.987	.998	.997	.997	1.00	.999	.999
(2.00)	.987	.985	.987	.997	.997	.997	.999	.999	.999

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 7 (continued). Traditional Power Comparisons for Skewed Distribution Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$, significance level 0.100, n = 20

Distribution	$k = 1.0$			$k = 2.0$			$k = 3.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.173	.080	.080	.354	.218	.220	.482	.336	.340
(6.62)	.097	.078	.078	.245	.214	.215	.364	.332	.334
Gamma(1.0,0.15)	.206	.092	.093	.402	.252	.254	.533	.377	.380
(5.16)	.112	.090	.090	.282	.248	.249	.408	.372	.374
IG(1.0,0.6)	.197	.089	.091	.457	.310	.317	.628	.495	.504
(3.87)	.106	.086	.088	.335	.304	.310	.519	.489	.497
Chi(2)	.183	.093	.103	.613	.503	.523	.833	.770	.785
(2.00)	.103	.090	.099	.518	.496	.515	.780	.765	.779

n = 20 (continued)

Distribution	$k = 4.0$			$k = 5.0$			$k = 6.0$		
(skewness)	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6	χ^2, χ_r^2	Zs	Z6
Weibull(1.0,0.5)	.578	.439	.443	.646	.516	.521	.699	.577	.582
(6.62)	.466	.433	.437	.541	.511	.514	.601	.572	.576
Gamma(1.0,0.15)	.615	.471	.473	.677	.546	.548	.722	.601	.604
(5.16)	.502	.466	.467	.574	.541	.542	.627	.597	.598
IG(1.0,0.6)	.741	.634	.643	.816	.731	.739	.863	.797	.805
(3.87)	.653	.629	.637	.747	.727	.734	.810	.794	.800
Chi(2)	.924	.893	.901	.964	.949	.953	.982	.973	.976
(2.00)	.898	.890	.897	.951	.947	.951	.974	.972	.975

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 8. Traditional Power Comparisons for Heavy-tail Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$ and significance level 0.100, n = 40

Distribution	$k = 1.0$				$k = 2.0$				$k = 3.0$			
(kurtosis)	χ^2	χ_r^2	Zs	Zh	Z6	χ^2	χ_r^2	Zs	Zh	Z6	χ^2	χ_r^2	Zs	Zh	Z6
Barnes3	.171	.066	.005	.065		.432	.312	.116	.317		.840	.827	.666	.846	
(75.1)	.088	.065	.004	.064		.344	.308	.113	.312		.836	.824	.659	.842	
T(5)	.159	.077	.053	.085		.863	.814	.768	.830		.990	.985	.972	.987	
(6.00)	.086	.076	.052	.083		.820	.811	.765	.827		.986	.985	.971	.987	
Laplace(2,1)	.178	.087	.067	.097		.857	.784	.736	.799		.954	.973	.958	.976	
(3.00)	.094	.085	.066	.095		.793	.781	.733	.795		.975	.973	.958	.975	
T(8)	.141	.086	.073	.097		.916	.889	.873	.901		.997	.995	.993	.996	
(1.50)	.090	.084	.071	.095		.891	.887	.871	.899		.995	.995	.993	.996	

n = 40 (continued)

Distribution	$k = 4.0$				$k = 5.0$				$k = 6.0$			
(kurtosis)	χ^2	χ_r^2	Zs	Zh	Z6	χ^2	χ_r^2	Zs	Zh	Z6	χ^2	χ_r^2	Zs	Zh	Z6
Barnes3	.994	.994	.871	.995		1.00	1.00	.894	1.00		1.00	1.00	.907	1.00	
(75.1)	.994	.993	.867	.995		1.00	1.00	.891	1.00		1.00	1.00	.903	1.00	
T(5)	.999	.999	.992	.999		1.00	1.00	.995	1.00		1.00	1.00	.996	1.00	
(6.00)	.999	.999	.992	.999		1.00	1.00	.995	1.00		1.00	1.00	.996	1.00	
Laplace(2,1)	.998	.997	.992	.997		1.00	1.00	.997	1.00		1.00	1.00	.999	1.00	
(3.00)	.997	.997	.992	.997		1.00	1.00	.997	1.00		1.00	1.00	.999	1.00	
T(8)	1.00	1.00	.999	1.00		1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	
(1.50)	1.00	1.00	.999	1.00		1.00	1.00	.999	1.00		1.00	1.00	1.00	1.00	

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs, Zh, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 8 (continued) Traditional Power Comparisons for Heavy-tail Upper-Tailed Rejection Region when $\sigma_x = k\sigma_0^2$ and significance level 0.100, n = 20

Distribution	$k = 1.0$				$k = 2.0$				$k = 3.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.132	.063	.004	.062	.287	.225	.062	.230	.596	.581	.425	.609
(75.1)	.078	.062	.004	.059	.238	.220	.058	.223	.588	.572	.410	.597
T(5)	.143	.080	.050	.091	.678	.607	.519	.634	.913	.885	.823	.898
(6.00)	.086	.077	.047	.087	.614	.600	.508	.626	.888	.882	.815	.894
Laplace(2,1)	.164	.090	.061	.102	.679	.584	.482	.609	.895	.852	.769	.864
(3.00)	.096	.086	.058	.098	.594	.577	.471	.600	.857	.848	.759	.860
T(8)	.134	.087	.068	.102	.741	.690	.636	.717	.945	.929	.899	.938
(1.50)	.090	.084	.065	.098	.692	.682	.627	.710	.931	.927	.894	.936

n = 20 (continued)

Distribution	$k = 4.0$				$k = 5.0$				$k = 6.0$			
(kurtosis)	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6	χ^2, χ_r^2	Zs	Zh	Z6
Barnes3	.912	.908	.779	.925	.985	.983	.869	.987	.997	.996	.812	.997
(75.1)	.910	.904	.768	.921	.984	.983	.862	.987	.996	.996	.886	.997
T(5)	.976	.967	.927	.972	.993	.990	.963	.991	.998	.997	.976	.997
(6.00)	.968	.966	.922	.970	.990	.989	.959	.991	.997	.996	.973	.997
Laplace(2,1)	.964	.947	.888	.952	.986	.979	.940	.982	.994	.991	.963	.992
(3.00)	.949	.945	.881	.950	.980	.978	.935	.981	.992	.991	.959	.992
T(8)	.988	.984	.967	.986	.997	.996	.986	.996	.999	.999	.992	.999
(1.50)	.983	.983	.965	.985	.996	.995	.984	.996	.999	.999	.991	.999

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs, Zh, and Z6 test using z_α, and $(z_\alpha + t_{\alpha,n-1})/2$ critical points (first, and second numbers in column Zs, Zh, and Z6) and chi-square and robust chi-square test (first and second) on the column χ^2, χ_r^2.

Table 9. Comparisons of Type I Error Rates among Zs & Z6 when n=30 Skewed Distributions

Distribution	α=0.10		α=0.05		α=0.02		α=0.01	
(skewness)	Zs	Z6	Zs	Z6	Zs	Z6	Zs	Z6
IG(1.0,0.1)	.0805	.0792	.0549	.0454	.0378	.0224	.0301	.0138
(9.49)	.0792	.0775	.0534	.0435	.0361	.0206	.0286	.0121
	.0779	.0759	.0518	.0416	.0348	.0189	.0273	.0108
Weibull(1,0.5)	.0802	.0804	.0517	.0437	.0305	.0184	.0234	.0110
(6.62)	.0788	.0786	.0500	.0416	.0288	.0168	.0219	.0095
	.0775	.0769	.0484	.0396	.0273	.0150	.0204	.0082
LN(0,1)	.0722	.0729	.0447	.0381	.0256	.0158	.0197	.0091
(6.18)	.0706	.0710	.0431	.0361	.0243	.0145	.0181	.0078
	.0693	.0693	.0415	.0342	.0231	.0132	.0166	.0069
IG(1.0,0.25)	.0833	.0835	.0512	.0432	.0324	.0198	.0231	.0104
(6.00)	.0818	.0816	.0494	.0409	.0305	.0181	.0214	.0091
	.0802	.0797	.0478	.0388	.0290	.0164	.0198	.0079
Gamma(1,.15)	.0877	.0890	.0538	.0472	.0298	.0200	.0212	.0110
(5.16)	.0856	.0863	.0517	.0448	.0280	.0179	.0196	.0098
	.0837	.0840	.0499	.0427	.0265	.0161	.0178	.0085
IG(1.0,0.5)	.0828	.0864	.0503	.0447	.0264	.0175	.0182	.0101
(4.24)	.0811	.0833	.0481	.0421	.0245	.0158	.0165	.0090
	.0803	.0814	.0464	.0397	.0227	.0141	.0149	.0080
Chi(1)	.0886	.0942	.0490	.0477	.0241	.0214	.0155	.0128
(2.83)	.0864	.0915	.0468	.0453	.0221	.0193	.0138	.0115
	.0843	.0890	.0448	.0430	.0203	.0176	.0126	.0102
Exp(1.0)	.0880	.0970	.0463	.0486	.0229	.0226	.0145	.0141
(2.00)	.0857	.0944	.0441	.0459	.0210	.0206	.0129	.0125
	.0835	.0918	.0420	.0437	.0195	.0189	.0115	.0110
Chi(2)	.0894	.0978	.0472	.0494	.0233	.0230	.0146	.0140
(2.00)	.0872	.0951	.0450	.0468	.0214	.0210	.0128	.0123
	.0848	.0930	.0428	.0446	.0196	.0191	.0116	.0111
Barnes2	.0933	.1048	.0518	.0570	.0265	.0284	.0169	.0181
(1.75)	.0912	.1022	.0497	.0546	.0245	.0262	.0151	.0161
	.0891	.0995	.0474	.0520	.0225	.0240	.0136	.0145
IG(1.0,25.0)	.0865	.1021	.0441	.0505	.0204	.0216	.0109	.0112
(0.60)	.0841	.0990	.0418	.0478	.0185	.0198	.0098	.0094
	.0816	.0963	.0398	.0452	.0168	.0178	.0087	.0081
Chi(24)	.0868	.1017	.0420	.0483	.0187	.0202	.0110	.0109
(0.58)	.0845	.0990	.0399	.0456	.0169	.0180	.0097	.0094
	.0821	.0963	.0377	.0433	.0153	.0163	.0086	.0079

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs and Z6).

Table 9 (continued). Comparisons of Type I Error Rates among Zs & Z6 when n=30 Heavy-tailed Distributions

Distribution	α=0.10		α=0.05		α=0.02		α=0.01	
(skewness)	Zs	Z6	Zs	Z6	Zs	Z6	Zs	Z6
Barnes3	.0644	.0631	.0390	.0303	.0261	.0135	.0196	.0067
(75.1)	.0630	.0613	.0379	.0286	.0249	.0121	.0186	.0056
	.0615	.0596	.0367	.0270	.0238	.0108	.0175	.0047
T(5)	.0795	.0887	.0385	.0388	.0170	.0144	.0103	.0075
(6.00)	.0775	.0861	.0365	.0365	.0157	.0128	.0088	.0065
	.0754	.0835	.0347	.0342	.0143	.0113	.0077	.0054
Barnes1	.1014	.1096	.0517	.0507	.0234	.0191	.0146	.0107
(6.00)	.0988	.1066	.0490	.0477	.0215	.0169	.0128	.0091
	.0965	.1035	.0468	.0448	.0197	.0151	.0113	.0076
T(6)	.0823	.0932	.0407	.0431	.0180	.0170	.0102	.0088
(3.00)	.0799	.0903	.0385	.0404	.0163	.0151	.0089	.0075
	.0777	.0875	.0365	.0381	.0148	.0134	.0078	.0062
Laplace(2,1)	.0879	.0911	.0444	.0474	.0203	.0199	.0124	.0113
(3.00)	.0857	.0893	.0423	.0448	.0186	.0179	.0108	.0098
	.0836	.0879	.0401	.0422	.0170	.0161	.0096	.0084
JTB(4.0,1.0)	.0894	.1045	.0455	.0516	.0203	.0212	.0117	.0114
(0.78)	.0872	.1008	.0431	.0490	.0185	.0193	.0103	.0099
	.0851	.0979	.0409	.0461	.0168	.0172	.0092	.0083
T(16)	.0882	.1035	.0441	.0504	.0195	.0205	.0112	.0107
(0.50)	.0859	.1007	.0417	.0476	.0179	.0184	.0099	.0092
	.0836	.0977	.0397	.0450	.0160	.0165	.0087	.0078
JTB(1.25,0.5)	.0895	.1059	.0441	.0518	.0190	.0203	.0116	.0114
(0.24)	.0856	.1017	.0419	.0486	.0172	.0183	.0100	.0098
	.0827	.0988	.0398	.0459	.0156	.0163	.0087	.0081
T(32)	.0884	.1049	.0436	.0501	.0186	.0196	.0107	.0103
(0.21)	.0859	.1019	.0415	.0476	.0169	.0175	.0093	.0086
	.0834	.0992	.0391	.0452	.0151	.0157	.0083	.0074
JTB(2.0,0.5)	.0769	.0943	.0350	.0408	.0131	.0131	.0067	.0055
(-0.30)	.0743	.0903	.0327	.0382	.0117	.0113	.0059	.0044
	.0705	.0868	.0306	.0355	.0105	.0098	.0049	.0034

NOTE: Entries are the estimated proportion of samples rejected in 100,000 simulated samples for Zs and Z6 test using z_α, $(z_\alpha + t_{\alpha,n-1})/2$, and $t_{\alpha,n-1}$ critical points (first, second, and third numbers in column Zs and Z6).

Conclusion

This study proposed a new right-tailed test of the variance of non-normal distributions. The test is adapted from Hall's inverse Edgeworth expansion for variance (1992) with the purpose to find a new test with fewer restrictions from assumptions and no need for the knowledge of the distribution type. To this end, the study compared Type I error rates and power of previously known tests to its own.

Of the previous tests and six new tests examined by the study, Z6 had the best performance for right-tailed tests. The Z6 test outperforms the χ^2 test by far while performing much better than the χ_r^2 test on skewed distributions and better with heavy-tailed distributions. The Z6 test does not need the original assumptions for the Zs test that the coefficient of skewness of the parent distribution is greater than $\sqrt{2}$ or that the distribution is skewed.

Additionally, the Z6 test performs better overall than the Zs test since Zs performs poorly with smaller alpha levels. Test Z6, unlike Zh, does not need the original assumptions that the population coefficient of skewness is zero in the heavy-tailed distribution or that the distribution is heavy-tailed. Also, the Z6 test performs better for skewed distributions than the Zh test, which has low power at lower alphas. Finally, when considering the Type I error rates, both distribution types, and power, the Z6 test is the best in performance overall. The Z6 test can be used for both types of distributions with good power performance and superior Type I error rates. Therefore, the Z6 test is a good choice for right-tailed tests of variance with non-normal distributions

References

Efron, B., & Tibshirani, R. J. (*1993*). *An Introduction to the bootstrap*. New York: Chapman & Hall.

Kendall, S. M. (*1994*). *Distribution theory*. New York: Oxford University Press Inc.

Lee, S. J., & Sa, P. (*1998*). Testing the variance of skewed distributions. *Communications in Statistics: Simulation and Computation, 27*(3), 807-822.

Lee, S. J., & Sa, P. (*1996*). Testing the variance of symmetric heavy-tailed distributions. *Journal of Statistical Computation and Simulation, 56*, 39-52.

Bickel, P. J., & Doksum, K. A. (*1977*). *Mathematical statistics -Basic idea and selected topics*. San Francisco: Holden-Day.

Hall, P. (*1983*). Inverting an edgeworth expansion. *The Annals of Statistics, 11*, 2, 569-576.

Kendall, S. M., & Stuart, A. (*1969*). *The advanced theory of statistics-distribution theory*, Vol.I, 4^{th} Edition, Griffith Inc.

Hall, P. (*1992*). *The bootstrap and edgeworth*. New York: Springer-Verlag.

Evans, M., Hastings, N., & Peacock, B. (*2000*). *Statistical distributions*, 3rd Edition, New York: John Wiley & Sons, Inc..

Chhikara, R. S., & Folks, J. L. (*1989*). *The inverse gaussian distribution-theory, methodology and applications*. Marcel Dekker, Inc.

Fleishman, A. I. (*1978*). A method of simulating non-normal distributions, *Psychometrica, 43*, 521-531

Johnson, M. E., Tietjen, G. L., & Beckman, R. J. (*1980*). A new family of probability distributions with applications to monte carlo studies. *Journal of the American Statistical Association, 75*, 370, 276-279

Long, M., & Sa, P., The Simulation Results for Right-tailed Testing of Variance for Non-Normal Distributions, Technical Report, #030503, Department of Mathematics and Statistics, University of North Florida, http://www.unf.edu/coas/math-stat/CRCS/CRTechRep.htm, *2003*.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 214-226

1538 – 9472/05/$95.00

Using Scale Mixtures Of Normals To Model Continuously Compounded Returns

Hasan Hamdan
Department of Mathematics and Statistics
James Madison University

John Nolan
Department of Mathematics and Statistics
American University

Melanie Wilson
Department of Mathematics
Allegheny College

Kristen Dardia
Department of Mathematics and Statistics
James Madison University

A new method for estimating the parameters of scale mixtures of normals (SMN) is introduced and evaluated. The new method is called UNMIX and is based on minimizing the weighted square distance between exact values of the density of the scale mixture and estimated values using kernel smoothing techniques over a specified grid of x-values and a grid of potential scale values. Applications of the method are made in modeling the continuously compounded return, CCR, of stock prices. Modeling this ratio with UNMIX proves promising in comparison with other existing techniques that use only one normal component, or those that use more than one component based on the EM algorithm as the method of estimation.

Key words: Expectation-Maximization algorithm, UNMIX, kernel density smoothing, expected return

Introduction

The study of univariate scale mixtures of normals, SMN, has long been of interest to statisticians continuously hunting for better methods to model probability density functions. Modeling using these mixtures has many applications from genetics and medicine to economic and populations studies. More specifically, one can use SMN to model any data that is seemingly normally distributed and has a high kurtosis. Using SMN allows for the tails of the density to be heavier than those in the normal density, giving a better coverage for data that varies greatly from the mean.

The most common estimation of the parameters of the mixtures is the EM algorithm of by Dempster, Larid, and Rubin (1977). This method is based on finding the maximum likelihood estimate of the parameters of a given data set. The EM algorithm performs well in cases where the distance between means of the components is relatively large. However, when estimating the parameters of a mixture of normals where all of the components have the same mean but different variances, the EM algorithm gives a poor estimation when these variances are small and close.

In this article, we elaborate on a new approach of estimation, UNMIX, proposed by Hamdan and Nolan (2004). The UNMIX program uses kernel smoothing techniques to get an empirical estimate of the density of the data. It then estimates the parameters of the mixture based on minimizing the weighted least squares of the distance between the values from the empirical density and the new scale mixture density over a pre-specified grid of x-values, and

Hasan Hamdan is Assistant Professor at James Madison University. His research interests are in mixture models, sampling and mathematical statistics. Email: hamdanhx@jmu.edu. John Nolan is a Professor at American University. His research interests are in probability, stochastic processes and mathematical statistics. Melanie Wilson is a graduate student in statistics at Duke and Kristen Dardia is graduating from James Madison University. This research was partially supported by NSF grant number NSF-DMS 0243845.

potential grid of σ values called r-grid. The UNMIX method will be used to estimate the density of the continuously compound return, CCR. The estimation of the density function is pertinent to knowing the probability that the closing stock price will stay within a certain interval during a given time period. The density function was first estimated simply by using a normal curve.

However, Mandelbrot (1963) and Fama (1965) showed that the normal estimation did not model market returns appropriately due to the excess kurtosis and volatility clustering that characterize returns in financial markets. Clark (1973) then tested the use of the lognormal distribution to estimate the density of the stock returns. This is analogous to using the normal distribution to estimate the density of the natural log of the stock returns (also called the continuously compounded return). Following Clark's estimation, Epps and Epps (1976) found that a better estimation is obtained when using a mixture of distributions.

However, their assumption used the transaction volume as the mixing variable thus introducing excess error. Another popular method evolved recently when Zangari (1996), Wilson (1998), and Glasserman, Heidelberger and Shahabuddin (2000), used the multivariate *t* distribution to estimate the stock return. Unfortunately, Glasserman, Heidelberger, and Shahabuddin (2000), pointed out that since most stock returns have equally fat tails, this model frequently comes up short. Additionally, the method involves solving non-linear equations to derive a numerical approximation of an input covariance matrix and requires the consuming and difficult job of inverting marginal distributions.

As proposed by Clark (1973) and Epps and Epps (1976), we look deeper into modeling the CCR (the natural log of the stock returns), we find that modeling the distribution using a simple normal curve should be avoided due to the fact that the CCR of most stock prices are mound shaped but have a high kurtosis (also known as a high volatility). Therefore, these ratios can be modeled using a SMN with mean zero, since the mean of the CCR of the prices is close to zero. A brief explanation of the concept of a random variable X having a density function of the form of a SMN is introduced in Section 2. Next, in Section 3, techniques of estimation of SMN are listed and brief background on the common EM algorithm is also presented. In Section 4, the density of CCR is estimated for different stocks with SMN using the UNMIX program and using a single normal. Also, the density is also estimated using the EM algorithm and the results are compared. Finally, some suggestions for improving this method are made in the conclusion section.

Methodology

A random variable X is a scale mixture of normals or SMN if $X \overset{d}{=} AZ$, where Z ~ N(0,1), $A > 0$, A and Z independent. Here N(0,1) is the standard normal variable with mean 0 and standard deviation 1. Therefore, X has a probability density function

$$f(x) = \int_0^\infty \frac{1}{\sigma}\phi(x/\sigma)\pi(d\sigma), \tag{1}$$

where ϕ is the standard normal density and the mixing measure π is the distribution of A.

An SMN can either be an infinite or a finite mixture, depending upon the mixing measure π. If our mixing measure is discrete and A takes on a finite number of values, say $\sigma_1,...,\sigma_m$ with respective probabilities $\pi_1,...,\pi_m$ then the probability density function can be rewritten as

$$f(x) = \sum_{j=1}^{m} \frac{1}{\sigma_j}\phi(x/\sigma_j)\pi_j \tag{2}$$

A common finite mixture, called the contaminated normal, occurs when A takes on two values, with $\sigma_1 < \sigma_2$ and $\pi_1 > \pi_2$. In this case our density function can be simplified to

$$f(x) = \pi_1\phi(x/\sigma_1) + (1-\pi_1)\phi(x/\sigma_2).$$

Some common examples of infinite SMN are the Generalized t distribution, Exponential power family, and Sub-Gaussian distributions. The following theorem gives the characteristics necessary for a distribution to be SMN with mean zero.

Theorem: (Schoenberg, 1938) Given any random variable X with density $f(x)$, X is a scale mixture of normals if and only if $h(x) = f\left(\sqrt{x}\right)$ is a completely monotone function. See Feller (1971) for definition of completely monotone function. As we have seen above when A takes on a finite number of values, the density of X can be written more simply in the same manner as equation (2). When π is not concentrated at a finite number of points, Hamdan and Nolan (2004) give a constructive method on how to discretize π so that equation (2) is uniformly close to equation (1).

Estimating Scale Mixtures of Normals

In estimating SMN one needs to find the following: number of components, estimated parameters of each component, and estimated weights of each component. We highlight some of the important developments in this area.

This problem of estimating SMN has been the subject of a large diverse body of literature. Dempster, Larid, and Rubin (1977) introduced the EM algorithm for approximating the maximum likelihood estimates. Because other methods have been developed based on the EM algorithm. A robust powerful approach based on minimizing distance estimation is analyzed by Beran (1977) and Donoho and Liu (1988). Zhang (1990) used Fourier methods to derive kernel estimators and provided lower and upper bounds for the optimal rate of convergence. Priebe (1994) developed a nonparametric maximum likelihood technique from related methods of kernel estimation and finite mixtures.

EM algorithm

The EM algorithm developed by Dempster, Larid, and Rubin (1977), is based on finding the maximum likelihood estimate of the components, parameters, and weights of a mixture of normals. It should be noted that though we will only use the EM algorithm for a mixture of normals, it can be generalized for other mixtures. However, differentiation problems become more complicated in the M step of the algorithm for non-normal mixtures. The EM algorithm does not assume that we are dealing with SMN and allows each density function to have a different mean. Therefore, given the data points, $x_1,...,x_n$, from the finite normal mixture of k components

$$\sum_{j=1}^{k} \frac{1}{\sigma_j} \phi\left(\frac{x-\mu_j}{\sigma_j}\right) \pi_j \ ,$$

the data are completed by letting each x_i correspond to a y_i. The new y_i is a vector giving the initial value x_i and also a sequence of values $z_1,...,z_k$ which tells the location of the x value as follows:

$$y_i = (x_i, z_{i1},..., z_{ik})$$

$$\text{where } z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is generated by the jth component;} \\ 0 & \text{otherwise} \end{cases}$$

Therefore the only missing values are the labels, $z_{i1},...,z_{ik}$. Next the maximum likelihood estimate of each y_i is found in the in the Expectation Step of the EM algorithm. An initial guesses for the parameters

$$\hat{\pi}_1,..., \hat{\pi}_k, \hat{\mu}_1,..., \hat{\mu}_k, \hat{\sigma}_1,..., \hat{\sigma}_k$$

are taken. Then an estimate of probability of category membership of the i^{th} observation, conditional on x_i is found based on using the parameter estimate

$$(\hat{\pi}_1,...,\hat{\pi}_k, \hat{\mu}_1,...,\hat{\mu}_k, \hat{\sigma}_1,...,\hat{\sigma}_k).$$

This estimation is noted by

$$\hat{\gamma}_{ji} = \frac{\hat{\pi}_j \frac{1}{\hat{\sigma}_j}\phi\left(x_i, \hat{\mu}_j, \hat{\sigma}_j\right)}{\sum_{j=1}^{k} \hat{\pi}_j \frac{1}{\hat{\sigma}_j}\phi\left(x_i, \hat{\mu}_j, \hat{\sigma}_j\right)}$$

$i = 1,..., n$ and $j = 1,..., k$.

The next step is to compute the weighted means and variances in the Maximization Step of the EM algorithm for mixtures of normals. Then the E and the M steps are iterated until the parameters converge, and the final values are used as the parameters estimates of the mixture of normals. The EM algorithm works well in modeling SMN where the variance of the components are relatively large, but as the variances approach zero, the algorithm shows a poor performance. In general, as shown in many simulation studies, when the components are not well-separated, estimation based on maximum likelihood is poor (Dick & Bowden, 1973).

There are also many practical difficulties in estimating SMN using the EM. Some of these are computationally difficult and intractable. For example, when the MLE of the mixing measure in the finite case is found, a large local maxima might be found that occurs as a consequence of a fitted component having a very small (but nonzero) variance. Moreover, it is not clear how to initialize the estimates, especially when the mixture is a scale mixture. Though, methods have recently been developed by Biernacki, Celeux, & Govaert (2003) in order to find the most efficient initializing conditions. Another key problem in finite mixture models is determining the number of components in the mixture. Several criteria based on the penalized log-likelihood, such as Akaike Information Criterion, AIC, the Bayesian Information Criterion, BIC, and the Information Complexity Criterion introduced by Bozdogan (1993), have been used.

UNMIX

The next approach, UNMIX, uses kernel smoothing techniques to estimate the empirical density of a sample. It then minimizes the weighted square distance between the kernel smoothing estimate and the density computed by discretizing the mixture over a pre-specified grid of x-values and potential grid of sigma values. Given a sample of size n from the mixture, we fix a grid of possible sigma values (called the σ-grid), and possible x values (called the x-grid), $x_1 ... x_k$, where $k \le m$.

In order to obtain an estimate $\hat{f}(x)$ of $f(x)$ for each x in the x-grid, we use kernel smoothing techniques discussed briefly at the end of the section. Our model is

$$y_i = \sum_{j=1}^{m} \frac{1}{\sigma_j}\phi\left(x_i / \sigma_j\right)\pi_j + \varepsilon_{,i},$$

where ε_i are independent with mean 0. That is solved for π_j by minimizing

$S(\pi)$ where $\pi^{\mathrm{T}} = (\pi_1,...,\pi_m)$ and

$$S(\pi) = \sum_{i=1}^{k}\left(w_i\left(y_i - \sum_{j=1}^{m}\phi_{ij}\pi_j \right)\right)^2,$$

with $\phi_{ij} = \frac{1}{\sigma_j}\phi\left(x_i / \sigma_j\right)$ and w_i are weights. We will use w_i =1 throughout. However, if the data are heavy-tailed then one can try different weights until he finds a good fit (in the heavy-tailed case, a good strategy might be weighting the points that are close to the mean of the x-grid less than those that are far from the mean of the x-grid. Next consider the problem as a quadratic programming problem with two constraints: $\sum \pi_j = 1$ and $\pi_j \ge 0$ for all j. Expanding $S(\pi)$:

$$\begin{aligned}
S(\pi) &= \sum_{i=1}^{k}\left[w_i^2 y_i^2 - 2w_i y_i \sum_{j=1}^{m}\phi_{ij}\pi_j + w_i^2\left(\sum_{j=1}^{k}\phi_{ij}\pi_j\right)^2\right]; \\
&= \sum_{i=1}^{k} w_1^2 y_i^2 - 2\sum_{i=1}^{k}\left(w_i y_i \sum_{j=1}^{k}\phi_{ij}\pi_j\right) + \sum_{i=1}^{k}\left(\sum_{j=1}^{k}\phi_{ij}\pi_j\right)^2; \\
&= \sum_{i=1}^{k} w_1^2 y_i^2 - 2\sum_{j=1}^{m}\left(\sum_{j=1}^{k} w_i y_i \phi_{ij}\right)\pi_j \\
&\quad + \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{j=1}^{m}\left(w_j\phi_{ij}\pi_j\right)\left(w_i\phi_{il}\pi_l\right).
\end{aligned}$$

Because $\sum_{i=1}^{k} w_1^2 y_i^2$ is independent of π, it is a constant. Reformulating the problem in a matrix environment, we let g be the $(m\times 1)$ vector defined as

$$g = -\left(\sum_{j=1}^{k} w_i y_i \phi_{i1}, ..., \sum_{j=1}^{k} w_i y_i \phi_{im} \right)^T,$$

and let $\mathbf{H}$ be the $(m \times m)$ matrix defined as

$$\mathbf{H} = \begin{pmatrix} \sum_{i=1}^{k} w_i^2 \phi_{i1}\phi_{i1} & \sum_{i=1}^{k} w_i^2 \phi_{i1}\phi_{i2} & \cdots & \sum_{i=1}^{k} w_i^2 \phi_{i1}\phi_{im} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{k} w_i^2 \phi_{im}\phi_{i1} & \sum_{i=1}^{k} w_i^2 \phi_{im}\phi_{i2} & \cdots & \sum_{i=1}^{k} w_i^2 \phi_{im}\phi_{im} \end{pmatrix}$$

To simplify, let $c = \frac{1}{2}\sum_{i=1}^{k} w_i^2 y_i^2$ be a constant, resulting in the following formula for $S(\pi)$:

$$S(\pi) = 2\left[c + g^T\pi + \frac{1}{2}\pi^T H\pi \right].$$

Therefore π can be found by minimizing $\left[g^T\pi + \frac{1}{2}\pi^T H\pi \right]$ subject to our two quadratic programming constraints $\sum_{j=1}^{m} \pi_j = 1$ and $\pi_j \geq 0$, this latter constraint can be rewritten in matrix form as $\mathbf{A}\pi \geq b$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

of order $(m\times m)$ and $b^T = (0\cdots 0)$ of order $(m\times 1)$. A quadratic programming routine, QPSOLVE which is a Fortran subroutine, is used to solve this problem. UNMIX is a Splus program that takes the sample, x-grid, r-grid, and a vector of weights as the input and calls QPSOLVE. The program's output is a vector of estimated weights over the given r-grid.

In obtaining an estimate for $f(x)$, kernel smoothing techniques were used. One important variable in density estimates using kernel smoothing techniques is the bandwidth. In general, using a large bandwidth over-smoothes the density curve, and small bandwidths can under-smooth the density curve. In essence, the bandwidth controls how wide the kernel function is spread about the point of interest. If there are a large number of values, x_i near x, then the weight of x is relatively large and the estimation of the density at x will also be large.

There are four sources of variability involved when using UNMIX to estimate a SMN. The first is the sampling variability, the second is due to the method of density estimation and bandwidth used. The third variability is the choice of the x-grid and finally, the fourth is the choice of the r-grid.

Controlling sampling variability can be done by increasing the sample size. However, controlling the variability introduced by the method of density estimation requires care and investigation of the sample and bandwidth used. For example, we can weight the observations by using their distance from the center. There is considerable literature on how to pick the most effective bandwidth including articles by Hardle and Marron (1985) and Muller (1985). For the purposes of this article, when using the UNMIX program, the default bandwidth based on the literature given in R-Software is used.

UNMIX performs well for estimating distributions with a high kurtosis but losses accuracy for data that is extremely concentrated about the mean. However, these difficulties can be overcome due to the flexibility of the program in terms of fitting the data. In particular, the r-grid can be changed and the weights interactively in a systematic way until a good fit is found. We have found that the most useful x-grid is evenly distributed and symmetric about the mode, where the distance from the mode on both sides is the absolute maximum of the sample data because the mode is 0 in this case. This allows the x-grid to cover all data points. Also in creating the σ-grid, a

simple guideline is to make it evenly distributed from a point close to zero to a point at least three sample standard deviations away from zero. This again allows for the σ-grid to cover a large percentage of potential sigma values no matter what the original distribution of sigma is. Here, we assume that the values in the x-grid and the σ-grid are the only possible values for each x and σ, therefore it is important to pick them within in the range of the sample.

Results

Estimating the density of stock returns has been important to statisticians and those interested in finance since the stock market opened. Fama (1965) and many others model stock prices based on simple random walk assumption. In other words the actual price of a stock will be a good estimate of its intrinsic value. The standard assumption is that the percentage changes in the stock price in a short period of time are normally distributed with parameters μ, expected return of the stock, and σ which is the volatility of the stock price. The expected return is estimated by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $x_i = \ln\left(\frac{S_i}{S_{i-1}}\right)$, where S_T is the current stock price. Therefore, the 1 period volatility is estimated by

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\ .$$

The continuously compounded return, CCR, can now be estimated as follows with $S_{T-\tau}$ as the stock price τ time units earlier:

$$\ln\left(\frac{S_T}{S_{T-\tau}}\right) \sim \phi\left[\left(\mu - \sigma^2/2\right)\tau, \sigma\sqrt{\tau}\right].$$

Comparing Normal Estimate to UNMIX Estimate

We now estimate and compare the density of the CCR using a single normal curve and a scale mixture of normals. Taking advantage of Yahoo's (an internet search engine) intensive finance sources, three stocks were found whose price quotes showed relatively high volatility: Ciber Inc, ExxonMobil, and Continental Airlines. For each of the stocks, we sampled the weekly closing prices over the past four years, from July 14, 2000 to July 14, 2004. The natural log of the return was taken to find the CCR for each stock.

Modeling with the single normal method described above and the UNMIX program, their performances were compared against the empirical density found using kernel smoothing techniques. The empirical density is then used to estimate the density over an x-grid of 51 equally-spaced points between -4S and 4S, where S is the sample standard deviation. Because the empirical density can be made very close to the true density at any given point, it is considered as the true density in each of the following error calculations which are presented in Table 1, Table 2 and Table 3.

Example 1: In this example, the density of the CCR, of Ciber Inc. stock, is estimated. The normal estimate based on the random walk assumption has a mean of -.00686 and standard deviation of .09041. The estimated SMN was found using the UNMIX program and has 4 components with weight vector of (.52951,.07374,.39415,.00260) and an estimated σ-vector of (.12266,.06048,.03885,.03750). The estimated densities were evaluated on the same x-grid and the results are shown in Figure 1.

The maximum and average error between each estimate and the empirical density can be seen in Table 1. In Figure 2, the three density estimates were found for an x-grid located in the tail of distribution of CCR and it consists 25 equally-spaced points between .2 and .45. Using the normal assumption, the probability of any sample point falling in such range is approximately .012 and approximately .035 when the scale mixture assumption is used. Though this probability is not high, most density estimation techniques do not recover the tails well where the most extreme occurrences can be found. This could be very problematic in finance and risk analysis.

Figure 1: Estimated Density of CCR for Ciber Inc stock.

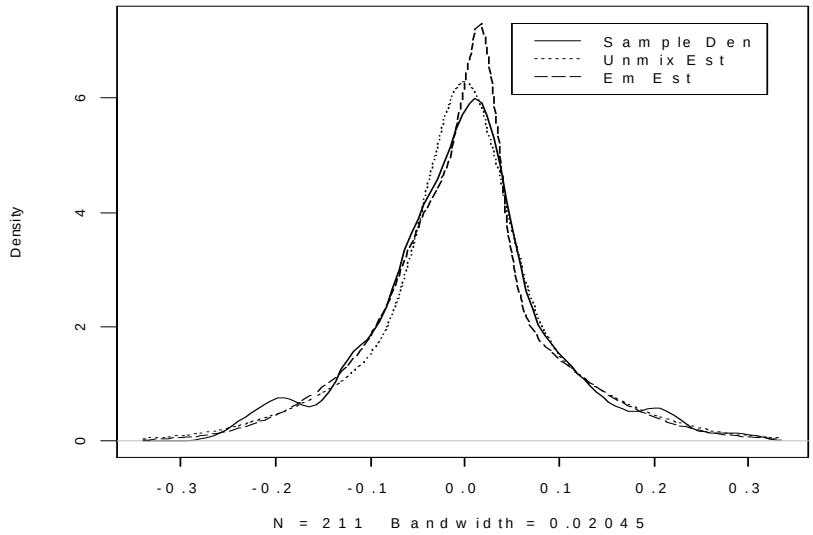
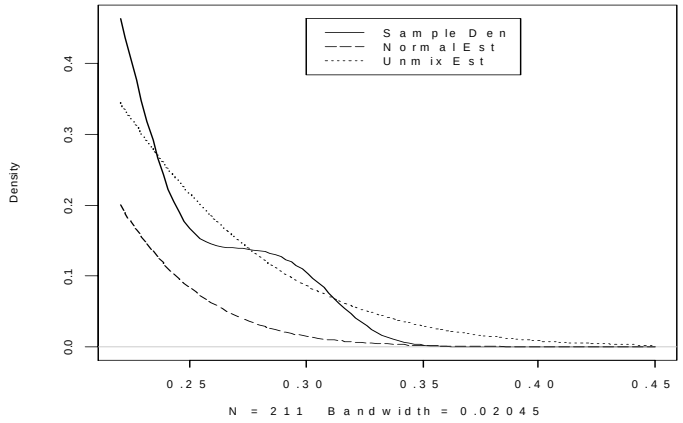


Figure 2: Estimated Density in the right tail of CCR of Ciber Inc.



Notice in Figure 2 that estimating with SMN produces a better fit in the tails. In contrast to overestimating the rate of return in the body, where around 95% of the data are located, the normal curve tends to underestimate the density in the tails. As in our examples, the distributions for the CCR tend to have fatter tails than the proposed normal has. Because the tails of the data are heavy, the scale mixture estimation will produce a better fit than the normal.

Under the single normal assumption, the 95% confidence interval for the mean of the CCR is (-.1767, .1767). Equivalently and by exponentiation, the interval for the mean rate of return is (.8381, 1.1933). The corresponding UNMIX estimate is found to be (.8469, 1.1808). In comparison to UNMIX, the normal curve tends to overestimate the rate of return in the body of the density. Though the gap does not seem large when investing a small amount, for

big time investors 1 penny off per dollar can translate to thousands of dollars lost when investing millions. In Table 4, we summarize the bounds for the middle 95% probability of the distribution for all three examples.

Examples 2 & 3: In these two examples, estimate the densities for the CCR of ExxonMobil and Continental Airline stocks are found based on a single normal and using UNMIX. The single normal and UNMIX estimates are plotted against the empirical density, as described in the previous example, the results are shown in Figure 3 and Figure 4 respectively. The single normal for the ExxonMobil case has an estimated mean of .000248 and an estimated standard σ of 0.038. The scale mixture for the ExxonMobile case has 4 components with a weight vector of (.04734,.47208,.4412,.03938) and the corresponding estimated σ-vector of (.091,.02587,.02542,.01276).

However, the single normal for the Continental Airline case has an estimated mean of -0.0076 and an estimated σ of 0.0930. Finally, the scale mixture for the Continental Airline case has 5 components with a weight vector of (.07006,.01495,.44468,.32486,.14544) and the corresponding estimated σ-vector of (.24215,.24074,.08372,.08232,.02343).

Notice in Figures 3 and 5 the empirical density tends to be negatively skewed. This is common in the densities of CCR since there is a greater probability of the stock market to produce large downward movements than large upward movements. This can be explained by the public's tendency to pull-out of a falling market thus causing prices to drop even further.

In the following tables, the maximum absolute difference between the empirical density and the estimated density over the selected grid using a single normal is indicated by Max. Norm., and the average value is indicated by Avg. Norm. Similarly, Max. UNMIX and Avg. UNMIX are the corresponding values when a scale mixture, with UNMIX as a method of estimation, is used rather than one single normal as a model for CCR.

Next, the performance of the UNMIX method is compared to the EM algorithm in estimating the density of the CCR of the same three stocks. The number of components to be used with the EM is also unknown, and there are many ways that can be used to estimate it. Here, we tried two, three, four and five component mixture.

There was no noticeable difference between the four-component mixture and the five-component mixture. Therefore, the four-component mixture was used for our examples. The parameters were then estimated using the EM algorithm and it was compared to that found using the UNMIX estimation. The initialization of the parameters was somewhat arbitrary because our goal is to find the best density fit and not to investigate the speed or the convergence of these estimation methods. The π's were initialized such that each component has an equal weight of .25, and the μ's were initialized such that = the mean of the sample, μ_1 and μ_2, μ_3 and μ_4 = .2,.4, and .8 times the mean of the sample respectively.

Then, the σ's were initialized for each component in the same manner as the μ's. For each of the three examples, the process was repeated 50 times and the mean of the parameter estimate was taken as the final EM estimate. The estimated densities of the stocks are shown in Figure 7.

Notice that the EM estimate tends to overestimate the mean of the empirical density which is a consequence of the fitted component having a very small variance. The EM captures the skewness of the density better but in general, UNMIX outperforms it. This is seen by the fact that in the three examples, the EM algorithm produces both a greater maximum and average error as summarized in Table 5.

Figure 3: Estimated Density of CCR for Exxon Mobile stock.

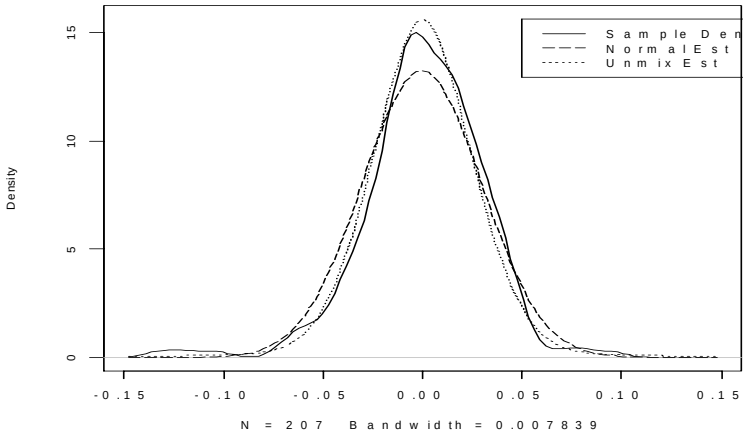


Figure 4: Estimated Density of CCR in the right tail of Exxon Mobile stock. Probability of being in the tail is approximately .0372.

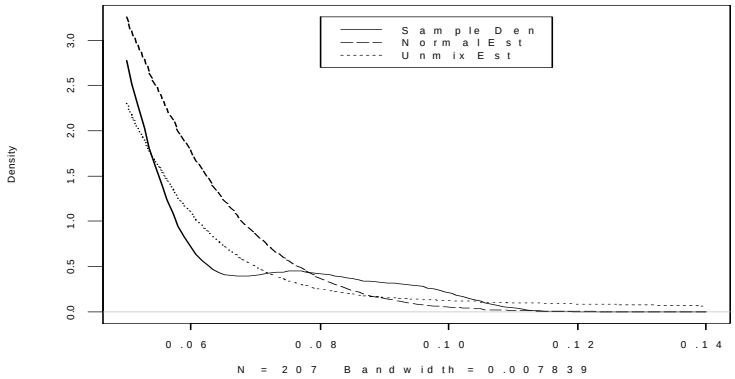


Figure 5: Estimated Density of CCR for Continental Airline.

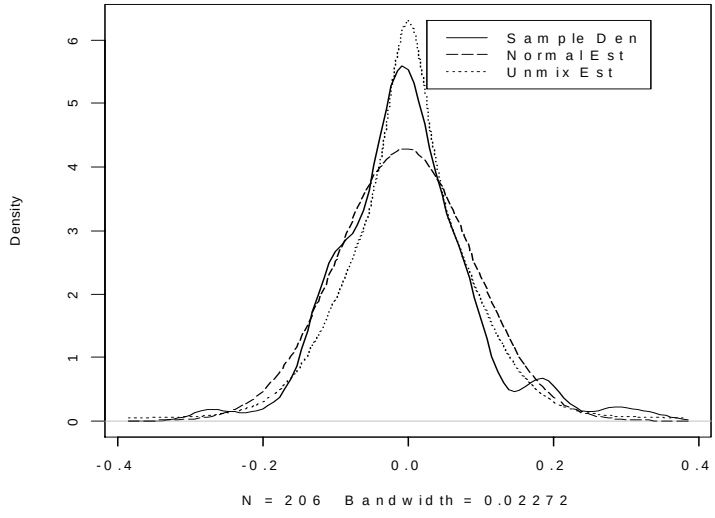


Figure 6: Estimated Density in the right tail of CCR for Continental Airlines stock. Probability of being in this tail is .0186.

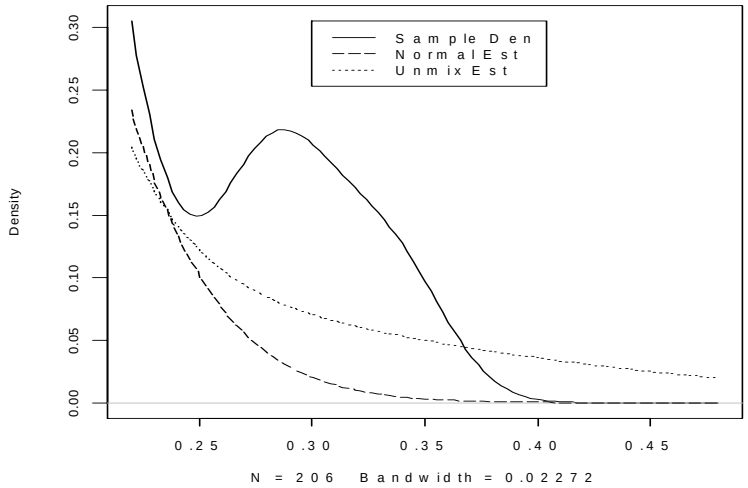


Table 1: Maximum and average errors of the Normal and UNMIX estimates of CCR for Ciber Inc.

Error	Body. Den.	Tail Den.
Max. Norm.	1.6440	.2630
Max. UNMIX	.7874	.1188
Avg. Norm.	.3645	.0499
Avg. UNMIX	.1559	.0203

Table 2: Maximum and average errors of the Normal and UNMIX estimates of CCR for ExxonMobile stock.

Error	Body. Den.	Tail Den.
Max. Norm.	1.9545	1 .0582
Max. UNMIX	1.7341	.4712
Avg. Norm.	.5838	.2268
Avg. UNMIX	.4015	.1283

Table 3: Maximum and average errors of the Normal and UNMIX estimates of CCR for Continental Airline stock**.**

Error	Body. Den.	Tail Den.
Max. Norm.	1.3104	.1897
Max. UNMIX	.8375	.1410
Avg. Norm.	.26911	.0699
Avg. UNMIX	.02090	.0579

Table 4: Bounds for the middle 95% probability of the distribution for the CCR of Ciber Inc., ExxonMobil, and Continental Airlines in both the normal and UNMIX estimates.

Stock	Normal	UNIMIX
Ciber Inc.	(.8381, 1.1933)	(.8469, 1.1808)
ExxonMobile	(.9428,1.0607)	(.9462, 1.0569)
Continental	(.8334, 1.200)	(.8416, 1.1882)

Figure 7: Estimated Density of CCR using the UNMIX program and the EM algorithm for (a) Ciber Inc.; (b) ExxonMobil; (c) Continental Airlines.

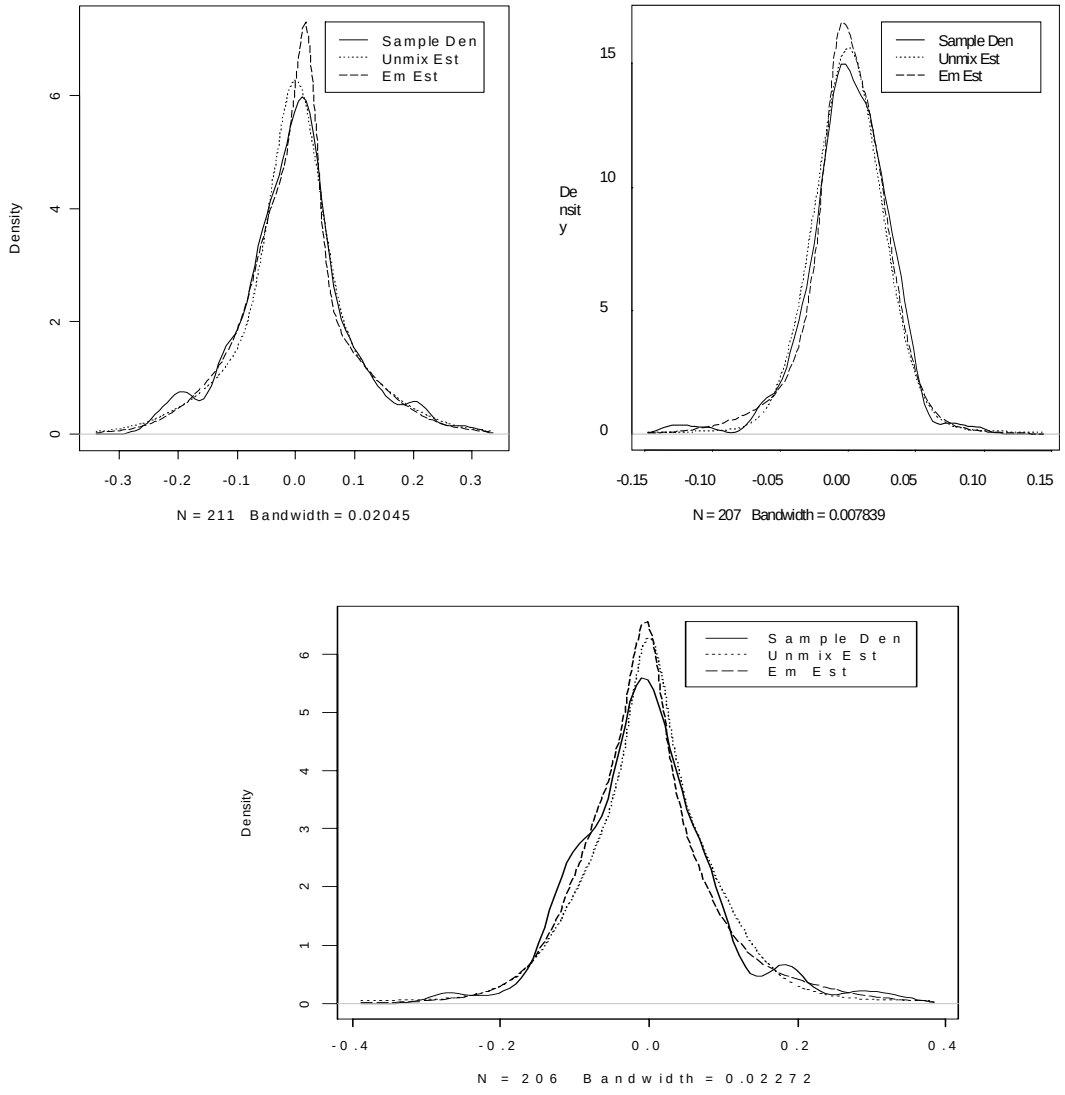


Table 2: Maximum and average errors of the UNMIX and EM estimates of CCR for all examples.

Error	Ciber Inc.	ExxonMobile	Continental
Max. EM	1.2832	2.143	1.2987
Max. UNMIX	.7971	1.7269	.8320
Av. EM	.1592	.4059	.2193
Avg. UNMIX	.1542	.3985	.2048

Conclusion

Estimation of the CCR of stocks has been an interest of both statisticians and financiers due to the importance of producing accurate models for the data. As evidenced by the previous examples, UNMIX allows for this analysis to occur with smaller error in comparison to the single normal assumption and the common methods based on the EM algorithm.

Although the EM algorithm is well developed and allows for different location and different scales, sometimes it has some practical difficulties. For example, when trying to find the MLE of the parameters, it might find a large local maxima that occurs as a consequences of a fitted component having a very small (but non-zero) variance. Also, there are still some problems associated with initializing the parameters including the number of components.

However, UNMIX fitted the data better than the EM. We believe that it will always fit the data well, because it is based on minimizing the weighted distance between empirical density and the mixture over a given grid. However, in terms of estimating the actual parameters, more work needs to be done because the EM still does a better job in estimating the actual values as we have seen in many simulated examples where the actual mixtures are known.

Here are some areas where we can improve UNMIX. First, make it most applicable is the possibility of handling not only scale, but location conditions. Also improvements to the program can be made by developing guidelines to choose the most optimal x-grid and r-grid. Finally, we can improve the empirical density estimate by using optimal kernel functions and bandwidths. Implications of the UNMIX program can apply beyond the scope of the stock market. This program can be used to model distributions with relatively high possibilities of outlying events. Staying in the realm of finance the program can be used to estimate exchange rates.

However, there are also many examples outside of the finance field including fitting extreme data. For example, the UNMIX program was used to fit the density of some heavy-tailed data. These data were generated from the class of stable densities that have infinite variance and known to be infinite variance mixture of normals such as Cauchy density. Although more work needs to be done, but the UNMIX method looks promising in fitting such data.

References

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, *6*, 127-132.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5, 445-463.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, *41*, 561-575.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the Inverse-Fisher Information Matrix, *Information and Classification*, 40-54.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, *41*, 135-155.

Dempster, A. P., Larid, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*, 1-38.

Dick, N. P., & Bowden, D. C. (1973). Maximum likelihood estimation for mixtures of two Normal distributions. *Biometrics, 29*, 781-790.

Donoho, D. L., & Liu, R. C., (1988). The 'Automatic' robustness of minimum distance functional. *Annals of Statistics, 16*, 552-586.

Epps, T. W., & Epps, M. L. (1976). The stochastic dependence of security price changes and transaction volumes: implications for the mixture-of-distributions hypothesis. *Econometrica, 44*, 305-321.

Fama, E. F. (1965). The behavior of stock market prices. *Journal of Business, 38*, 34-105.

Feller, W. J. (1971). An introduction to probability and its applications. (2nd ed.). *Vol. II.* NY: Wiley.

Fix, E., & Hodges, J. L. (1989). Discriminatory data analysis - nonparametric discrimination: consistency properties., *International Statistics Review, 57*, 238-247.

Glasserman, P., Heidelberger, P., & Shahabuddin, P. (2000). Portfolio value-at-risk with heavy-tailed risk factors. *IBM Research Report*.

Hamdan, H., & Nolan, J. (2004). Approximating scale mixtures. *Stochastic Processes and Functional Analysis, 283*, 161-169.

Hamdan, H., & Nolan J (2004, in press). Estimating the parameters of infinite scale mixtures of normals. *Computing Science & Statistics. 36, the proceeding of the 36th Symposium on the Interface.*

Hardle, W., & Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function Estimation. *Annals of Statistics, 13*, 1465-1481.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business, 36*, 394-419.

Muller, H. G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statistical Decisions, 2*, 193-206.

Priebe, E. C. (1994). Adaptive mixtures. *Journal of the American Statistical Association., 89*, 796-806.

Schoenberg, I .J. (1938). Metric spaces and completely monotonic functions. *Annals of Math, 39*, 811-841.

Wilson, T. C. (1998). Value at risk. *Risk Management and Analysis, 1*, 61-124.

Zangari, P. (1996). An improved methodology for measuring VaR. *Risk Metrics Monitor*, 7-25.

Zhang, C. (1990). Fourier methods for estimating mixing densities and distributions. *Annals of Statistics, 18*, 806-831.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 227-239

1538 – 9472/05/$95.00

Enhancing The Performance Of A Short Run Multivariate Control Chart For The Process Mean

Michael B.C. Khoo T. F. Ng
School of Mathematical Sciences
Universiti Sains, Malaysia

Short run production is becoming more important in manufacturing industries as a result of increased emphasis on just-in-time (JIT) techniques, job shop settings and synchronous manufacturing. Short run production or more commonly short run is characterized by an environment where the run of a process is short. To meet these new challenges and requirements, numerous univariate and multivariate control charts for short run have been proposed. In this article, an approach of improving the performance of a short run multivariate chart for individual measurements will be proposed. The new chart is based on a robust estimator of process dispersion.

Key words: Short run, process mean, process dispersion, quality characteristic, in-control, out-of-control

Introduction

Let $\boldsymbol{X}_n = (X_{n1}, X_{n2}, ..., X_{np})'$ denotes the $p \times 1$ vector of quality characteristics made on a part. Assume that $\boldsymbol{X}_n$, n = 1, 2, …, are independent and identically distributed (i.i.d.) multivariate normal, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, observations where X_{nj} is the observation on variable (quality characteristic) j at time n. Define the estimated mean vector obtained from a sequence of $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ random multivariate observations as $\overline{\boldsymbol{X}}_n = (\overline{X}_1, \overline{X}_2, ..., \overline{X}_p)'$ where $\overline{X}_j = \sum_{i=1}^{n} X_{ij} \Big/ n$ is the estimated mean for variable j made from the first n observations. Table 1 gives the additional notations that are required in the article.

Michael B. C. Khoo (Ph.D., University Science of Malaysia, 2001) is a lecturer at the University of Science of Malaysia. His research interests are statistical process control and reliability analysis. Email: mkbc@usm.my. T. F. Ng is a graduate student in the school of Mathematical Sciences, University Science of Malaysia.

The following four cases (see Khoo & Quah, 2002) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ known and unknown give the standard normal V statistics for the short run multivariate chart based on individual measurements: Because V statistics follow a standard normal distribution, this feature makes it suitable for the limits of the chart to be based on the 1-of-1, 3-of-3, 4-of-5 and EWMA tests which will be discussed in the later section.

Case KK: $\boldsymbol{\mu} = \boldsymbol{\mu}_0, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, both known

$$T_n^2 = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

and

$$V_n = \Phi^{-1}\{H_p(T_n^2)\}, n = 1, 2, \ldots \quad (1)$$

Case UK: $\boldsymbol{\mu}$ unknown, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ known

$$T_n^2 = (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})$$

and

$$V_n = \Phi^{-1}\left\{H_p\left[\left(\frac{n-1}{n}\right) T_n^2\right]\right\}, n = 2, 3, \ldots \quad (2)$$

Case KU: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ known, $\boldsymbol{\Sigma}$ unknown

$$T_n^2 = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}_{0,n-1}^{-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

where

$$\boldsymbol{S}_{0,n} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)'$$

and

$$V_n = \Phi^{-1}\left\{F_{p,n-p}\left[\left(\frac{n-p}{p(n-1)}\right)T_n^2\right]\right\}, \quad n = p+1, p+2, \ldots \tag{3}$$

Case UU: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ both unknown

$$T_n^2 = (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})' \boldsymbol{S}_{n-1}^{-1} (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})$$

where

$$\boldsymbol{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)'$$

and

$$V_n = \Phi^{-1}\left\{F_{p,n-p-1}\left[\left(\frac{(n-1)(n-p-1)}{np(n-2)}\right)T_n^2\right]\right\}, \quad n = p+2, p+3, \ldots \tag{4}$$

In Eq. (1) – (4), p represents the number of quality characteristics that are monitored simultaneously, i.e., $p \geq 2$.

Enhanced Short Run Multivariate Control Chart for Individual Measurements

The short run multivariate chart statistics in Eq. (1) and (2) are based on the known covariance matrix while that of Eq. (3) and (4) are based on the estimated covariance matrix, a.k.a., the sample covariance matrix. It is shown in Ref. 1 that the performance of the chart based on the V statistics in Eq. (3) and (4) are inferior to that of cases KK and UK in Eq. (1) and (2) respectively.

Thus, in this article an approach to enhance the performance of the short run multivariate chart for cases KU and UU is proposed by replacing the estimators of the process dispersion, i.e., $\boldsymbol{S}_{0,n}$ and $\boldsymbol{S}_n$ in Eq. (3) and (4) respectively with a robust estimator of scale based on a modified mean square successive difference (MSSD) approach. Holmes and Mergen (1993) and Seber (1984) provided discussion about the MSSD approach. The new estimator of the process dispersion is denoted by $\boldsymbol{S}_{\text{MSSD}}$ while the new V statistic is represented by V_{MSSD}.

Table 1. Notations for Cumulative Distribution Functions.

$\Phi(.)$	- The standard normal cumulative distribution function
$\Phi^{-1}(.)$	- The inverse of the standard normal cumulative distribution function
$H_v(.)$	- The chi-squared cumulative distribution function with v degrees of freedom
$F_{v_1,v_2}(.)$	- The Snedecor-F cumulative distribution function with (v_1, v_2) degrees of freedom

The following formulas give the new standard normal V_{MSSD} statistics for cases KU and UU: Note that all the notations which are used here are similar to that defined in the previous section.

Case KU: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ known, $\boldsymbol{\Sigma}$ unknown

For odd numbered observations, i.e., *n*, is an odd number,

$$T^2_{\text{MSSD},n} = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

where

$$\boldsymbol{S}_{\text{MSSD},n-1} = \frac{1}{2} \sum_{i=2,4,6}^{n-1} (\boldsymbol{X}_i - \boldsymbol{X}_{i-1})(\boldsymbol{X}_i - \boldsymbol{X}_{i-1})'$$

and

$$V_{\text{MSSD},n} = \Phi^{-1}\left\{ F_{p,\frac{1}{2}(n-2p+1)} \left[\frac{n-2p+1}{2p} \right] T^2_{\text{MSSD},n} \right\}, \quad n = 2p + 1, 2p + 3, \ldots \tag{5a}$$

For even numbered observations, i.e., *n*, is an even number,

$$T^2_{\text{MSSD},n} = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-2} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

where

$$\boldsymbol{S}_{\text{MSSD},n-2} = \frac{1}{2} \sum_{i=2,4,6}^{n-2} (\boldsymbol{X}_i - \boldsymbol{X}_{i-1})(\boldsymbol{X}_i - \boldsymbol{X}_{i-1})'$$

and

$$V_{\text{MSSD},n} = \Phi^{-1}\left\{ F_{p,\frac{1}{2}(n-2p)} \left[\frac{n-2p}{2p} \right] T^2_{\text{MSSD},n} \right\}, \quad n = 2p + 2, 2p + 4, \ldots \tag{5b}$$

Case UU: μ and Σ both unknown

For odd numbered observations, i.e., *n*, is an odd number,

$$T^2_{\text{MSSD},n} = (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})$$

where

$$\boldsymbol{S}_{\text{MSSD},n-1} = \frac{1}{2} \sum_{i=2,4,6}^{n-1} (\boldsymbol{X}_i - \boldsymbol{X}_{i-1})(\boldsymbol{X}_i - \boldsymbol{X}_{i-1})'$$

and

$$V_{\text{MSSD},n} = \Phi^{-1}\left\{ F_{p,\frac{1}{2}(n-2p+1)} \left[\frac{(n-2p+1)(n-1)}{2np} \right] T^2_{\text{MSSD},n} \right\}, \quad n = 2p + 1, 2p + 3, \ldots \tag{6a}$$

For even numbered observations, i.e., *n*, is an even number,

$$T^2_{MSSD,n} = (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})' \boldsymbol{S}^{-1}_{\text{MSSD},n-2} (\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1})$$

where

$$\boldsymbol{S}_{\text{MSSD},n-2} = \frac{1}{2} \sum_{i=2,4,6}^{n-2} (\boldsymbol{X}_i - \boldsymbol{X}_{i-1})(\boldsymbol{X}_i - \boldsymbol{X}_{i-1})'$$

and

$$V_{\text{MSSD},n} = \Phi^{-1}\left\{ F_{p,\frac{1}{2}(n-2p)} \left[\frac{(n-2p)(n-1)}{2np} \right] T^2_{\text{MSSD},n} \right\}, \quad n = 2p + 2, 2p + 4, \ldots \tag{6b}$$

For the V_{MSSD} statistics in eqs. (5a), (5b), (6a) and (6b) above, *p* is the number of quality characteristics monitored simultaneously, hence $p \geq 2$.

Tests for Shifts in the Mean Vector μ

Because all the V_{MSSD} statistics are standard normal random variables, the following tests will be used in the detection of shifts in the mean vector. Given a sequence of V_{MSSD} statistics, i.e., $V_{\text{MSSD},a+1}, V_{\text{MSSD},a+2}, ..., V_{\text{MSSD},m}, ...$, where $V_{\text{MSSD},a}$ represents the control chart statistic, V_{MSSD}, at observation *a*, the tests are defined as follow:

The 1-of-1 Test: When $V_{\text{MSSD},m}$ is plotted, the test signals a shift in μ if $V_{\text{MSSD},m} > 3\sigma$, i.e., $V_{\text{MSSD},m} > 3$.

The 3-of-3 Test: When $V_{\text{MSSD},m}$ is plotted, the test signals a shift in μ if $V_{\text{MSSD},m}$, $V_{\text{MSSD},m-1}$ and $V_{\text{MSSD},m-2}$ all exceed 1σ (i.e., 1). This test requires the availability of three consecutive V_{MSSD} statistics.

The 4-of-5 Test: When $V_{\text{MSSD},m}$ is plotted, the test signals a shift in μ if at least four of the five values $V_{\text{MSSD},m}$, $V_{\text{MSSD},m-1}$, …, $V_{\text{MSSD},m-4}$ exceed 1σ (i.e., 1). This test can only be used if five consecutive V_{MSSD} statistics are available.

In addition to these tests, the EWMA chart computed from a sequence of the V_{MSSD} statistics is also considered. The EWMA chart is defined as follows:

$$Z_{\text{MSSD},m} = \alpha V_{\text{MSSD},m} + (1-\alpha) Z_{\text{MSSD},m-1}, \quad m = a, a+1, \ldots \tag{7}$$

where $Z_{\text{MSSD},a-1} = 0$ and a is an integer representing the starting point of the monitoring of a process. The *UCL* of an EWMA chart is $K\sqrt{\alpha/(2-\alpha)}$, where α is the smoothing constant and K is the control limit constant. For the simulation study in this paper, the values of (α, K) used are (0.25, 2.90) which gives *UCL* = 1.096, i.e., similar to that in Ref. 1.

Evaluating the Performance of the Enhanced Short Run Multivariate Chart

A simulation study is performed using SAS version 8 to study the performance of the enhanced short run multivariate chart for individual measurements. To enable a comparison to be made between the performance of the new short run chart with the chart proposed in Ref. 1, the simulation study of the new bivariate chart is conducted under the same condition as that of Ref. 1. The on-target mean vector vector is $\boldsymbol{\mu}_0 = (0, 0)'$ while the in-control covariance matrix is $\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where ρ is the correlation coefficient between the two quality characteristics. For every value of $c \in \{10, 20, 50\}$, c in-control observations are generated from a $N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ distribution followed by 30 additional observations from a $N_2(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_0)$ distribution. The V_{MSSD} statistics for cases KU and UU in Eq. (5a), (5b), (6a) and (6b) are computed as soon as enough values are available to define its statistics for the particular case.

This procedure is repeated 5000 times and the proportion of times an o.o.c. signal is observed from $c + 1$ to $c + 30$ for the first time is recorded. All of the tests defined in the previous section are used in evaluating the performance of the chart. Note that the new chart is also directionally invariant. Thus, the chart's performance is determined solely by the square root of the noncentrality parameter (see Ref. 1). Because of the directionally invariant property of the new short run multivariate chart, only $\boldsymbol{\mu}_S = (\delta,0)'$ based on ρ = 0 and 0.5 are considered in the simulation study.

The results of cases KU and UU for the enhanced short run multivariate chart are given in Tables 2 and 3 for ρ = 0 and 0.5 respectively. Tables 4 and 5 give the corresponding results of the short run multivariate chart proposed in Ref. 1. The results show that the approach incorporating the new estimator of process dispersion, i.e., $\boldsymbol{S}_{\text{MSSD}}$, are superior to that proposed in Ref. 1.

For example, if δ = 1.5, c = 10 and ρ = 0, the probabilities of detecting an o.o.c. for case KU are 0.225, 0.721, 0.681 and 0.739 for the enhanced chart based on the 1-of-1, 3-of-3, 4-of-5 and EWMA tests respectively (see Table 2). From the results in Table 4, the corresponding probabilities that are computed for these four tests are 0.056, 0.253, 0.172 and 0.157 respectively. Clearly, these probabilities are much lower than those of the enhanced chart. Note also that the Type-I error of the enhanced chart based on the 3-of-3, 4-of-5 and EWMA tests are higher than those in Ref. 1. However, from Tables 2 and 3, it is observed that the probabilities of signaling a false o.o.c. for these three tests decrease as the values of c increase. The probabilities of a false alarm for the 1-of-1

Table 2. Simulation Results of the Enhanced Short Run Multivariate Chart for Cases KU and UU based on $\boldsymbol{\mu}_0 = (0,0)'$, $\boldsymbol{\mu}_s = (\delta,0)'$ and $\rho = 0$.

$\rho = 0$		$c = 10$				$c = 20$				$c = 50$			
$\boldsymbol{\mu}_s = (\delta,0)'$		1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA
δ													
0.0	KU	0.039	0.152	0.111	0.116	0.032	0.130	0.086	0.087	0.040	0.113	0.066	0.064
	UU	0.036	0.156	0.111	0.118	0.035	0.123	0.079	0.087	0.037	0.113	0.063	0.069
0.5	KU	0.055	0.220	0.169	0.173	0.049	0.194	0.130	0.140	0.070	0.187	0.121	0.126
	UU	0.040	0.171	0.126	0.133	0.037	0.149	0.100	0.108	0.054	0.158	0.096	0.102
1.0	KU	0.111	0.423	0.360	0.394	0.123	0.422	0.343	0.396	0.167	0.440	0.352	0.420
	UU	0.049	0.221	0.168	0.174	0.063	0.239	0.171	0.189	0.114	0.305	0.228	0.266
1.5	KU	0.225	0.721	0.681	0.739	0.277	0.746	0.703	0.779	0.390	0.790	0.746	0.846
	UU	0.064	0.308	0.247	0.261	0.112	0.396	0.329	0.362	0.240	0.578	0.505	0.594
2.0	KU	0.409	0.919	0.910	0.947	0.510	0.943	0.931	0.970	0.665	0.972	0.968	0.991
	UU	0.091	0.434	0.371	0.387	0.189	0.611	0.550	0.609	0.431	0.841	0.813	0.893
2.5	KU	0.611	0.986	0.986	1.000	0.740	0.994	0.994	0.998	0.882	0.999	0.998	1.000
	UU	0.126	0.574	0.516	0.534	0.293	0.799	0.769	0.815	0.660	0.969	0.968	0.989
3.0	KU	0.787	0.998	0.999	1.000	0.894	1.000	0.999	1.000	0.974	1.000	1.000	1.000
	UU	0.173	0.718	0.678	0.681	0.430	0.927	0.914	0.939	0.849	0.997	0.998	1.000
4.0	KU	0.965	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.292	0.910	0.897	0.897	0.695	0.996	0.995	0.998	0.988	1.000	1.000	1.000
5.0	KU	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.423	0.980	0.981	0.978	0.883	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 3. Simulation Results of the Enhanced Short Run Multivariate Chart for Cases KU and UU based on $\boldsymbol{\mu}_0 = (0,0)'$, $\boldsymbol{\mu}_s = (\delta,0)'$ and $\rho = 0.5$.

$\rho = 0$		$c = 10$				$c = 20$				$c = 50$			
$\boldsymbol{\mu}_s = (\delta,0)'$		1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA
δ													
0.0	KU	0.039	0.152	0.111	0.116	0.032	0.130	0.086	0.087	0.040	0.113	0.066	0.064
	UU	0.036	0.156	0.111	0.118	0.035	0.123	0.079	0.087	0.037	0.113	0.063	0.069
0.5	KU	0.063	0.238	0.189	0.191	0.058	0.214	0.149	0.166	0.078	0.210	0.138	0.157
	UU	0.040	0.179	0.132	0.137	0.040	0.163	0.113	0.118	0.061	0.171	0.108	0.120
1.0	KU	0.141	0.513	0.459	0.499	0.164	0.520	0.447	0.519	0.227	0.546	0.462	0.571
	UU	0.055	0.245	0.191	0.198	0.076	0.281	0.216	0.237	0.148	0.387	0.295	0.359
1.5	KU	0.304	0.826	0.805	0.859	0.382	0.863	0.842	0.901	0.518	0.900	0.885	0.949
	UU	0.078	0.364	0.307	0.317	0.144	0.498	0.429	0.484	0.322	0.709	0.662	0.760
2.0	KU	0.525	0.971	0.968	0.988	0.648	0.988	0.985	0.995	0.821	0.995	0.996	0.999
	UU	0.112	0.522	0.466	0.478	0.255	0.734	0.692	0.744	0.572	0.932	0.923	0.973
2.5	KU	0.750	0.998	0.997	1.000	0.864	0.999	0.999	1.000	0.961	1.000	1.000	1.000
	UU	0.157	0.679	0.639	0.645	0.404	0.900	0.888	0.924	0.810	0.994	0.995	0.999
3.0	KU	0.894	1.000	1.000	1.000	0.965	1.000	1.000	1.000	0.997	1.000	1.000	1.000
	UU	0.217	0.822	0.796	0.793	0.556	0.976	0.974	0.985	0.941	1.000	1.000	1.000
4.0	KU	0.994	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.371	0.962	0.958	0.959	0.826	1.000	1.000	1.000	0.999	1.000	1.000	1.000
5.0	KU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.537	0.996	0.995	0.996	0.959	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4. Simulation Results of the Short Run Multivariate Chart in Ref. 1 for Cases KU and UU based on $\boldsymbol{\mu}_0 = (0,0)'$, $\boldsymbol{\mu}_s = (\delta,0)'$ and $\rho = 0$.

$\rho = 0$		$c = 10$				$c = 20$				$c = 50$			
$\boldsymbol{\mu}_s = (\delta,0)'$		1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA
δ													
0.0	KU	0.041	0.102	0.052	0.038	0.037	0.103	0.046	0.044	0.042	0.103	0.056	0.042
	UU	0.040	0.103	0.052	0.039	0.039	0.100	0.049	0.041	0.038	0.101	0.050	0.043
0.5	KU	0.048	0.120	0.069	0.056	0.049	0.133	0.073	0.066	0.057	0.153	0.088	0.088
	UU	0.041	0.100	0.054	0.040	0.040	0.106	0.053	0.049	0.051	0.131	0.070	0.069
1.0	KU	0.052	0.178	0.110	0.093	0.072	0.233	0.149	0.151	0.113	0.312	0.221	0.263
	UU	0.043	0.112	0.062	0.051	0.052	0.143	0.084	0.080	0.087	0.225	0.154	0.167
1.5	KU	0.056	0.253	0.172	0.157	0.093	0.387	0.286	0.321	0.184	0.581	0.493	0.617
	UU	0.041	0.128	0.074	0.065	0.067	0.216	0.141	0.148	0.144	0.417	0.320	0.393
2.0	KU	0.069	0.340	0.248	0.247	0.132	0.558	0.469	0.536	0.292	0.821	0.785	0.903
	UU	0.049	0.164	0.104	0.091	0.096	0.329	0.241	0.270	0.233	0.652	0.585	0.713
2.5	KU	0.096	0.434	0.337	0.342	0.193	0.713	0.650	0.741	0.445	0.949	0.943	0.991
	UU	0.064	0.215	0.145	0.133	0.151	0.468	0.381	0.428	0.368	0.841	0.809	0.921
3.0	KU	0.131	0.522	0.425	0.442	0.290	0.833	0.789	0.882	0.617	0.991	0.991	1.000
	UU	0.096	0.269	0.184	0.181	0.232	0.611	0.528	0.603	0.539	0.947	0.942	0.991
4.0	KU	0.268	0.663	0.561	0.605	0.569	0.949	0.933	0.984	0.914	1.000	1.000	1.000
	UU	0.194	0.372	0.258	0.292	0.484	0.804	0.733	0.854	0.873	0.996	0.997	1.000
5.0	KU	0.473	0.747	0.652	0.730	0.832	0.984	0.980	0.999	0.996	1.000	1.000	1.000
	UU	0.355	0.448	0.304	0.397	0.769	0.900	0.851	0.957	0.987	1.000	1.000	1.000

Table 5. Simulation Results of the Short Run Multivariate Chart in Ref. 1 for Cases KU and UU based on $\boldsymbol{\mu}_0 = (0,0)'$, $\boldsymbol{\mu}_s = (\delta,0)'$ and $\rho = 0.5$.

$\rho = 0$		$c = 10$				$c = 20$				$c = 50$			
$\boldsymbol{\mu}_s = (\delta,0)'$		1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA	1-of-1	3-of-3	4-of-5	EWMA
δ													
0.0	KU	0.041	0.102	0.052	0.038	0.037	0.103	0.046	0.044	0.042	0.103	0.056	0.042
	UU	0.040	0.103	0.052	0.039	0.039	0.100	0.049	0.041	0.038	0.101	0.050	0.043
0.5	KU	0.047	0.124	0.072	0.059	0.052	0.141	0.082	0.082	0.065	0.166	0.101	0.102
	UU	0.042	0.102	0.055	0.041	0.041	0.115	0.063	0.049	0.054	0.144	0.079	0.079
1.0	KU	0.054	0.196	0.124	0.120	0.077	0.274	0.190	0.201	0.129	0.391	0.295	0.355
	UU	0.042	0.115	0.068	0.050	0.056	0.165	0.098	0.097	0.098	0.281	0.197	0.217
1.5	KU	0.061	0.286	0.202	0.199	0.109	0.465	0.374	0.428	0.234	0.700	0.638	0.789
	UU	0.047	0.139	0.087	0.077	0.079	0.266	0.181	0.199	0.181	0.527	0.440	0.553
2.0	KU	0.085	0.399	0.308	0.305	0.171	0.650	0.588	0.679	0.387	0.916	0.903	0.976
	UU	0.062	0.182	0.119	0.121	0.126	0.416	0.325	0.364	0.314	0.785	0.744	0.870
2.5	KU	0.127	0.501	0.402	0.421	0.269	0.804	0.769	0.857	0.589	0.984	0.985	0.999
	UU	0.091	0.250	0.167	0.173	0.218	0.578	0.495	0.564	0.508	0.927	0.922	0.983
3.0	KU	0.187	0.590	0.490	0.527	0.418	0.900	0.884	0.951	0.789	0.998	0.998	1.000
	UU	0.139	0.317	0.217	0.229	0.341	0.717	0.645	0.733	0.719	0.979	0.983	0.999
4.0	KU	0.394	0.724	0.626	0.686	0.751	0.977	0.970	0.996	0.981	1.000	1.000	1.000
	UU	0.293	0.424	0.288	0.354	0.678	0.883	0.831	0.935	0.965	1.000	1.000	1.000
5.0	KU	0.653	0.801	0.700	0.802	0.944	0.995	0.993	1.000	1.000	1.000	1.000	1.000
	UU	0.518	0.489	0.325	0.473	0.909	0.949	0.911	0.989	0.999	1.000	1.000	1.000

Table 6. V_{MSSD} and V Statistics for Case UU.

Observation No., n	X_1	X_2	V_n	$V_{MSSD,n}$	Observation No., n	X_1	X_2	V_n	$V_{MSSD,n}$
1	1.404	0.268	-	-	21	0.819	-0.277	0.395	0.580
2	0.624	1.392	-	-	22	1.706	0.564	0.780	1.085
3	0.454	0.755	-	-	23	1.198	-1.313	2.181	2.434
4	-1.768	-1.902	1.162	-	24	2.863	0.211	2.049	2.737
5	-0.224	0.140	-1.452	-1.650	25	2.141	0.438	0.545	1.657
6	-0.082	0.734	-0.585	-1.214	26	1.823	0.474	-0.023	0.987
7	1.146	0.484	-0.190	-0.327	27	1.609	0.414	-0.366	0.630
8	1.816	0.906	0.222	0.058	28	2.811	2.192	1.191	1.650
9	-1.245	-1.555	0.482	0.296	29	0.170	-0.650	-0.987	-0.676
10	-0.976	-0.340	-0.199	0.023	30	-0.776	-1.186	-0.193	0.347
11	-0.621	-1.058	-0.266	-0.393	31	-0.111	-0.613	-1.216	-0.838
12	-0.080	-0.710	-0.507	-0.800	32	1.400	0.302	-0.656	0.313
13	0.742	-0.146	-0.202	0.042	33	1.584	0.337	-0.403	0.609
14	-0.543	-0.818	-0.824	-0.654	34	2.047	0.585	0.080	1.203
15	-2.335	-2.801	1.437	1.507	35	0.481	0.690	-0.153	0.667
16	-0.848	-1.176	-0.808	-0.415	36	3.773	2.495	1.693	2.545
17	-0.431	0.590	0.836	0.742	37	1.891	1.871	0.673	1.256
18	1.369	1.863	0.769	0.955	38	2.169	1.073	-0.160	0.420
19	0.283	0.197	-1.659	-1.405	39	1.761	1.191	-0.400	0.049
20	0.850	0.149	-0.155	0.028	40	1.184	-0.113	-0.531	0.132

Figure 1. Plotted V_{MSSD} Statistics for Case UU

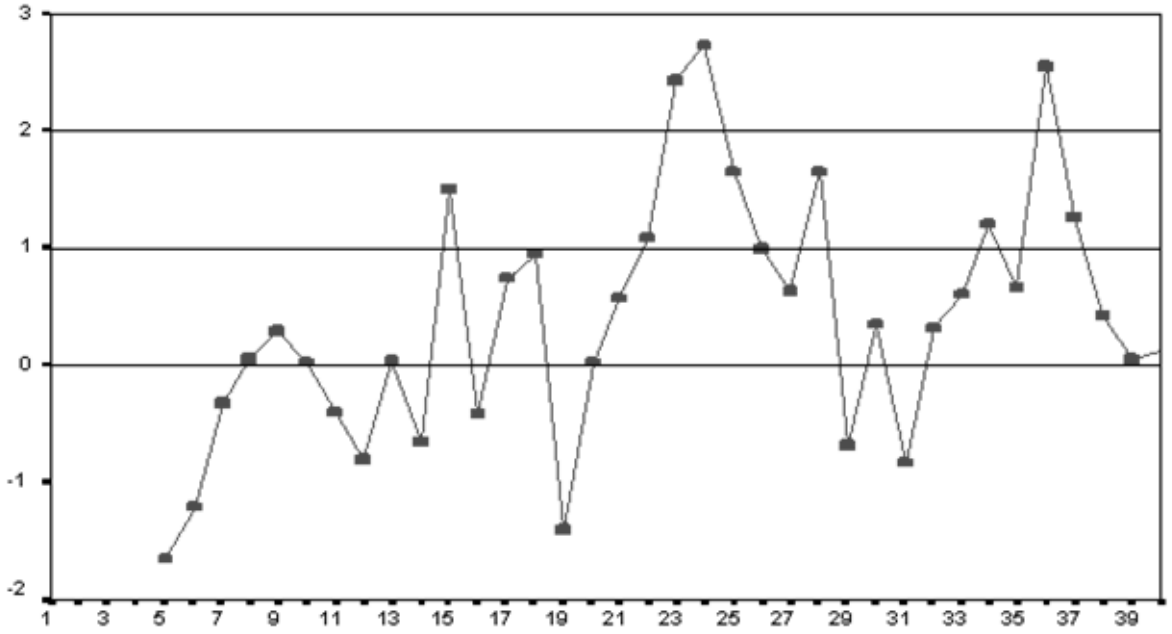
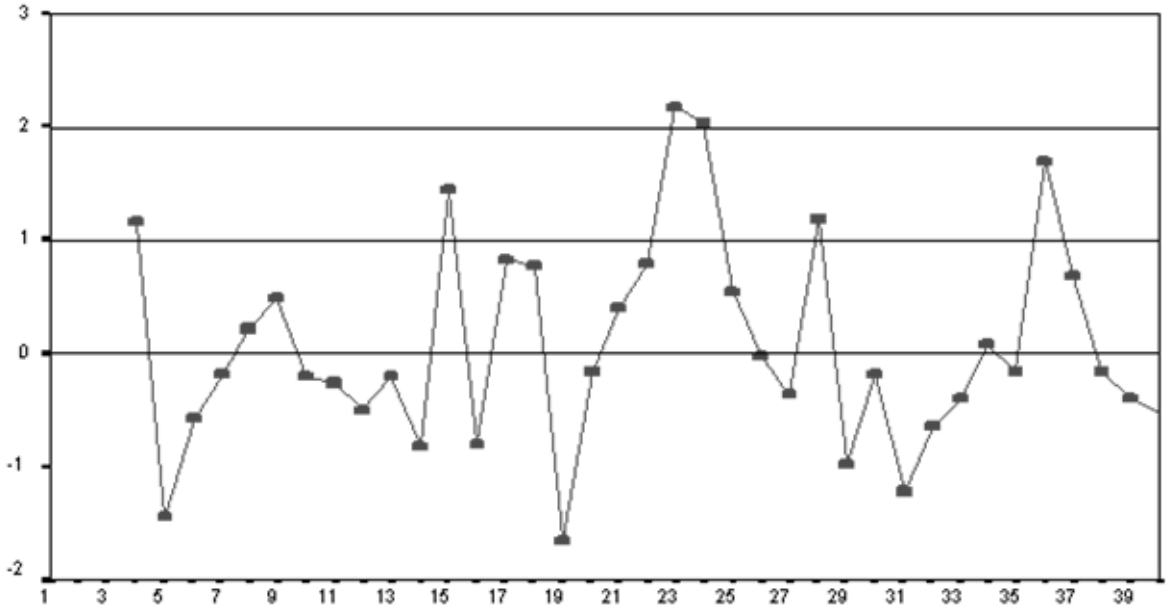


Figure 2. Plotted *V* Statistics for Case UU



test in Tables 2, 3, 4 and 5 are almost the same. The results also show that the performance of the enhanced chart based on the basic 1-of-1 rule is superior to the chart proposed in Ref. 1.

An Example of Application

An example will be given to show how the proposed enhanced short run multivariate chart is put to work. To simulate an in-control process, 20 bivariate observations are generated using SAS version 8 from a $N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ distribution. For an o.o.c. process, with a shift in the mean vector, the next 20 bivariate observations are generated from a $N_2(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_0)$ distribution. Here, $\boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{\mu}_S = \begin{pmatrix} 1.3 \\ 0 \end{pmatrix}$, $\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where ρ = 0.8. The 40 observations generated are substituted in eqs. (6a) and (6b) to compute the corresponding V_{MSSD} statistics for case UU. Similarly, these 40 observations are substituted in Eq. (4) to compute the corresponding *V* statistics for case UU. The computed *V* and V_{MSSD} statistics are summarized in Table 6. Figures 1 and 2 show the plotted V_{MSSD} and *V* statistics respectively. For the enhanced chart based on the V_{MSSD} statistics, the 3-of-3 test signals an o.o.c. at observation 24 while the 4-of-5 test signals at observation 25. The chart proposed in Ref. 1 based on the *V* statistics fails to detect a shift in the mean vector.

Conclusion

It is shown in this paper that the enhanced chart based on a robust estimator of scale, i.e., $\boldsymbol{S}_{\text{MSSD}}$ gives excellent improvement over the existing short run multivariate chart proposed in Khoo & Quah (2002). The proofs of how the V_{MSSD} statistics for cases KU and UU are derived are shown in the Appendix.

References

Khoo, M. B. C. & Quah, S. H. (2002). Proposed short runs multivariate control charts for the process mean. *Quality Engineering*, *14* (4), 603 – 621.

Holmes, D. S., & Mergen, A. E. (1993). Improving the performance of the T^2 control chart. *Quality Engineering*, *5* (4), 619 – 625.

Seber, G. A. F. (1984). *Multivariate observations*. New York : John Wiley and Sons.

Appendix

In this section, it will be shown that the V_{MSSD} statistics in eqs. (5a), (5b), (6a) and (6b) are $N(0,1)$ random variables. All the notations used here are already defined in the earlier sections. The following theorems taken from Seber (1984) are used:

Theorem A. Suppose that $y \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, $W \sim W_p(n, \mathbf{\Sigma})$, and y and W are statistically independent. Assumed that the distribution are nonsingular, i.e., $\mathbf{\Sigma} > \mathbf{O}$, and $n \geq p$, so that $\boldsymbol{W}^{-1}$ exists with probability 1.

Let

$$T^2 = n\boldsymbol{y}'\boldsymbol{W}^{-1}\boldsymbol{y}', \tag{A1}$$

then

$$\frac{(n-p+1)}{p}\frac{T^2}{n} \sim F_{p,n-p+1} \tag{A2}$$

Theorem B. Suppose that $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ are independently and identically distributed (i.i.d.) as $N_p(\mathbf{0}, \mathbf{\Sigma})$, then

$$\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \sim W_p(n, \mathbf{\Sigma}) \tag{A3}$$

where $W_p(n, \mathbf{\Sigma})$ is the Wishart distribution with n degrees of freedom.

Equation (5a): Case KU

We need to show that for odd numbered observations, i.e., when n is an odd number,

$$T^2_{\text{MSSD},n} = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

$$\sim \frac{2p}{n-2p+1} F_{p,\frac{1}{2}(n-2p+1)}$$

Proof:

If $\boldsymbol{X}_j$, j = 1, 2, 3, …, are i.i.d. $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$

variables, then

$$\boldsymbol{X}_i - \boldsymbol{X}_{i-1} \sim N_p(\mathbf{0}, 2\mathbf{\Sigma}),\ i = 2, 4, 6, \ldots$$

and

$$\frac{1}{\sqrt{2}}(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}) \sim N_p(\mathbf{0}, \mathbf{\Sigma}),\ i = 2, 4, 6, \ldots .$$

Thus, from eq. (A3) of Theorem B,

$$\frac{1}{2}\sum_{i=2,4,6}^{n-1} (\boldsymbol{X}_i - \boldsymbol{X}_{i-1})(\boldsymbol{X}_i - \boldsymbol{X}_{i-1})' \sim W_p\left(\frac{n-1}{2}, \mathbf{\Sigma}\right),$$

i.e.,

$$\boldsymbol{S}_{\text{MSSD},n-1} \sim W_p\left(\frac{n-1}{2}, \mathbf{\Sigma}\right). \tag{A4}$$

Because $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is known, then

$$\boldsymbol{X}_n - \boldsymbol{\mu}_0 \sim N_p(\mathbf{0}, \mathbf{\Sigma}) \tag{A5}$$

Substituting Eq. (A4) and (A5) into Eq. (A1) and (A2) of Theorem A,

$$\frac{\left(\frac{n-1}{2}-p+1\right)\left(\frac{n-1}{2}\right)}{p\left(\frac{n-1}{2}\right)} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

$$\sim F_{p,\frac{n-1}{2}-p+1}$$

i.e.,

$$\frac{(n-2p+1)}{2p} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)$$

$$\sim F_{p,\frac{1}{2}(n-2p+1)}.$$

Define

$$T^2_{\text{MSSD},n} = (\boldsymbol{X}_n - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1}_{\text{MSSD},n-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0);$$

then

$$T^2_{\text{MSSD},n} \sim \frac{2p}{n-2p+1} F_{p,\frac{1}{2}(n-2p+1)} \text{ for } n > 2p-1,$$

i.e., $n = 2p+1, 2p+3, \ldots$.

Equation (5b): Case KU

We need to show that for even numbered observations, i.e., when n is an even number,

$$T^2_{\mathrm{MSSD},n} = \left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1}_{\mathrm{MSSD},n-2}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right) \sim \frac{2p}{n-2p} F_{p,\frac{1}{2}(n-2p)}$$

Proof:

If $\boldsymbol{X}_j$, j = 1, 2, 3, …, are i.i.d. $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ variables, then

$$\boldsymbol{X}_i - \boldsymbol{X}_{i-1} \sim N_p(\boldsymbol{0},2\boldsymbol{\Sigma}),\ i = 2, 4, 6, \ldots$$

and

$$\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right) \sim N_p(\boldsymbol{0},\boldsymbol{\Sigma}),\ i = 2, 4, 6, \ldots .$$

Thus, from Eq. (A3) of Theorem B,

$$\frac{1}{2}\sum_{i=2,4,6}^{n-2}\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right)\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right)' \sim W_p\left(\frac{n-2}{2},\boldsymbol{\Sigma}\right),$$

i.e.,

$$\boldsymbol{S}_{\mathrm{MSSD},n-2} \sim W_p\left(\frac{n-2}{2},\boldsymbol{\Sigma}\right). \tag{A6}$$

Because $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is known, then

$$\boldsymbol{X}_n - \boldsymbol{\mu}_0 \sim N_p(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{A7}$$

Substituting Eq. (A6) and (A7) into Eq. (A1) and (A2) of Theorem A,

$$\frac{\left(\frac{n-2}{2}-p+1\right)\left(\frac{n-2}{2}\right)}{p\left(\frac{n-2}{2}\right)}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1}_{\mathrm{MSSD},n-2}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right) \sim F_{p,\frac{n-2}{2}-p+1}$$

i.e.,

$$\frac{(n-2p)}{2p}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1}_{\mathrm{MSSD},n-2}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right) \sim F_{p,\frac{1}{2}(n-2p)}.$$

Define

$$T^2_{\mathrm{MSSD},n} = \left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1}_{\mathrm{MSSD},n-2}\left(\boldsymbol{X}_n - \boldsymbol{\mu}_0\right);$$

then

$$T^2_{\mathrm{MSSD},n} \sim \frac{2p}{n-2p} F_{p,\frac{1}{2}(n-2p)} \text{ for } n > 2p,$$

i.e., $n = 2p+2, 2p+4, \ldots$.

Equation (6a): Case UU

We need to show that for odd numbered observations, i.e., when n is an odd number,

$$T^2_{\mathrm{MSSD},n} = \left(\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1}\right)' \boldsymbol{S}^{-1}_{\mathrm{MSSD},n-1}\left(\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1}\right) \sim \frac{2np}{(n-2p+1)(n-1)} F_{p,\frac{1}{2}(n-2p+1)}$$

Proof:

If $\boldsymbol{X}_j$, j = 1, 2, 3, …, are i.i.d. $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ variables, then

$$\boldsymbol{X}_i - \boldsymbol{X}_{i-1} \sim N_p(\boldsymbol{0},2\boldsymbol{\Sigma}),\ i = 2, 4, 6, \ldots$$

and

$$\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right) \sim N_p(\boldsymbol{0},\boldsymbol{\Sigma}),\ i = 2, 4, 6, \ldots .$$

Thus, from Eq. (A3) of Theorem B,

$$\frac{1}{2}\sum_{i=2,4,6}^{n-1}\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right)\left(\boldsymbol{X}_i - \boldsymbol{X}_{i-1}\right)' \sim W_p\left(\frac{n-1}{2},\boldsymbol{\Sigma}\right),$$

i.e.,

$$\boldsymbol{S}_{\mathrm{MSSD},n-1} \sim W_p\left(\frac{n-1}{2},\boldsymbol{\Sigma}\right). \tag{A8}$$

Because μ is unknown,

$$\overline{\boldsymbol{X}}_{n-1} \sim N_p\left(\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{n-1}\right)$$

Then,

$$\boldsymbol{X}_n - \overline{\boldsymbol{X}}_{n-1} \sim N_p\left[\boldsymbol{0},\left(1+\frac{1}{n-1}\right)\boldsymbol{\Sigma}\right] \equiv N_p\left(\boldsymbol{0},\frac{n}{n-1}\boldsymbol{\Sigma}\right)$$

and

$$\sqrt{\frac{n-1}{n}}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim N_p\left(\mathbf{0},\boldsymbol{\Sigma}\right) \tag{A9}$$

Substituting Eq. (A8) and (A9) into Eq. (A1) and (A2) of Theorem A,

$$\frac{\left(\frac{n-1}{2}-p+1\right)\left(\frac{n-1}{2}\right)\left(\frac{n-1}{n}\right)}{p\left(\frac{n-1}{2}\right)}$$

$$\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-1}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim F_{p,\frac{n-1}{2}-p+1}$$

i.e.,

$$\frac{(n-2p+1)(n-1)}{2np}$$

$$\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-1}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim F_{p,\frac{1}{2}(n-2p+1)}$$

Define

$$T^2_{\mathrm{MSSD},n}=\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-1}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right);$$

then

$$T^2_{\mathrm{MSSD},n}\sim\frac{2np}{(n-2p+1)(n-1)}F_{p,\frac{1}{2}(n-2p+1)}$$

for $n > 2p-1$, i.e., $n = 2p+1, 2p+3, \ldots$.

Equation (6b): Case UU

We need to show that for even numbered observations, i.e., when n is an even number,

$$T^2_{\mathrm{MSSD},n}=\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-2}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim\frac{2np}{(n-2p)(n-1)}F_{p,\frac{1}{2}(n-2p)}$$

Proof:

If $\boldsymbol{X}_j$, j = 1, 2, 3, …, are i.i.d. $N_p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ variables, then

$$\boldsymbol{X}_{\boldsymbol{i}}-\boldsymbol{X}_{\boldsymbol{i}-1}\sim N_p\left(\mathbf{0},2\boldsymbol{\Sigma}\right),\ i = 2, 4, 6, \ldots$$

and

$$\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_{\boldsymbol{i}}-\boldsymbol{X}_{\boldsymbol{i}-1}\right)\sim N_p\left(\mathbf{0},\boldsymbol{\Sigma}\right),\ i = 2, 4, 6, \ldots .$$

Thus, from Eq. (A3) of Theorem B,

$$\frac{1}{2}\sum_{\boldsymbol{i}=2,4,6}^{\boldsymbol{n}-2}\left(\boldsymbol{X}_{\boldsymbol{i}}-\boldsymbol{X}_{\boldsymbol{i}-1}\right)\left(\boldsymbol{X}_{\boldsymbol{i}}-\boldsymbol{X}_{\boldsymbol{i}-1}\right)'\sim W_p\left(\frac{n-2}{2},\boldsymbol{\Sigma}\right),$$

i.e.,

$$\boldsymbol{S}_{\mathrm{MSSD},\boldsymbol{n}-2}\sim W_p\left(\frac{n-2}{2},\boldsymbol{\Sigma}\right). \tag{A10}$$

Because μ is unknown,

$$\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\sim N_p\left(\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{n-1}\right)$$

Then,

$$\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\sim N_p\left[\mathbf{0},\left(\frac{n}{n-1}\right)\boldsymbol{\Sigma}\right]$$

and

$$\sqrt{\frac{n-1}{n}}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim N_p\left(\mathbf{0},\boldsymbol{\Sigma}\right) \tag{A11}$$

Substituting Eq. (A10) and (A11) into Eq. (A1) and (A2) of Theorem A,

$$\frac{\left(\frac{n-2}{2}-p+1\right)\left(\frac{n-2}{2}\right)\left(\frac{n-1}{n}\right)}{p\left(\frac{n-2}{2}\right)}$$

$$\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-2}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim F_{p,\frac{n-2}{2}-p+1}$$

i.e.,

$$\frac{(n-2p)(n-1)}{2np}$$

$$\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-2}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)\sim F_{p,\frac{n-2p}{2}}$$

Define

$$T^2_{\mathrm{MSSD},\,n}=\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right)'\boldsymbol{S}^{-1}_{\mathrm{MSSD},\boldsymbol{n}-2}\left(\boldsymbol{X}_{\boldsymbol{n}}-\overline{\boldsymbol{X}}_{\boldsymbol{n}-1}\right);$$

then

$$T^2_{\mathrm{MSSD},n}\sim\frac{2np}{(n-2p)(n-1)}F_{p,\frac{1}{2}(n-2p)}\ \text{for } n > 2p,$$

i.e., $n = 2p+2, 2p+4, \ldots$ ■

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 240-250

1538 – 9472/05/$95.00

An Empirical Evaluation Of The Retrospective Pretest: Are There Advantages To Looking Back?

Paul A. Nakonezny
Center for Biostatistics and Clinical Science
University of Texas Southwestern Medical Center

Joseph Lee Rodgers
Department of Psychology
University of Oklahoma

This article builds on research regarding response shift effects and retrospective self-report ratings. Results suggest moderate evidence of a response shift bias in the conventional pretest-posttest treatment design in the treatment group. The use of explicitly worded anchors on response scales, as well as the measurement of knowledge ratings (a cognitive construct) in an evaluation methodology setting, helped to mitigate the magnitude of a response shift bias. The retrospective pretest-posttest design provides a measure of change that is more in accord with the objective measure of change than is the conventional pretest-posttest treatment design with the objective measure of change, for the setting and experimental conditions used in the present study.

Key words: Response shift bias, quasi-experimentation, retrospective pretest-posttest design, retrospective pretest, measuring change

Introduction

More than 30 years after Cronbach and Furby (1970) posited their compelling question, "How we should measure change—or should we?," the properties of the change score continue to attract much attention in educational and psychological measurement. Self-report evaluations are frequently used to measure change in treatment and educational training interventions. In using self-report instruments, it is assumed that a subject's understanding of the standard of measurement for the dimension being measured will not change from pretest to posttest (Cronbach & Furby, 1970).

Paul A. Nakonezny is an Assistant Professor of Biostatistics in the Center for Biostatistics and Clinical Science at the University of Texas Southwestern Medical Center, 6363 Forest Park Rd., Suite 651, Dallas, TX 75235. Email: paul.nakonezny@utsouthwestern.edu. Joseph Lee Rodgers is a Professor of Psychology in the Department of Psychology at the University of Oklahoma, Norman, OK, 73019.

If the standard of measurement is not comparable between the pretest and posttest scores, however, then self-report evaluations in pretest-posttest treatment designs may be contaminated by a response shift bias (Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Maxwell & Howard, 1981). A response shift becomes a bias if the experimental intervention changes the subject's internal evaluation standard for the dimension measured and, hence, changes the subject's interpretation of the anchors of a response scale.

When a response shift is presumably a result of the treatment, a treatment-induced response shift bias should occur in the treatment group and not in the control group. However, another possible source of contamination in response shifts, for both the treatment and control groups, is exposure to the conventional pretest, which could have a priming effect and confounding influence on subsequent self-report ratings (Hoogstraten, 1982; Spranger & Hoogstraten, 1989). A response shift, nevertheless, results in different scale units (metrics) at the posttest than at the pretest, which could produce systemic errors of measurement that threaten evaluation of the basic treatment effect.

When self-report evaluations must be used to measure change, the traditional pretest-posttest treatment design can be modified to include a retrospective pretest at the time of the posttest (e.g., Howard & Dailey, 1979; Howard, Millham, Slaten, & O'Donnell, 1981; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck, & Bray, 1979). After filling out the posttest, subjects then report their memory or perception of what their score would have been prior to the treatment (this is referred to as a retrospective self-report pretest).

Because it is presumed that the self-report posttest and the retrospective self-report pretest would be filled out with respect to the same internal standard, a comparison of the traditional pretest with the retrospective pretest scores within the treatment group would provide an indication of the presence of a response shift bias (Howard et al., 1979). If a response shift bias is present, as indicated by an appreciable mean difference between scores on the conventional pretest and the retrospective pretest, then comparison of the posttest with the retrospective pretest scores would eliminate treatment-induced response shifts and, thus, provide an unconfounded and unbiased estimate of the treatment effect (Howard et al. 1979).

Thus, the retrospective self-report pretest is a method that can be used to obtain pretreatment estimates of subjects' level of functioning (on cognitive, behavioral, and attitudinal dimensions) that are measured with respect to the same internal standard (i.e, in a common metric) as the posttest rating. Retrospective self-report pretests could be used in at least three evaluation research settings: (a) to attenuate a response shift bias (as mentioned above), (b) when conventional pretest data or concurrent data are not available, or (c) when researchers want to measure change on dimensions not included in earlier-wave longitudinal data.

However, the use of retrospective self-reports in the measurement of change has not gained popular acceptance among social scientists. There seem to be at least two possible, yet related, reasons for skepticism and reservation concerning the use of retrospective ratings. First, retrospective self-reports may be perceived to be counter to the paradigm of objective measurement that is rooted in the philosophy of logical positivism (an epistemology in the social sciences that views subjective measures as obstacles toward an objective science of measurement). Second, retrospective self-reports are susceptible to a response-style bias (e.g., memory distortion, subjects' current attitudes and moods, subject acquiescence, social desirability), which could presumably affect ratings in both the treatment and control groups.

Nonetheless, in self-report pretest-posttest treatment designs, previous psychometric research has demonstrated empirical support for the retrospective pretest-posttest difference scores over the traditional pretest-posttest change scores in providing an index of change more in agreement with objective measures of change on both cognitive and behavioral dimensions (e.g., Hoogstraten, 1982; Howard & Dailey, 1979; Howard, Millham, Slaten, & O'Donnell, 1981; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Howard, Schmeck, & Bray, 1979; Spranger & Hoogstraten, 1989).

The purpose of this article is to build on a previous line of research, by Howard and colleagues and Hoogstraten and Spranger, on response shift effects and retrospective self-report ratings. Specifically, the current study examined (a) response shift bias in the self-report pretest-posttest treatment design in an evaluation setting, (b) the validity of the retrospective pretest-posttest design in estimating treatment effects, (c) the effect of memory distortion on retrospective self-report pretests, and (d) the effect of pretesting on subsequent and retrospective self-report ratings.

Methodology

A cross-sectional quasi-experimental pre-post treatment design (Cook & Campbell, 1979) with data from 240 participants was used to address the research objectives of this study. The design included a treatment group and a no-treatment comparison group. Participants in the treatment group were 124 students enrolled in an undergraduate epidemiology course (Class A) and participants in the no-treatment comparison group were 116 students enrolled in an

undergraduate health course (Class B). The 240 participants were undergraduate students who attended a large public University in the state of Texas during the Spring semester of 2002 and who met the following criteria for inclusion in the study:

(a) at least 18 years of age,
(b) must not have taken an epidemiology course or a course that addressed infectious disease epidemiology, and
(c) must not have been concurrently enrolled in Class A and Class B.

Participants signed a consent form approved by the Institutional Review Board of the University and received bonus class points for participating. The gender composition was 29 males and 211 females, and the age range was 18 to 28 years (with an average age of 20.61 years, *SD* = 2.46). The racial distribution of the study sample included 181 (75.4 %) Caucasians, 37 (15.4 %) African Americans, 13 (5.4 %) Hispanics, and 9 (3.8 %) Asians. Participant characteristics by group are reported in Table 1.

The treatment in this design was a series of lectures on infectious disease epidemiology that was part of the course content in Class A, but not in Class B. Participants' knowledge of infectious disease epidemiology—the basic construct in this study—was measured with a one-item self-report instrument and with a ten-item objective instrument, and the same item-scale instruments were used for both the treatment and no-treatment comparison groups. Each instrument was operationalized as the mean of the items measuring each scale, and was scored so that a higher score equaled more knowledge of infectious disease epidemiology.

The conventional self-report instrument, which was used in both the pretest and posttest measurement settings, consisted of one-item that asked participants to respond to the following question: "How much do you know about the principles of Infectious Disease Epidemiology?" The current study measured this one-item using a six-point Likert-type scale that ranged from 0 (not much at all) to 5 (very very much), with verbal labels for the intermediate scale points.

The retrospective self-report pretest, which was similar to the conventional self-report pretest, consisted of one-item that asked participants to respond to the following question: "Three months ago, at the beginning of the semester, you were asked how much you knew about Infectious Disease Epidemiology. Thinking back 3 months ago, to the beginning of the semester, how much did you know about Infectious Disease Epidemiology at that time?"

The current study measured this one retrospective item using a six-point Likert-type scale like that mentioned above. The objective instrument, which was used in both the pretest and posttest measurement settings, consisted of 10 multiple choice items/questions that tapped the participants' knowledge level of infectious disease epidemiology.

Participants within each group—treatment group and no-treatment comparison group—were randomly assigned to four pretesting conditions, which represented the pretesting main effect. Participants in condition 1 completed both the self-report and objective pretests. Participants in condition 2 completed the objective pretest. Participants in condition 3 completed the self-report pretest. Participants in condition 4 completed neither the self-report pretest nor the objective pretest.

All participants, regardless of the assigned condition, completed the posttests as well as the retrospective and recalled self-report pretests. The sample size per condition by group was approximately equal, and the participants across the four conditions were not significantly different in age, F's < .91, p's > .43, gender, race, and academic classification (e.g., freshman, sophomore, junior, senior), respectively, χ^2's < 1.08, p's > .29.

At the outset of the academic semester (time 1), before the treatment, all participants in the assigned condition completed the pretest(s) which measured their baseline knowledge level of infectious disease epidemiology. The pretests were collected immediately after they were completed and then the treatment was begun (for participants in the treatment group). At the conclusion of the instruction on infectious disease epidemiology (the treatment), which occurred at about the end of the 12th week of classes (time 2), participants in the treatment

Table 1. Participant Characteristics

	Treatment Group (n = 124)			Comparison Group (n = 116)			
Variable	Mean	SD	n (%)	Mean	SD	n (%)	p
Age (years)	20.5	1.9	124 (51.7)	20.6	2.9	116 (48.3)	.66[a]
Gender							.68[b]
Male			16 (12.9)			13 (11.2)	
Female			108 (87.1)			103 (88.8)	
Race							.58[b]
White			91 (73.4)			90 (77.6)	
Black			21 (16.9)			16 (13.8)	
Hispanic			08 (06.5)			05 (04.3)	
Asian			04 (03.2)			05 (04.3)	
Classification							.65[b]
Freshman			17 (13.7)			22 (19.0)	
Sophomore			42 (33.9)			41 (35.3)	
Junior			49 (39.5)			40 (34.5)	
Senior			16 (12.9)			13 (11.2)	

*a*F statistic was used to test for mean age differences between the treatment group and the no-treatment comparison group.
*b*Chi-Square statistic was used to test for differences between the treatment group and the no-treatment comparison group on gender, race and classification, respectively.

group and participants in the no-treatment comparison group (who were not exposed to the treatment) completed the objective posttest. The objective posttest was identical to the objective pretest.

One week after completion of the objective posttest (time 3), participants in both the treatment and no-treatment comparison groups completed the self-report posttest and the retrospective self-report pretest. Participants first completed the self-report posttest and, while keeping the posttest in front of them, they then filled out the retrospective self-report pretest.

The self-report posttest was identical to the conventional self-report pretest. The retrospective self-report pretest was similar to the conventional self-report pretest, but the wording of the question accounted for the retrospective time frame.

Lastly, about one month after completion of the self-report posttest and retrospective self-report pretest, at the end of the academic semester (time 4), participants in both the treatment and no-treatment comparison groups completed the recalled self-report pretest, which permitted a memory test of the initial/conventional self-report pretest completed at the outset of the academic semester (time 1)

and, thus, yielded a test for a response-style bias of the retrospective self-report pretest rating.

The recalled self-report pretest consisted of one-item that asked participants to respond to the following question: "Four months ago, at the beginning of the semester, you were asked how much you knew about Infectious Disease Epidemiology. Please recall, remember, and be as accurate as possible, how you responded at that time regarding your knowledge level of Infectious Disease Epidemiology (i.e., how did you respond at that time?)." The current study measured this one-item using a six-point Likert-type scale similar to that described above.

The research objectives of this study were addressed by analyzing the series of pretest and posttest ratings using the dependent *t* test, the Pearson product-moment correlation (*r*), and analysis of variance (ANOVA). Estimates of the magnitude of the effect size were also computed (Rosenthal, Rosnow, & Rubin, 2000). The effect size estimators that accompanied the dependent *t* test and the ANOVA were Cohen's (1988) *d* and eta-square (η^2), respectively.

The Pearson product-moment correlation (*r*) was also used as the effect size estimator in the specific regression analyses. To test the response shift hypothesis, the dependent *t* test was carried out comparing the retrospective self-report pretest to the conventional self-report pretest within the treatment and no-treatment comparison groups. The dependent *t* test also was used to compare the recalled self-report pretest to the conventional self-report pretest, which tested for the effect of memory distortion in the retrospective pretest-posttest design.

The Pearson correlation between the recalled self-report pretest and the conventional self-report pretest and between the recalled self-report pretest and the retrospective self-report pretest also was used to test for memory distortion. To examine the relative validity of the retrospective pretest-posttest design in estimating treatment effects, a simple correlation analysis was further used to assess the relationship between the self-reported measures of change and the objective measure of change in both the conventional and retrospective pretest-posttest designs for the treatment and no-treatment comparison groups.

One-way ANOVA was used to assess the pretesting main effect (the four pretesting conditions) on the conventional self-report posttest, the retrospective self-report pretest, and the recalled self-report pretest. The Ryan-Einot-Gabriel-Welsch multiple-range test was used to carry out the cell means tests for the pretesting main effect for the ANOVA. A separate ANOVA was performed for the treatment group and the no-treatment comparison group.

Results

Response Shift

Using the conventional pre/post self-report change score and the objective pre/post change score, effects were found in the treatment group, *t*'s > 8.60, *p*'s < .0001, but not in the no-treatment group, *t*'s < .84, *p*'s > .40. The dependent *t* test, averaged across conditions 1 and 3, revealed a marginally significant mean difference between the retrospective self-report pretest and the conventional self-report pretest in the treatment group, $t(61) = -1.56$, $p < .10$, $M = -0.16$, $SD = .81$, $d = -0.20$, and, unexpectedly, a significant mean difference in the no-treatment comparison group, $t(54) = -2.99$, $p < .004$, $M = -0.30$, $SD = .76$, $d = -0.39$. These findings provide moderate support for the response shift hypothesis. Means and standard deviations for the pretests and posttests by condition and group are reported in Table 2.

Treatment Effects

To assess the relative validity of the retrospective pretest-posttest design in estimating treatment effects, the self-reported measures of change were compared with the objective measure of change in both the conventional and retrospective pretest-posttest designs for the treatment and no-treatment comparison groups. For the treatment group, averaged across conditions 1 and 2, the Pearson correlation results indicated that the retrospective pre/post self-report change score was somewhat more in accord with the objective pre/post measure of change ($r = .32$, $p < .01$) than was the conventional pre/post self-report

Table 2. Means and Standard Deviations for the Pretests and Posttests by Condition and Group.

	Treatment Group (n = 124)					
	Self-Report				Objective	
Pretest Condition	Pretest	Posttest	Retro	Recalled	Pretest	Posttest
Condition 1						
M	1.10	2.34	0.86	1.13	1.82	4.06
SD	1.01	0.81	0.87	0.91	0.76	0.98
Condition 2						
M		2.81	1.18	1.65	1.89	3.78
SD		0.82	0.93	1.09	0.82	0.73
Condition 3						
M	0.99	2.21	0.91	1.30		3.48
SD	1.14	0.92	0.80	0.95		1.01
Condition 4						
M		2.43	1.03	1.26		3.66
SD		0.77	0.93	0.86		0.93

	No-Treatment Comparison Group (n = 116)					
	Self-Report				Objective	
Pretest Condition	Pretest	Posttest	Retro	Recalled	Pretest	Posttest
Condition 1						
M	0.79	0.86	0.52	0.83	1.67	1.50
SD	0.82	0.87	0.78	0.85	0.66	0.79
Condition 2						
M		1.09	0.71	0.99	1.68	1.67
SD		0.98	0.86	0.89	0.56	0.67
Condition 3						
M	1.07	1.19	0.73	1.03		1.82
SD	0.84	0.75	0.87	0.91		0.61
Condition 4						
M		1.43	0.67	0.83		1.55
SD		0.89	0.71	0.79		0.72

Note. Retro = retrospective self-report pretest; Recalled = recalled self-report pretest (used to test for the threat of memory distortion). Participants in condition 1 completed both the self-report and objective pretests; Participants in condition 2 completed the objective pretest; Participants in condition 3 completed the self-report pretest; Participants in condition 4 completed neither the self-report pretest nor the objective pretest. All participants, regardless of the assigned condition, completed the posttests as well as the retrospective and recalled self-report pretests. The sample size per condition by group was approximately equal.

change score with the objective pre/post measure of change (r = .26, p < .18).

Conversely, as anticipated, for the no-treatment comparison group averaged across conditions 1 and 2, the magnitude of the correlation between the conventional pre/post self-report change score and the objective pre/post measure of change, r = .27, p < .16, was greater than the correlation between the retrospective pre/post self-report change score and the objective pre/post change score, r = .04, p < .75, albeit neither was significant.

Memory Distortion

The effect of memory distortion within the retrospective pretest-posttest design was also examined. For the treatment group, averaged across conditions 1 and 3, the results of the dependent t test revealed no significant mean difference between the recalled self-report pretest (M = 1.22, SD = .93) and the conventional self-report pretest (M = 1.05, SD = 1.07), t(61) = 1.56, p < .12, M = .17, SD = .89, d = 0.19 (Table 2). Further, the no-treatment comparison group had nearly identical average scores on the recalled self-report pretest (M = .933, SD = .882) and the conventional self-report pretest (M = .935, SD = .832), averaged across conditions 1 and 3, suggesting no significant mean difference, t(54) = -0.01, p < .99, M = -0.002, SD = .85, d = -0.002 (Table 2). The dependent t test results suggest no significant presence of memory distortion in the retrospective pretest-posttest treatment design.

A simple correlation analysis also was used to test for memory distortion. The Pearson correlations between the recalled pre/post self-report change score and the conventional pre/post self-report change score, averaged across conditions 1 and 3, and between the recalled pre/post self-report change score and the retrospective pre/post self-report change score, averaged across all four conditions, were significant and reasonably high in the treatment group (r = .64 and r = .63, respectively, p's <.0001) and in the no-treatment comparison group (r = .54 and r = .56, respectively, p's <.0001).

Further, the Pearson correlations between the recalled self-report pretest and the conventional self-report pretest, averaged across conditions 1 and 3, and between the recalled self-report pretest and the retrospective self-report pretest, averaged across all four conditions, were significant and fairly high in the treatment group (r = .61 and r = .62, respectively, p's <.0001) and in the no-treatment comparison group (r = .60 and r = .68, respectively, p's <.0001).

Pretesting Effects

The ANOVA revealed a significant pretesting effect on the conventional self-report posttest in the treatment group, F(3, 120) = 3.04, p < .03, η^2 = .07, but not in the no-treatment comparison group, F(3, 112) = 2.11, p < .10, η^2 = .05. The cell means tests, however, indicated no significant difference between the conventional self-report pretest condition and the no-pretest condition on the conventional self-report posttest score in the treatment and no-treatment comparison groups, $t$is < 1.05, p's > .30. Further, the ANOVA revealed no significant pretesting effect on the retrospective self-report pretest and on the recalled self-report pretest in the treatment group, $F$is(3, 120) < 1.64, p's > .18, η^2s < .04, and in the no-treatment comparison group, F's(3, 112) < 0.46, p's > .70, η^2 = .01. The ANOVA results suggest that pretesting had little effect on the subsequent and retrospective self-report ratings. Means and standard deviations for the pretests and posttests by condition and group are reported in Table 2.

Response Shift

Do treatments in evaluation research alter participants' perceptions in a manner which contaminates self-report assessment of the treatment? The findings of the current study indicate moderate evidence of a response shift bias in the conventional pretest-posttest treatment design in the treatment group, suggesting that the knowledge ratings from self-report pretest to posttest were partially a result of respondents recalibrating their internal evaluation standard for the dimension measured. A plausible interpretation of this moderate response shift bias in the treatment group is that the use of explicitly worded anchors on response scales in measuring the participant's self-

reported knowledge of infectious disease epidemiology—a cognitive construct—in a classroom setting helped to mitigate the magnitude of a response shift effect.

The degree of a response shift bias is, in part, conditional upon the experimental setting, the type of constructs measured, and the response scale anchors. Previous research (e.g., Collins et al., 1985; Finney, 1981; Howard, Schmeck, & Bray, 1979; Maisto et al., 1982) suggests that the magnitude of a response shift bias seems to be smaller when cognitive constructs are measured (such as knowledge ratings) and when questions and anchors on response scales are explicit.

Although no treatment effects were found in the no-treatment comparison group, as expected, a significant mean difference between the retrospective self-report pretest and the conventional self-report pretest was found, suggesting a non-treatment-related response shift. Typically, a response shift is a result of respondents changing their internal evaluation standard for the dimension measured between pretest and posttest because of exposure to the treatment. There are, however, alternative sources of bias in response shifts—such as a pretesting effect, memory distortion, and subject acquiescence—which could presumably affect ratings in both the treatment and no-treatment comparison groups (Collins et al., 1985; Howard & Dailey, 1979; Sprangers & Hoogstraten, 1989).

Because the results of the current study suggest that memory distortion and pretesting had little effect on subsequent self-report ratings, a plausible explanation for the response shift bias in the no-treatment comparison group is subject acquiescence. In the case of subject acquiescence, participants in the no-treatment comparison group might have realized that their knowledge level had not changed since their initial pretest rating, but their desire to provide the experimenter with a favorable set of results (given that bonus grade points were given for participation in the study) led them to lower their retrospective self-report rating. The retrospective rating was administered at the same time as the self-report posttest, allowing participants in the no-treatment comparison group the opportunity to adjust their retrospective preratings in a downward direction.

Treatment Effects in the Retrospective Pre/Post Design

The principal focus of the current study was to evaluate the validity of the retrospective pretest-posttest design in estimating treatment effects. The findings of the present study favor the retrospective pre/post self-report measure of change in providing a measure of self-reported change that better reflects the objective index of change on a construct of knowledge rating. This finding is in line with previous psychometric research (e.g., Hoogstraten, 1982; Howard & Dailey, 1979; Howard et al., 1979; Howard, Schmeck, & Bray, 1979; Spranger & Hoogstraten, 1989), and is most likely a result of the self-report posttest and the retrospective self-report pretest being filled out with respect to the same internal standard, the same metric. This, therefore, mitigates the treatment-induced response shift bias, minimizes errors of measurement, and provides an unconfounded and unbiased estimate of the treatment effect (Howard et al., 1979).

Although there is empirical support for the retrospective pretest-posttest difference scores over the conventional pretest-posttest change scores in providing an index of change more in agreement with objective measures of change, this is not to suggest that the conventional self-report pretest should be substituted by the retrospective self-report rating. Rather, in light of the findings of this study as well as those from previous studies, the suggestion put forward is that retrospective self-report pretests could be used in at least three evaluation research settings: (a) to test for and attenuate a response shift bias in the conventional pretest-posttest treatment design, (b) when conventional pretest data or concurrent data are not available, or (c) when researchers want to measure change on dimensions not included in earlier-wave longitudinal data.

Testing for Threats to Validity

Also evaluated were the potential threats of memory distortion and pretesting effect to the internal validity of the retrospective pretest-posttest treatment design in the current study.

Retrospective self-report ratings could be limited by memory lapses and pretests could exert a confounding influence on subsequent self-report ratings, including retrospective ratings, which could threaten evaluation of the treatment effect (Collins et al., 1985; Howard & Dailey, 1979; Sprangers & Hoogstraten, 1989). In general, the present study found no significant presence of memory distortion or a pretesting effect in the retrospective pretest-posttest treatment design used in the current study.

This is not to suggest that memory distortion or a pretesting effect should not be accounted for as potential threats to the basic retrospective pretest-posttest design. Rather, what this finding suggests is that memory distortion and pretesting are not influencing the interpretation of the treatment effect in the type of retrospective pretest-posttest design used in the present study. The conventional self-report pretest and the recalled self-report pretest were only separated by four months, which may have in part mitigated the effect of memory distortion. Previous research (e.g., Finney, 1981; Howard, Dailey, & Gulanick, 1979; Howard, Schmeck, & Bray, 1979; Maisto et al., 1982), nonetheless, suggests that a pretesting effect can be mitigated and moderate-to-high recall accuracy is possible when cognitive constructs are measured (such as knowledge ratings) and when retrospective questions are specific and anchors on response scales are explicit (these conditions are consistent with those used in this study).

An Application of the Retrospective Pre/Post Design

In this section, a study by Nakonezny, Rodgers, and Nussbaum (2003) which applied the retrospective pretest-posttest treatment design to a unique research setting is briefly described.

Nakonezny et al. (2003) examined the effect of later life parental divorce on solidarity in the relationship between the adult child and older parent. This examination was achieved by testing the buffering hypothesis that greater levels of predivorce solidarity in the adult child/older parent relationship buffers damage to postdivorce solidarity. The unique and uncommon nature of the phenomenon of later life parental divorce, however, precluded access to these atypical divorcees prior to their divorce, which led to the necessity to use a retrospective pretest-posttest treatment design by Nakonezny et al. (2003).

As mentioned earlier, one research scenario under which retrospective self-report pretests could be used is when conventional pretest data are not available, which was the case in the Nakonezny et al. (2003) study.

In the retrospective design used in Nakonezny et al. (2003), predivorce/pretest solidarity included retrospective measures of the same scale-item instruments that were used to measure postdivorce/posttest solidarity. The wording of the questions, however, was changed to account for the retrospective time frame. Parents in the divorced group were asked to remember the period before their divorce and to provide a retrospective self-report account of solidarity in the relationship with their oldest living adult child during the predivorce period. The average number of years from the divorce decree to the date of data collection was about 8 years.

Also, parents in the intact two-parent family group (the no-treatment comparison group) were asked to remember back approximately five years from the date of participation in the study and to provide a retrospective self-report account of solidarity in the relationship with their oldest living adult child during that period, which represented the pretest period for the intact group. The basic findings of Nakonezny et al. (2003), using a retrospective pretest-posttest treatment design, were in the hypothesized directions for both groups. Nakonezny et al. (2003) can be consulted for a complete explanation of this application of the retrospective pretest-posttest treatment design in a social science evaluation research setting.

Future Research

The current study and previous research suggest that, under certain conditions, the retrospective pretest-posttest treatment design provides a more accurate assessment of change than that of the conventional pretest-posttest treatment design. However, the retrospective pretest-posttest treatment design still remains something of an enigma, and future research

concerning the validity of the retrospective pretest-posttest design is still needed. Further research is needed to address the effect of subject acquiescence and other extraneous sources of invalidity on self-report ratings in the retrospective pretest-posttest treatment design.

Further research also is needed to determine the different types of retrospective pretest-posttest designs, experimental conditions, treatment interventions, constructs, and time lapses that are most susceptible to a response shift bias and that most affect recall accuracy of retrospective self-report ratings. Most importantly, a next step in this line of evaluation research is to continue to explore the research settings and applications in both the social and behavioral sciences under which retrospective self-report ratings are appropriate and under which the retrospective pretest-posttest design produces unbiased estimates of treatment effects.

Conclusion

The empirical findings support that a moderate response shift bias occurred in the conventional pretest-posttest treatment design in the treatment group, and are highly suggestive that the knowledge ratings from self-report pretest to posttest were partially a result of respondents recalibrating their internal evaluation standard for the dimension measured (presumably because of exposure to the treatment). The results further suggest that the use of explicitly worded anchors on response scales as well as the measurement of knowledge ratings (a cognitive construct) in an evaluation methodology setting mitigated the magnitude of a response shift bias. Subject acquiescence is a likely explanation of the unexpected non-treatment-related response shift bias that occurred in the no-treatment comparison group.

Further, the current study suggests that the retrospective pretest-posttest treatment design provides a more accurate assessment of change than that of the conventional pretest-posttest treatment design for the setting and experimental conditions used in the present study. Based on these results, it is suggested that researchers collect both a conventional self-report pretest and a retrospective self-report pretest when using a conventional pretest-posttest treatment design in evaluation research settings. Retrospective self-report pretests could be used, however, when conventional self-report pretest data are not available. In support of this scenario, we present an example of an innovative application of the retrospective pretest-posttest treatment design in a social science research setting. Finally, the ultimate value of this work may lie in its ability to renew interest in the retrospective pretest-posttest treatment design, to motivate future research, and to sharpen the empirical focus of that research.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Collins, L. M., Graham, J. W., Hansen, W. B., & Johnson, C. A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. *Applied Psychological Measurement, 9*, 301-309.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Cronbach, L. J., & Furby, L. (1970). How we should measure change--or should we? *Psychological Bulletin*, 74, 68-80.

Finney, H. C. (1981). Improving the reliability of retrospective survey measures: Results of a longitudinal field survey. *Evaluation Review, 5*, 207-229.

Hoogstraten, J. (1982). The retrospective pretest in an educational training context. *Journal of Experimental Education, 50*, 200-204.

Howard, G. S., & Dailey, P. R. (1979). Response shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*, 144-150.

Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subjective response style effects on retrospective measures. *Applied Psychological Measurement, 5*, 89-100.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, *3*, 1-23.

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, *16*, 129-135.

Maisto, S. A., Sobell, L. C., Cooper, A. M., & Sobell, M. B. (1982). Comparison of two techniques to obtain retrospective reports of drinking behavior from alcohol abusers. *Addictive Behaviors*, *7*, 33-38.

Maxwell, S. E. & Howard, G. S. (1981). Change scores: Necessarily anathema? *Educational and Psychological Measurement*, *41*, 747-756.

Nakonezny, P. A., Rodgers, J. L., & Nussbaum, J. F. (2003). The effect of later life parental divorce on adult-child/older-parent solidarity: A test of the buffering hypothesis. *Journal of Applied Social Psychology*, *33*, 1153-1178.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge, UK: Cambridge University Press.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*, 265-272.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 251-274

1538 – 9472/05/$95.00

An Exploration of Using Data Mining in Educational Research

Yonghong Jade Xu
Department of Counseling, Educational Psychology, and Research
The University of Memphis

Technology advances popularized large databases in education. Traditional statistics have limitations for analyzing large quantities of data. This article discusses data mining by analyzing a data set with three models: multiple regression, data mining, and a combination of the two. It is concluded that data mining is applicable in educational research.

Key words: Data mining, large scale data analysis, quantitative educational research, Bayesian network, prediction

Introduction

In the last decade, with the availability of high-speed computers and low-cost computer memory (RAM), electronic data acquisition and database technology have allowed data collection methods that are substantially different from the traditional approach (Wegman, 1995). As a result, large data sets and databases are becoming increasingly popular in every aspect of human endeavor including educational research. Different from the small, low-dimensional homogeneous data sets collected in traditional research activities, computer-based data collection results in data sets of large volume and high dimensionality (Hand, Mannila, & Smyth, 2001; Wegman, 1995).

Yonghong Jade Xu is an Assistant Professor at the Department of Counseling, Educational Psychology, and Research, College of Education, the University of Memphis. The author wishes to thank Professor Darrell L. Sabers, Head of the Department of Educational Psychology at the University of Arizona, and Dr. Patricia B. Jones, Principal Research Specialist at the Center for Computing Information Technology at the University of Arizona, for their invaluable input pertaining to this study. Correspondence concerning this article should be addressed to Yonghong Jade Xu, Email: yxu@memphis.edu

Many statisticians (e.g., Fayyad, 1997; Hand et al., 2001; Wegman, 1995) noticed some drawbacks of traditional statistical techniques when trying to extract valid and useful information from a large volume of data, especially those of a large number of variables. As Wegman (1995) argued, applying traditional statistical methods to massive data sets is most likely to fail because "homogeneity is almost surely gone; any parametric model will almost surely be rejected by any hypothesis testing procedure; fashionable techniques such as bootstrapping are computationally too complex to be seriously considered for many of these data sets; random subsampling and dimensional reduction techniques are very likely to hide the very substructure that may be pertinent to the correct analysis of the data" (p. 292). Moreover, because most of the large data sets are collected from convenient or opportunistic samples, selection bias puts in question any inferences from sample data to target population (Hand, 1999; Hand et al., 2001).

The statistical challenge has stimulated research aiming at methods that can effectively examine large data sets to extract valid information (e.g., Daszykowski, Walczak, & Massart, 2002). New analytical techniques have been proposed and explored. Among them, some statisticians (e.g., Elder & Pregibon, 1996; Friedman, 1997; Hand, 1998, 1999, 2001; Wegman, 1995) paid attention to a new data analysis tool called data mining and knowledge discovery in database. Data mining is a process

of nontrivial extraction of implicit, previously unknown, and potentially useful information from a large volume of data (Frawley & Piatetsky-Shapiro, 1991).

Although data mining has been used in business and scientific research for over a decade, a thorough literature review has found no educational study that used data mining as the method of analysis. To explore the usefulness of data mining in quantitative research, the current study provides a demonstration of the analysis of a large education-related data set with several different approaches, including traditional statistical methods, data mining, and a combination of these two. With different analysis techniques laid side-by-side working on the same data set, the virtue of the illustrated methods, models, outputs, conclusions, and unique characteristics is ready for assessment.

Research Background

According to its advocates, data mining has prevailed as an analysis tool for large data sets because it can efficiently and intelligently probe through an immense amount of material to discover valuable information and make meaningful predictions that are especially important for decision-making under uncertain conditions.

Data mining uses many statistical techniques, including regression, cluster analysis, multidimensional analysis, stochastic models, time series analysis, nonlinear estimation techniques, just to name a few (Michalski, Bratko, & Kubat, 1998).

However, data mining is not a simple rework of statistics; it implements statistical techniques through an automated machine learning system and acquires high-level concepts and/or problem-solving strategies through examples (input data) in a way analogous to human knowledge induction to attack problems that lack algorithmic solutions or have only ill-defined or informally stated solutions (Michalski et al., 1998).

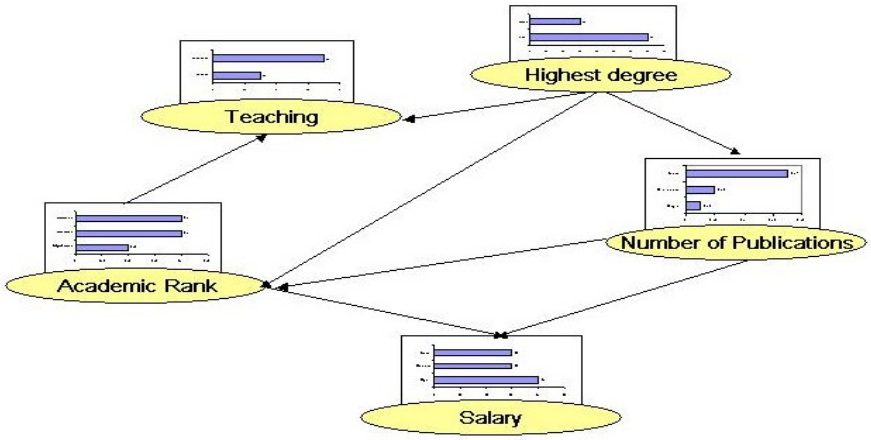
Data mining generates descriptions of rules as output using algorithms such as Bayesian probability, artificial neural networks, decision trees, and generic algorithms that do not assume any parametric form of the appropriate model. Automated analysis processes that reduce or eliminate the need for human interventions become critical when the volume of data goes beyond human ability of visualization and comprehension.

Due to its applied importance, data mining as an academic discipline continues to grow with input from statistics, machine learning, and database management (Fayyad, 1997; Zhou, 2003). One popular algorithm in recent research is the Bayesian Belief Network (BBN), which started from a set of probability rules discovered by Thomas Bayes in the 18th century. The tree-like network based upon Bayesian probability can be used as a prediction model (Friedman et al., 1997). To build such a model, various events (variables) have to be defined, along with the dependencies among them and the conditional probabilities (CP) involved in those dependencies.

Once the variables are ready and the topology is defined, they become the information used to calculate the probabilities of various possible paths being the actual path leading to an event or a particular value of a variable. Through an extensive iteration, a full joint probability distribution is to be constructed over the product state space (defined as the complete combinations of distinct values of all variables) of the model variables. The computational task is enormous because elicitation at a later stage in the sequence results in back-tracking and changing the information that has been elicited at an earlier point (Yu & Johnson, 2002). With the iterative feedback and calculation, a BBN is able to update the prediction probability, the so-called belief values, using probabilistic inference.

BBN combines a sound mathematical basis with the advantages of an intuitive visual representation. The final model of a BBN is expressed as a special type of diagram together with an associated set of probability tables (Heckerman, 1997), as shown in the example in Figure 1. The three major classes of elements are a set of uncertain variables presented as nodes, a

Figure 1. An example of a BBN model. This graph illustrates the three major classes of elements of a Bayesian network; all variables, edges, and CP tables are for demonstration only and do not reflect the data and results of the current study in any way.



set of directed edges (arcs) between variables showing the causal/relevance relationships between variables, and also, a CP table P(A| B_1, B_2,…, B_n) attached to each variable A with parents B_1, B_2, …, B_n. The CPs describe the strength of the beliefs given that the prior probabilities are true.

Because in learning a previously unknown BBN, the calculation of the probability of any branch requires all branches of the network to be calculated (Niedermayer, 1998), the practical difficulty of performing the propagation, even with the availability of high-speed computers, delayed the availability of software tools that could interpret the BBN and perform the complex computation until recently. Although the resulting ability to describe the network can be performed in linear time, given a relatively large number of variables and their product state space, the process of network discovery remains computationally impossible if an exhaustive search in the entire model space is required for finding the network of best prediction accuracy.

As a compromise, some algorithms and utility functions are adopted to direct random selection of variable subsets in the BBN modeling process and to guide the search for the optimal subset with an evaluation function tracking the prediction accuracy (measured by the classification error rate) of every attempted model (Friedman et al., 1997). That is, a stochastic variable subset selection is embedded into the BBN algorithms. The variable selection function conducts a search for the optimal subset using the BBN itself as a part of the evaluation function, the same algorithm that will be used to induce the final BBN prediction model.

Some special features of the BBN are considered beneficial to analyzing large data sets. For instance, to define a finite product state space for calculating the CPs and learning the network, all continuous variables have to be discretized into a number of intervals (bins). With such discretization, variable relationships are measured as associations that do not assume linearity and normality, which minimizes the negative impacts of outliers and other types of irregularities inherent in secondary data sources. Variable discretization also makes a BBN flexible in handling different types of variables and eliminates the sample size as a factor influencing the amount of computation.

With large databases available for research and policy making in education, this study is designed to assess whether the data mining approach can provide educational researchers with extra means and benefits in analyzing large-scale data sets.

Methodology

To examine the usefulness of data mining in educational research, the current study demonstrated the analysis of a large post-secondary faculty data set with three different approaches, including data mining, traditional statistical methods, and a combination of these two. Because data mining shares a few common concerns with traditional statistics, such as estimation of uncertainty, construction of models in defined problem scope, prediction, and so on (Glymour, Madigan, Pregibon, & Smyth, 1997), in order to narrow down the research problem, prediction functions were chosen as a focus of this article to see whether data mining could offer any unique outlook when processing large data sets.

To be specific, all three models were set to search for factors that were most efficient in predicting post-secondary faculty salary. On the statistical side, multiple linear regression was used because it is an established dynamic procedure of prediction; for data mining, prediction was performed with a BBN. Although the major concern of faculty compensation studies is the evaluation of variable importance in salary determination rather than prediction, the purpose of this study was to illustrate a new data analysis technique, rather than to advance the knowledge in the area of faculty compensation. Unless specified otherwise, α = .01 was used in all significance tests.

Data Set

In order to compare different data analysis approaches, the post-secondary faculty data set collected using the National Survey of Postsecondary Faculty 1999 (NSOPF:99) was chosen as a laboratory setting for demonstrating the statistical and data mining methods.

The NSOPF:99 was a survey conducted by the National Center for Education Statistics (NCES) in 1999. The initial sample included 960 degree granting postsecondary institutions and 27,044 full and part-time faculty employed at these institutions. Both the sample of institutions and the sample of faculty were stratified and systematic samples. Approximately 18,000 faculty and instructional staff questionnaires were completed at a weighted response rate of 83 percent. The response rate for the institution survey was 93 percent.

In this study, only faculty data were used which included 18,043 records and 439 original and derived measures. Information was available on faculty demographic backgrounds, workloads, responsibilities, salaries, benefits, and more. The data set was considered appropriate because it is an education-related survey data set, neither too large for traditional analysis approaches nor too small for data mining techniques.

To focus on the salary prediction of regular faculty in postsecondary institutions, only respondents who reported fulltime faculty status were included. Faculty assigned by religious order was excluded as well as those having affiliated or adjunct titles. Also, some respondents were removed from the data set to eliminate invalid salary measures. As a result, the total number of records available for analysis was 9,963. Two-thirds of the records were randomly selected as training data and used to build the prediction models; the remaining one-third were saved as testing data for purpose of cross-validation.

Variables in the data set were also manually screened so that only the most salient measures of professional characteristics were kept to quantify factors considered relevant in determining salary level according to the general guidelines of salary schema in postsecondary institutions and to the compensation literature in higher education. At the end, only 91 (including salary) were kept in the study out of the entire set of variables.

Among them, a few variables were derived from the original answers to the questionnaire in order to avoid redundant or overly specific information. However, multiple measures were kept on teaching, publication, and some other constructs because they quantified different aspects of the underlying constructs; the redundant information among them also offered a chance of testing the differentiation power of the variable selection procedures. Table 1 provides a list of all the 91 variables and their definitions.

Table 1. Name, Definition, and Measurement Scale of the 91 Variables from NSOPF:99.

Variable name	Variable definition	Scale
Q25	Years teaching in higher education institution	Interval
Q26	Positions outside higher education during career	Interval
Q29A1	Career creative works, juried media	Interval
Q29A2	Career creative works, non-juried media	Interval
Q29A3	Career reviews of books, creative works	Interval
Q29A4	Career books, textbooks, reports	Interval
Q29A5	Career exhibitions, performances	Interval
Q29B1	Recent sole creative works, juried media	Interval
Q29B2	Recent sole creative works, non-juried media	Interval
Q29B3	Recent sole reviews of books, works	Interval
Q29B4	Recent sole books, textbooks, reports	Interval
Q29B5	Recent sole presentations, performances	Interval
Q29C1	Recent joint creative works, juried media	Interval
Q29C2	Recent joint creative works, non-juried media	Interval
Q29C3	Recent joint reviews of books, creative works	Interval
Q29C4	Recent joint books, reports	Interval
Q29C5	Recent joint presentations, performances	Interval
Q2REC	Teaching credit or noncredit courses	Ordinal
Q30B	Hours/week unpaid activities at the institution	Interval
Q30C	Hours/week paid activities not at the institution	Interval
Q30D	Hours/week unpaid activities not at the institution	Interval

Table 1 Continued.

Variable name	Variable definition	Scale
Q31A1	Time actually spent teaching undergrads (percentage)	Ratio
Q31A2	Time actually spent teaching graduates (percentage)	Ratio
Q31A3	Time actually spent at research (percentage)	Ratio
Q31A4	Time actually spent on professional growth (percentage)	Ratio
Q31A5	Time actually spent at administration (percentage)	Ratio
Q31A6	Time actually spent on service activity (percentage)	Ratio
Q31A7	Time actually spent on consulting (percentage)	Ratio
Q32A1	Number of undergraduate committees served on	Interval
Q32A2	Number of graduate committees served on	Interval
Q32B1	Number of undergraduate committees chaired	Interval
Q32B2	Number of graduate committees chaired	Interval
Q33	Total classes taught	Interval
Q40	Total credit classes taught	Interval
Q50	Total contact hours/week with students	Interval
Q51	Total office hours/week	Interval
Q52	Any creative work/writing/research	Categorical
Q54_55RE	PI / Co-PI on grants or contracts	Ordinal
Q58	Total number of grants or contracts	Interval
Q59A	Total funds from all sources	Ratio
Q61SREC	Work support availability	Ordinal
Q64	Union status	Categorical

Table 1 Continued.

Variable name	Variable definition	Scale
Q76G	Consulting/freelance income	Ratio
Q7REC	Years on current job	Interval
Q80	Number of dependents	Interval
Q81	Gender	Categorical
Q85	Disability	Categorical
Q87	Marital status	Categorical
Q90	Citizenship status	Categorical
Q9REC	Years on achieved rank	Interval
X01_3	Principal activity	Categorical
X01_60	Overall quality of research index	Ordinal
X01_66	Job satisfaction: other aspects of job	Ordinal
X01_82	Age	Interval
X01_8REC	Academic rank	Ordinal
X01_91RE	Highest educational level of parents	Ordinal
DISCIPLINE	Principal field of teaching/researching	Categorical
X02_49	Individual instruction w/grad &1st professional students	Interval
X03_49	Number of students receiving individual instructions	Interval
X04_0	Carnegie classification of institution	Categorical
X04_41	Total classroom credit hours	Interval
X04_84	Ethnicity in single category	Categorical
X08_0D	Doctoral, 4-year, or 2-year institution	Ordinal

Table 1 Continued.

Variable name	Variable definition	Scale
X08_0P	Private or public institution	Categorical
X09_0RE	Degree of urbanization of location city	Ordinal
X09_76	Total income not from the institution	Ratio
X10_0	Ratio: FTE enrollment / FTE faculty	Ratio
X15_16	Years since highest degree	Interval
X21_0	Institution size: FTE graduate enrollment	Interval
X25_0	Institution size: Total FTE enrollment	Interval
X37_0	Bureau of Economic Analysis (BEA) regional codes	Categorical
X46_41	Undergraduate classroom credit hours	Interval
X47_41	Graduate and First professional classroom credit hours	Interval
SALARY	Basic academic year salary	Ratio

Note. All data were based on respondent' reported status during the 1998-99 academic year.

Analysis

Three different prediction models were constructed and compared through the analysis of NSOPF:99; each of them had a variable reduction procedure and a prediction model based on the selected measures. The first model, Model I, was a multiple regression model with variables selected through statistical data reduction techniques; Model II was a data mining BBN model with an embedded variable selection procedure. A combination model, Model III, was also a multiple regression model, but built on variables selected by the data mining BBN approach.

Model I. The first model started with variable reduction procedures that reduced the 90 NSOPF:99 variables (salary measure excluded) to a smaller group that can be efficiently manipulated by a multiple regression procedure, and resulted in an optimal regression model based on the selected variables. According to the compensation theory and characteristics of the current data set, basic salary of the academic year as the dependent variable was log-transformed to improve its linear relationship with candidate independent variables.

The variable reduction for Model I was completed in two phases. In the first phase, the dimensional structure of the variable space was examined with Exploratory Factor Analysis (EFA) and K-Means Cluster (KMC) analysis; based on the outcomes of the two techniques, variables were classified into a number of major dimensions. Because EFA measures variable relationships by linear correlation and KMC by Euclidian distance, only 82 variables on

dichotomous, ordinal, interval, or ratio scales were included. Two different techniques were used to scrutinize the underlying variable structure such that any potential bias associated with each of the individual approaches could be reduced.

In EFA, different factor extraction methods were tried and followed by both orthogonal and oblique rotations of the set of extracted factors. The variable grouping was determined based on the matrices of factor loadings: variables that had a minimum loading of .35 on the same factor were considered as belonging to the same group. In the KMC analysis, the number of output clusters usually needs to be specified. When the exact number of variable clusters is unknown, the results of other procedures (e.g., EFA) can provide helpful information for estimating a range of possible number of clusters. Then the KMC can be run several times, each time with a different number of clusters specified within the range. The multiple runs of the KMC can also help to reduce the chance of getting a local optimal solution. Because variables were separated into mutually exclusive clusters, the interpretation of cluster identity was based on variables that had short distance from the cluster seed (the centroid).

The results of the KMC analysis were compared with that of the EFA for similarities and differences. A final dimensional structure of the variable space was determined based on the consensus of the EFA and KMC outputs; each of the variable dimensions was labeled with a meaningful interpretation.

During the second phase, one variable was selected from each dimension. Because of the different clustering methods used, variables in the same dimension might not share linear relationships. Taking into consideration that the final model of the analysis was of linear prediction, a method of extracting variables that account for more salary variance was desirable. Thus, for each cluster, the log-transformed salary was regressed on the variables within that cluster, and only one variable was chosen that associated with the greatest partial R^2 change.

Variables that did not show any strong relationships with any of the major groups, along with multilevel nominal variables that could not be classified, were carried directly into the second stage of multiple regression modeling as candidate predictors and tested for their significance. Nominal variables were recoded into binary variables and possible interactions among the predictor variables were checked and included in the model if significant. Both forced entry and stepwise selection were used to search for the optimal model structure; if any of the variables was significant in one variable selection method, but nonsignificant in the other, a separate test on the variable was conducted in order to decide whether to include the variable in the final regression model. Finally, the proposed model was cross-checked with All Possible Subsets regression techniques including Max R and C_p evaluations to make sure the model was a good fit in terms of the model R^2, adjusted R^2, and the C_p value.

Model II. The second prediction model was a BBN-based data mining model. To build the BBN model, all 91 original variables were input into a piece of software called the Belief Network Powersoft ; variables on interval and ratio scales were binned into category-like intervals because the network-learning algorithms require discrete values for a clear definition of a finite product state space of the input variables. Rather than logarithmical transformation, salary was binned into 24 intervals for the following reasons: first, log-transformation was not necessary because BBN is a robust nonmetric algorithm independent of any monotonic variable transformation. And second, a finite number of output classes is required in a Bayesian network construction. During the modeling process, variable selection was performed internally to find the subset with the best prediction accuracy.

The BBN model learning was an automated process after reading in the input data. According to Chen and Greiner (1999), the authors of the software, two major tasks in the process are learning the graphical structure (variable relationships) and learning the parameters (CP tables). Learning the structure is the most computationally intensive task. The BBN software used in this study takes the network structure as a group of CP relationships (measured by statistical functions such as χ^2 statistic or mutual information test) connecting

the variables, and proceeds with the model construction by identifying the CPs that are stronger than a specified threshold value.

The output of the BBN model was a network in which the nodes (variables) were connected by arcs (CP relationships between variables) and a table of CP entries (probability) for each arc. Only the subset of variables that was evaluated as having the best prediction accuracy stayed in the network. The prediction accuracy was measured by the percentage of correct classifications of all observations in the data set.

Model III. Finally, a combination model was created that synchronized data mining and statistical techniques: the variables selected by the data mining BBN model were put into a multiple regression procedure for an optimal prediction model. The final BBN model contained a subset of variables that was expected to have the best prediction accuracy. Once the BBN model was available, the variables in that model were put through a multiple regression procedure for another prediction model. If it results in a better model, it would be evident that BBN could be used together with traditional statistical techniques when appropriate. As in Model I, categorical variables were recoded and salary as the dependent variable was log-transformed. Multiple variable selection techniques were used including forced entry and stepwise selection.

Model Comparison

The algorithms, input variables, final models, outputs, and interpretations of the three prediction models were presented. The two multiple regression models were comparable because they shared some common evaluation criteria, including the model standard error of estimate, residuals, R^2, and adjusted R^2. The data mining BBN model offered a different form of output, and is less quantitatively comparable with the regression models because they had little in common.

Software

SAS and SPSS were used for the statistical analyses. The software for learning the BBN model is called Belief Network Powersoft, a shareware developed and provided by Chen and Greiner (1999) on the World Wide Web. The Belief Network Powersoft was the winner of the yearly competition of the Knowledge Discovery and Data mining (KDD) – KDDCup 2001 Data Mining Competition Task One, for having the best prediction accuracy among 114 submissions from all over the world.

Results

Model I

The result of the variable space simplification through EFA and KMC was that 70 of the 82 variables were clustered into 17 groups. Ten of the groups were distinct clusters that did not seem to overlap with each other: academic rank, administrative responsibility, beginning work status, education level, institution parameter, other employment, research, teaching, experience, and work environment index. Another seven groups were 1) teaching: undergraduate committee, 2) teaching: graduate, 3) teaching: individual instruction, 4) publications: books, 5) publications: reviews, 6) publication: performances and presentations, and 7) institutional parameters: miscellaneous. In general, the dimensional structure underlying the large number of variables provided a schema of clustering similar measures and therefore made it possible to simplify the data modeling by means of variable extraction.

Following the final grouping of variables, one variable was extracted from each of the clusters by regressing the log-transformed salary on variables within the same cluster and selecting the variable that contributed the greatest partial R^2 change in the dependent variable. The 17 extracted variables, along with the 20 variables that could not be clustered, are listed in Table 2 as the candidate independent variables for a multiple regression model.

After a thorough model building and evaluation process, a final regression model was selected having 16 predictor variables (47 degrees of freedom due to binary-coded nominal measures) from the pool of 37 candidates. The parameter estimates and model summary information are in Tables 3 and 5. The model R^2 is .5036 and adjusted R^2 .5001.

Table 2. Candidate Independent Variables of Model I.

Variable name	Variable Definition	*df*
	Variables from the clusters	
Q29A1	Career creative works, juried media	1
X15_16	Years since highest degree	1
Q31A1	Time actually spent teaching undergraduates (percentage)	1
Q31A2	Time actually spent at teaching graduates (percentage)	1
X02_49	Individual instruction w/grad &1st professional students	1
Q32B1	Number of undergraduate committees chaired	1
Q31A5	Time actually spent at administration (percentage)	1
Q16A1REC	Highest degree type	1
Q24A5REC	Rank at hire for 1st job in higher education	1
Q29A3	Career reviews of books, creative works	1
Q29A5	Career presentations, performances	1
X08_0D	Doctoral, 4-year, or 2-year institution	1
Q29A4	Career books, textbooks, reports	1
X10_0	Ratio: FTE enrollment / FTE faculty	1
Q76G	Consulting/freelance income	1
X01_66	Job satisfaction: other aspects of job	1
X01_8REC	Academic rank	1

Table 2 Continued.

Variable name	Variable definition	*df*
	Variables from the original set	
DISCIPLINE	Principal field of teaching/research	10
Q12A	Appointments: Acting	1
Q12E	Appointments: Clinical	1
Q12F	Appointments: Research	1
Q19	Current position as primary employment	1
Q26	Positions outside higher education during career	1
Q30B	Hours/week unpaid activities at the institution	1
Q31A4	Time actually spent on professional growth (percentage)	1
Q31A6	Time actually spent on service activity (percentage)	1
Q64	Union status	3
Q80	Number of dependents	1
Q81	Gender	1
Q85	Disability	1
Q87	Marital status	3
Q90	Citizenship status	3
X01_3	Principal activity	1
X01_91RE	Highest educational level of parents	1
X04_0	Carnegie classification of institution	14
X04_84	Ethnicity in single category	3
X37_0	Bureau of Economic Analysis (BEA) region code	8

Table 3. Parameter Estimates of Model I.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
Intercept	Intercept	10.0399	0.0485	207.10	<.0001
Q29A1	Career creative works, juried media	0.0019	0.0002	11.87	<.0001
X15_16	Years since highest degree	0.0077	0.0004	17.82	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0011	0.0002	-6.04	<.0001
Q31A5	Time actually spent at administration (%)	0.0017	0.0003	5.95	<.0001
Q16A1REC	Highest degree type	0.0841	0.0050	16.68	<.0001
Q29A3	Career reviews of books, creative works	0.0018	0.0004	4.22	<.0001
Q76G	Consulting/freelance income	0.0000037	0.0000	5.75	<.0001
X01_66	Other aspects of job	0.0519	0.0058	8.89	<.0001
X01_8REC	Academic rank	0.0510	0.0031	16.27	<.0001
Q31A4	Time actually spent on professional growth (%)	-0.0023	0.0006	-3.86	0.0001
Q31A6	Time actually spent on service activity (%)	0.0013	0.0003	3.80	0.0001
Q81	Gender	-0.0667	0.0084	-7.97	<.0001
<u>BEA region codes (Baseline: Far West)</u>					
BEA1	New England	-0.0608	0.0058	8.89	0.0021
BEA2	Mid East	0.0082	0.0031	16.27	0.5788
BEA3	Great Lakes	-0.0545	0.0006	-3.86	0.0001
BEA4	Plains	-0.0868	0.0003	3.80	<.0001

Table 3 Continued.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
BEA5	Southeast	-0.0921	0.0084	-7.97	<.0001
BEA6	Southwest	-0.0972	0.0198	-3.07	<.0001
BEA7	Rocky Mountain	-0.1056	0.0148	0.56	<.0001
BEA8	U.S. Service schools	0.1480	0.0142	-3.82	0.2879
Principal field of teaching/research (Baseline: legitimate skip)					
DSCPL1	Agriculture & home economics	-0.0279	0.0306	-0.91	0.3624
DSCPL2	Business	0.1103	0.0228	4.84	<.0001
DSCPL3	Education	-0.0643	0.0216	-2.98	0.0029
DSCPL4	Engineering	0.0695	0.0246	2.82	0.0048
DSCPL5	Fine arts	-0.0449	0.0241	-1.86	0.0627
DSCPL6	Health sciences	0.0933	0.0182	5.12	<.0001
DSCPL7	Humanities	-0.0641	0.0195	-3.29	0.001
DSCPL8	Natural sciences	-0.0276	0.0190	-1.45	0.148
DSCPL9	Social sciences	-0.0249	0.0202	-1.23	0.2173
DSCPL10	All other programs	0.0130	0.0194	0.67	0.502
Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	0.0053	0.0236	0.22	0.8221
STRATA2	Private comprehensive	-0.0377	0.0263	-1.43	0.1525
STRATA3	Public liberal arts	-0.0041	0.0341	-0.12	0.9039
STRATA4	Private liberal arts	-0.0917	0.0260	-3.52	0.0004

Table 3 Continued.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
STRATA5	Public medical	0.2630	0.0326	8.07	<.0001
STRATA6	Private Medical	0.2588	0.0444	5.82	<.0001
STRATA7	Private religious	-0.1557	0.0523	-2.98	0.0029
STRATA8	Public 2-year	0.0386	0.0247	1.56	0.1185
STRATA9	Private 2-year	-0.0061	0.0574	-0.11	0.9155
STRATA10	Public other	-0.0207	0.0563	-0.37	0.7127
STRATA11	Private other	-0.0879	0.0428	-2.06	0.0399
STRATA12	Public research	0.0792	0.0228	3.47	0.0005
STRATA13	Private research	0.1428	0.0259	5.51	<.0001
STRATA14	Public other Ph.D.	0.0005	0.0254	0.02	0.984
Primary activity (Baseline: others)					
PRIMACT1	Primary activity: teaching	-0.0541	0.0169	-3.21	0.0013
PRIMACT2	Primary activity: research	-0.0133	0.0199	-0.67	0.5039
PRIMACT3	Primary activity: administration	0.0469	0.0203	2.31	0.0211

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Model II

To make the findings of the data mining BBN model comparable to the result of regression Model I, the second model started without any pre-specified knowledge such as the order of variables in some dependence relationships, forbidden relations, or known causal relations. To evaluate variable relationships and simplify model structure, the data mining software makes it possible for users to provide a threshold value that determines how strong a mutual relationship between two variables is considered meaningful; relationships below this threshold are omitted from subsequent network structure learning (Chen & Greiner, 1999).

In the current analysis, a number of BBN learning processes were completed, each with a different threshold value specified, in order to search for an optimal model structure. Because generalizability to new data sets is an

important property of any prediction models, the model parameters were cross-validated with the testing data set. The results suggested that the model of best prediction power was the one having six variables connected by 10 CP arcs as shown in Figure 2. The prediction accuracy, quantified as the percentage of correct classification of the cases, was 25.66% for training data and 11.57% for testing data.

Model III

The final prediction model produced by the Belief Network Powersoft had six predictor variables. However, one of six, number of years since achieved tenure (Q10AREC), was only connected to another predictor variable (i.e., years since the highest degree), a strong relationship substantiated by their Pearson correlation ($r = .64$). Q10AREC also had a strong correlation with academic rank ($r = .43$), another variable in the model. After a test confirmed that Q10AREC was not a suppressor variable, it was excluded from the combination model. Therefore, Model III started with only five independent variables. Among them, the Carnegie classification of institutions as the only categorical measure was recoded into binary variables. With log-transformed salary as the dependent variable, the process of building Model III was straightforward because all five variables were significant at $p < .0001$ with both forced entry and stepwise variable selections. The model has R^2 of .4214 and adjusted R^2 .4199 (summary information is presented in Tables 4 and 5).

Model Comparison

Model I and Model II are comparable in many ways. First, both models are result of data-driven procedures; second, theoretically, they both selected the predictors from the original pool of 90 variables; and third, they share the same group of major variables even though Model I had a much larger group. With the common ground they share, the differences between the two models provide good insight to the differences between traditional statistics and data mining BBN in make predictions with large-scale data sets.

The differences between Model I and Model III are informative about the effects of statistical and data mining approaches in simplifying the variable space and identifying the critical measures in making accurate prediction, given both models used multiple regression for the final prediction. Models II and III share the same group of predictor variables; their similarities and differences shed light on the model presentations and prediction accuracy of different approaches as well.

Variable Selection and Transformation

Model I started with all 90 variables in the pool, and identified 17 of the 70 variables that could be clustered with EFA and KMC procedures. Along with the ungrouped 20 variables, a total of 37 independent variables were available as initial candidates, and 16 of them stayed in the final model with an R^2 of .5036 (df = 47 and adjusted $R^2 = .5001$). With a clear goal of prediction, the modeling process was exploratory without theoretical considerations from variable reduction through model building. During this process, variable relationships were measured as linear correlations; consequently, the dependent variable was transformed to improve its linear relationships with the independent variables. Also, multilevel categorical measures were recoded into binary variables.

The data mining model, Model II, also started with all 90 variables. An automated random search was performed internally to select a subset of variables that provided the most accurate salary prediction. In contrast to regression models that explicitly or implicitly recode categorical data, data mining models usually keep the categorical variables unchanged, but bin continuous variables into intervals. The information loss associated with variable downgrade in binning is a threat to model accuracy, but it helps to relax model assumptions and as a result BBN requires no linear relationships among variables. The network structure discovery uses some statistical tests (e.g., χ^2 test of statistical independence) to compare how frequently different values of two variables are associated with how likely they happen to be together by random chance in order to build conditional probability statistics among variables (Chen, Greiner, Kelly, Bell, & Liu, 2001).

Figure 2. The BBN model of salary prediction. Some of the directional relationships may be counterintuitive (e.g., Q31A1 → X04_0) as a result of data-driven learning. The CP tables are not included to avoid complexity. The definitions of the seven variables are

a. SALARY: Basic salary of the academic year.
b. Q29A1: Career creative works, juried media
c. Q31A1: Percentage of time actually spent teaching undergrads
d. X15_16: Years since highest degree
e. X01_8REC: Academic rank
f. X04_0: Carnegie classification of institutions
g. Q10AREC: Years since achieved tenure

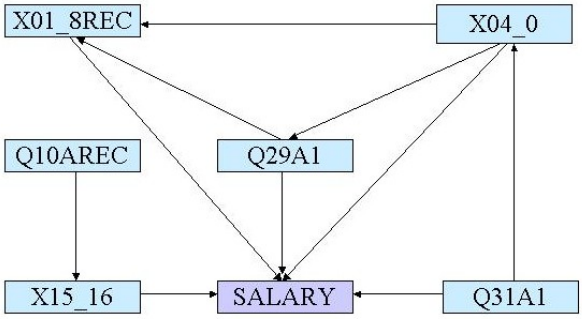


Table 4. Parameter Estimates of Model III.

Variable	Label	Parameter estimate	Standard error	*t* value	*p* > |t|
Intercept	Intercept	10.5410	0.0272	387.28	<.0001
Q29A1	Career creative works, juried media	0.0024	0.0002	15.34	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0030	0.0002	-20.06	<.0001
X01_8REC	Academic rank	0.0664	0.0032	21.01	<.0001
X15_16	Years since highest degree	0.0088	0.0004	19.97	<.0001

Table 4 Continued.

Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	-0.0385	0.0250	-1.54	0.1236
STRATA2	Private comprehensive	-0.0645	0.0281	-2.29	0.0218
STRATA3	Public liberal arts	-0.0315	0.0363	-0.87	0.3853
STRATA4	Private liberal arts	-0.1221	0.0276	-4.42	<.0001
STRATA5	Public medical	0.2933	0.0339	8.66	<.0001
STRATA6	Private Medical	0.2915	0.0471	6.20	<.0001
STRATA7	Private religious	-0.2095	0.0551	-3.80	0.0001
STRATA8	Public 2-year	-0.0403	0.0258	-1.56	0.1179
STRATA9	Private 2-year	-0.0371	0.0611	-0.61	0.544
STRATA10	Public other	-0.0245	0.0594	-0.41	0.6802
STRATA11	Private other	-0.0871	0.0456	-1.91	0.0563
STRATA12	Public research	0.0479	0.0242	1.98	0.0472
STRATA13	Private research	0.1543	0.0276	5.60	<.0001
STRATA14	Public other Ph.D.	-0.0496	0.0268	-1.85	0.0648

Variable	Label	Parameter estimate	Standard error	*t* value	$p > \|t\|$

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Table 5. Summary Information of Multiple Regression Models I and III

Source	*df*	Sum of squares	Mean square	F	Pr > F
Model I: Multiple regression with statistical variable selection					
Model	47	621.4482	13.2223	142.46	<.0001
Error	6599	612.4897	0.0928		
Corrected total	6646	1233.9379			

Table 5 Continued.

Source	*df*	Sum of Squares	Mean square	F	Pr > F
Model III: Multiple regression with variables selected through BBN					
Model	18	520.2949	28.90527	268.4	<.0001
Error	6632	714.3279	0.10769		
Corrected total	6651	1234.6228			

Note:
1. For Model I, R^2 = .5036, adjusted R^2 = .5001, and the standard error of estimate is 0.305.
2. For Model II, R^2 = .4214, adjusted R^2 = .4199 and the standard error of estimate is 0.328

Given the measures of variable associations that do not assume any probabilistic forms of variable distributions, neither linearity nor normality was required in the analysis. Consequently, the non-metric algorithms used to build the BBN model binned the original SALARY measure as the predicted values.

Model Selection

In the multiple regression analysis, every unique combination of the independent variables theoretically makes a candidate prediction model, albeit the modeling techniques produce candidate models that are mostly in a nested structural schema. Model comparison is part of the analysis process; human intervention is necessary to select the final model that usually has a higher R^2 along with simple and stable structure. In contrast, the learning of an optimal BBN model is a result of search in a model space that consists of candidate models of substantially different structures. In the automated model discovery process, numerous candidate models were constructed, evaluated with criteria called score functions, and the one with best prediction accuracy is output as the optimal choice.

Model Presentation

As a result of different approaches to summarizing data and different algorithms of analyzing data, the outputs of the multiple regression and the BBN models are different. The final result of a multiple regression analysis is usually presented as a mathematical equation. For example, Model III can be written as:

Log (Salary) = 10.5410 + 0.0024 × Q29A1 - 0.0030 × Q31A1 + 0.0664 × X01_8REC + 0.0088 × X15_16 - 0.0385 × STRATA1 - 0.0645 × STRATA2 - 0.0315 × STRATA3 - 0.1221 × STRATA4 + 0.2933 × STRATA5 + 0.2915 × STRATA6 - 0.2095 × STRATA7 - 0.0403 × STRATA8 - 0.0371 × STRATA9 - 0.0245 × STRATA10 - 0.0871 × STRATA11 + 0.0479 × STRATA12 + 0.1543 × STRATA13 - 0.0496 × STRATA14 + error. (1)

If a respondent received the highest degree three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of work time teaching undergraduate classes (Q31A1 = 20) as an assistant professor (X01_8REC = 4) in a public research institution (STRATA12 = 1 and all other STRATA variables were 0), the predicted value of this individual's log-transformed salary should be 10.83 according to Equation 1 (about $50,418), with an estimated standard error indicating the level of uncertainty.

The result of the BBN model is presented in a quite different way. For the above case, the BBN model would make a prediction

of salary for such faculty with a salary conditional probability table as shown in Table 6. The predicted salary fell in a range between $48,325 and $50,035 because it has the highest probability (p= 15.9%) in the CP table for this particular combination of variable values. A CP table like this is available for every unique combination of variable values (i.e., an instance in the variable product state space).

Using the conditional mean as a point estimator in most statistical predictions implicitly expresses the prediction uncertainty with a standard error of estimate based on the assumption of normal distribution. In contrast, the BBN model makes predictions based on the distributional mode of the posterior probability of the predicted variable. The prediction based on the mode of a probabilistic distribution is a robust feature of BBN; the mode is not sensitive to outliers or skewed distribution as the arithmetic mean is. Moreover, the presentation of posterior probability as a random variable explicitly expresses the prediction uncertainty in terms of probability. Without the assumption of normality, the conditional probability of a predicted value is the outcome of binning continuous variables and treating all variables as on a nominal scale in the computation. However, one problem of the classification approach is that it is difficult to tell how far the predicted value missed the observed value when a case was misclassified.

Prediction Accuracy

In multiple regression, predication accuracy is usually quantified by residuals or studentized residuals. Also, the model R^2 is an index of how well the model fits the data. For example, Model III had a R^2 of .4214, which was considered an acceptable level of explained variance in regression given such a complex data set. The predication accuracy of the BBN model was the ratio of the number of correct classifications to the total number of predictions. In this study, the prediction accuracy of the BBN model was only 25.66% on the same training data.

Several explanations are available for this relatively low prediction accuracy of Model II compared to that of Model III. First, information was lost when continuous variables were binned: five of the six predictors were on an interval or ratio scale. Second, the final class identity of an individual case was algorithmically determined to be the salary bin that had the highest probability, which might not be substantially strong when the predictor variable was divided into many narrow bins (as in the above example p = .16). Third, when the bin widths are relatively narrow, misclassification may increase due to weakened differences among the levels of a variable. Finally, scoring functions used for model evaluation in the Bayesian network learning could be another factor. According to Friedman et al. (1997), when the structure of the network is not constrained with any prior knowledge as in the current case, nonspecialized scoring functions may result in a poor classifier function when there are many attributes.

Dimensional Simplification

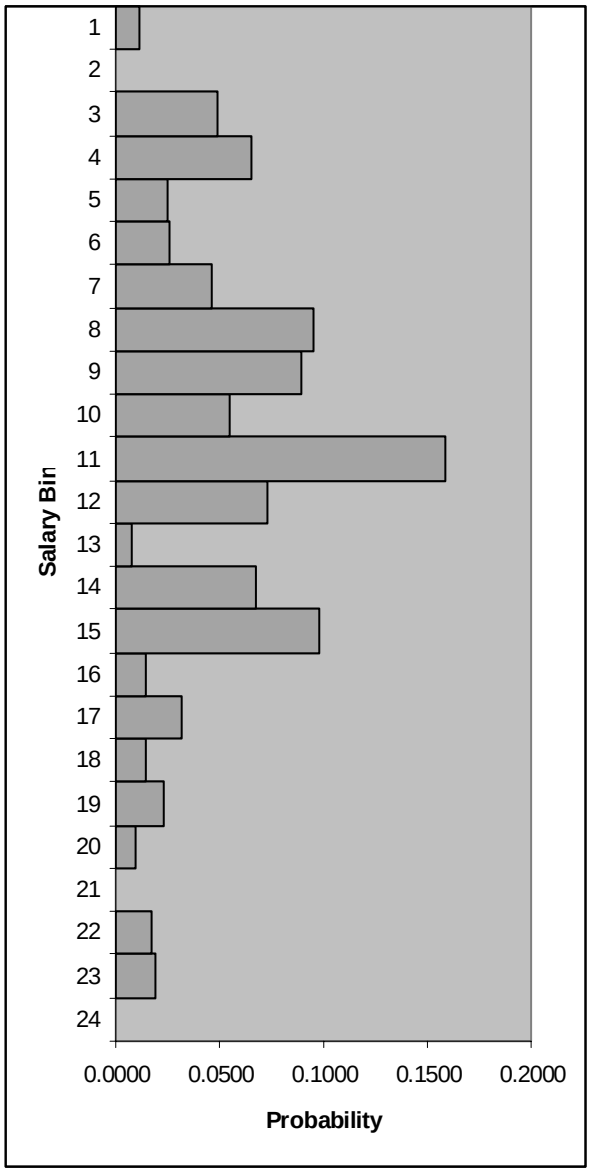
One important similarity between Models I and III is the final predictor variables. Model III had only five variables selected by the BBN model, and they were among the top six variables in the stepwise selection of Model I. Both models captured variables that shared strong covariance with the predicted variable. The overlap of the predictor variables is an indication that they both can serve the purpose of dimensional simplification.

In comparison to the automated process of variable selection and dimensional simplification in the BBN algorithms, the statistical approach was relatively laborious. However, the automation in BBN learning blinded researchers from having a detailed picture of variable relationships. In the statistical variable reduction, the clustering structure of variables was clear, and so were the variables that were similar or dissimilar to each other. Therefore, the high automation is only desirable when the underlying variable relationships are not of concern, or when the number of variables is extremely large.

The BBN data mining Model II identified five predictor variables that were subsequently used in Model III for prediction, all five independent variables were significant at $p < 0.0001$, and resulted in a final model with an

Table 6. An Example of the BBN Conditional Probability Tables.

Bin #	Salary range	Probability
1	Salary < 29600	0.0114
2	29600 < Salary < 32615	0.0012
3	32615 < Salary < 35015	0.0487
4	35015 < Salary < 37455	0.0655
5	37455 < Salary < 39025	0.0254
6	39025 < Salary < 40015	0.0263
7	40015 < Salary < 42010	0.0460
8	42010 < Salary < 44150	0.0950
9	44150 < Salary < 46025	0.0894
10	46025 < Salary < 48325	0.0552
11	48325 < Salary < 50035	0.1590
12	50035 < Salary < 53040	0.0728
13	53040 < Salary < 55080	0.0081
14	55080 < Salary < 58525	0.0672
15	58525 < Salary < 60010	0.0985
16	60010 < Salary < 64040	0.0140
17	64040 < Salary < 68010	0.0321
18	68010 < Salary < 72050	0.0142
19	72050 < Salary < 78250	0.0228
20	78250 < Salary < 85030	0.0098
21	85030 < Salary < 97320	0.0005
22	97320 < Salary < 116600	0.0170
23	116600 < Salary < 175090	0.0190
24	175090 < Salary	0.0005



Note. Salary was binned into 24 intervals. For this particular case, the product state is that the highest degree was obtained three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of the time teaching undergraduate classes (Q31A1 = .2) as an untenured (Q10AREC = 0) assistant professor (X01_8REC = 5) in a public research institution (STRATA = 12 and all other binary variables were 0).

R^2 = .4214 (df = 18 and adjusted R^2 = .4199). Although Model I has a greater R^2 than Model III, it also has more model degrees of freedom (47 vs. 18). Given an R^2 about .0822 higher than that of Model III at the expense of 29 more variables, each additional variable in Model I only increased the model R^2 by .0028 on average.

One of the negative effects associated with large numbers of independent variables in a multiple regression model is the threat of multicollinearity caused by possible strong correlations among the predictors. Model R^2 never decreases when the number of predictor variables increases, but if the variables bring along multicollinearity, estimated model parameters can have large standard errors, leading to an unreliable model. For the two regression models, Model I has 31 out of 47 variable with a VIF > 1.5 (66%). Model II has 10 out of 18 variables with a VIF > 1.5 (55%), and most of high VIF values are associated with the binary variables recoded from categorical variables.

Because the ordinary least square (OLS) method in prediction analysis produces a regression equation that is optimized for the training data, model generalizability should be considered as another important index of good prediction models. Model generalizability was measured by cross validating the proposed models with the holdout testing data set. Model I and III were applied to the 3,311 records to obtain their predicted values, and the R^2s of the testing data set were found to be .5055, and .4489, respectively, as compared with .5036 and .4214 in the original data set.

Large Data Volume

Multiple regression models have some problems when applied to massive data sets. First, many graphical procedures, including scatter plots for checking variable relationships become problematic when the large number of observations turns the plots into indiscernible black clouds. Second, with a large number of observations the statistical significance tests are oversensitive to minor differences. For example, a few variables with extremely small partial R^2s had significant p values in the stepwise selection of Model I. One particular case was the union status, which had a partial R^2 = .0009, given a sample size of 6,652, the variable was still added at a significant p = 0.0073.

Data mining models usually respond to large samples positively due to their inductive learning nature. Data mining algorithms rarely use significance tests, but rely on the abundant information in large samples to improve the accuracy of the rules (descriptions of data structure) summarized from the data. In addition, more data are needed to validate the models and to avoid optimistic bias (overfit).

Conclusion

In the field of education, large data sets recorded in the format of computer databases range from student information in a school district to national surveys of some defined population. Although data are sometimes collected without predefined research concerns, they become valuable resources of information for collective knowledge that can inform educational policy and practice. The critical step is how to effectively and objectively turn the data into useful information and valid knowledge. Educational researchers have not been able to take full advantage of those large data sets, partly because data sets of very large volume have presented practical problems related to statistical and analytical techniques.

The objective of this article is to explore the potentials of using data mining techniques in studying large data sets or databases in educational research. Data analysis methods that can effectively handle a large number of variables is one of the major concerns in this study of 91 variables (one was salary, the predicted variable).

The major findings are as follows. The multiple regression models were cumbersome with a large number of independent variables. Although the loss of degrees of freedom was not a concern given a large sample size, a thorough examination of variable interactions became unrealistic. The data mining model BBN needed much less human intervention in its automated learning and selection process. With the BBN algorithm inductively studying and summarizing variable relationships without probabilistic assumptions, the defense against normality and

linearity was dismissed, and significance tests were rarely necessary. However, the BBN model had some drawbacks as well. First, the BBN model, as most data mining models, is adaptive to categorical variables. Continuous measures had to be binned to be appropriately handled. The downgrade of measurement scale definitely cost information accuracy.

It also became clear in the process of this study that the ability to identify the most important variable from a group of highly correlated measures is an important criterion for evaluating applied data analysis methods when handling a large number of variables because redundant measures on the same constructs are common in large data sets and databases. The findings of this study indicate that BBN is capable to perform such a task because Model II identified five variables from groups of measures on teaching, publication, experience, academic seniority, and institution parameter, the same five as those selected by the data reduction techniques in building Model I for the reason that the five variables accounted for more variance of the predicted variables than their alternatives.

In general, data mining has some unique features that can help to explore and analyze enormous amount of data. Combining statistical and machine learning techniques in automated computer algorithms, data mining can be used to explore very large volumes of data with robustness against poor data quality such as nonnormality, outliers, and missing data. The inductive nature of data mining techniques is very practical to overcome limitations of traditional statistics when dealing with large sample sizes. The random selection of subset variables in making accurate predictions simplifies the problem associated with large number of variables. Nevertheless, the applicability of this new technique in educational and behavioral science has to be tailored for the specific needs of individual researchers and the goal of their studies.

By introducing data mining, a tool that has been widely used in business management and scientific research, this study demonstrated an alternative approach to analyzing educational databases. A clear-cut answer is difficult regarding the differences and advantages of the individual approaches. However, looking at a problem from different viewpoints itself is the essence of the study, and hopefully it can provide critical information for researchers to make their own assessment about how well these different models work to provide insight into the structure of and to extract valuable information from large volumes of data. Using confirmatory analysis to follow up the findings generated by data mining, educational researchers can virtually turn their large collection of data into a reservoir of knowledge to serve public interests.

References

Chen, J., & Greiner, R. (1999). Comparing Bayesian network classifiers. *Proceedings of the Fifteenth Conference on Uncertainty In Artificial Intelligence (UAI), Sweden*, 101-108.

Chen, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2001). Learning Bayesian networks from data: An information-theory based approach. *Artificially Intelligence, 137*(1-2), 43-100.

Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Representative subset selection. *Analytical Chimica Acta, 468*(1), 91-103.

Elder, J. F. & Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, R. Smyth, & R. Uthurusamy (Eds.) *Advances in knowledge discovery and data mining* (pp.83-113). Menlo Park, California: AAAI Press.

Fayyad, U. M. (1997, August). *Data mining and knowledge discovery in databases: Implications for scientific databases.* Papered presented at 9th International Conference on Scientific and Statistical Database Management (SSDBM'97), Olympia, WA**.**

Frawley, W. J., Piatetsky-Shapiro, G., & Matheu, C. J. (1991). Knowledge discovery in database: An overview. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.) *Knowledge Discovery in Databases* (pp. 1-27). MIT: AAAI Press.

Friedman, J. H. (1997). Data mining and statistics: What's the connection? In D. W. Scott (Ed.), *Computing Science and Statistics*: *Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 3-9). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Glymour, C., Madigan, D., Pregibon, D. & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery, 1*, 11-28.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52*, 112-118.

Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *SIGKDD Exploration, 1*, 16-19.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discover, 1*, 79-119.

Michalski, R. S., Bratko, I., & Kubat, M. (1998). *Machine learning and data mining: Methods and applications.* Chichester: John Wiley & Sons.

National Center of Education Statistics. (2002). National Survey of Postsecondary Faculty 1999 (NCES Publication No. 2002151), [Restricted-use data file, CD-ROM]. Washington, DC: Author.

Niedermayer, D. (1998). *An introduction to Bayesian networks and their contemporary applications.* Retrieved on September 24, 2003 from http://www.niedermayer.ca/papers/bayesian/

Thearling, K. (2003). *An Introduction to Data Mining: Discovering hidden value in your data warehouse.* Retrieved on July 6, 2003 from http://www.thearling.com/text/dmwhite/ dmwhite.htm.

Wegman, E., J. (1995). Huge data sets and the frontiers of computational feasibility. *Journal of Computational and Graphical Statistics, 4*(4), 281-295.

Yu, Y., & Johnson, B. W. (2002). *Bayesian belief network and its applications* (Tech. Rep. UVA-CSCS-BBN-001). Charlottesville, VA: University of Virginia, Center for Safety-Critical Systems.

Zhou, Z. (2003). Three perspectives of data mining. *Artificial Intelligence*, *143*(1), 139-146.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 275-282

1538 – 9472/05/$95.00

Manifestation Of Differences In Item-Level Characteristics In Scale-Level Measurement Invariance Tests Of Multi-Group Confirmatory Factor Analyses

Bruno D. Zumbo
University of British Columbia, Canada

Kim H Koh
Nanyang Technological University, Singapore

If a researcher applies the conventional tests of scale-level measurement invariance through multi-group confirmatory factor analysis of a PC matrix and MLE to test hypotheses of strong and full measurement invariance when the researcher has a rating scale response format wherein the item characteristics are different for the two groups of respondents, do these scale-level analyses reflect (or ignore) differences in item threshold characteristics? Results of the current study demonstrate the inadequacy of judging the suitability of a measurement instrument across groups by only investigating the factor structure of the measure for the different groups with a PC matrix and MLE. Evidence is provided that item level bias can still be present when a CFA of the two different groups reveals an equivalent factorial structure of rating scale items using a PC matrix and MLE.

Key words: multi-group confirmatory factor analysis, item response formats

Introduction

Broadly speaking, there are two general classes of statistical and psychometric techniques to examine measurement invariance across groups: (1) scale-level analyses, and (2) item-level analyses. The groups investigated for measurement invariance are typically formed by gender, ethnicity, or translated/adapted versions of a test. In scale-level analyses, the set of items comprising a test are often examined together using multi-group confirmatory factor analyses (Byrne, 1998; Jöreskog, 1971) that involve testing strong and full measurement invariance hypotheses. In the item-level analyses the focus is on the invariant characteristics of each item, one item at a time.

In setting the stage for this study, which involves a blending of ideas from scale- and item-level analyses (i.e., multi-group confirmatory factor analysis and item response theory), it is useful to compare and contrast overall frameworks for scale-level and item-level approaches to measurement invariance. Recent examples of this sort of comparison can be found in Raju, Laffitte, & Byrne (2002), Reise, Widaman, & Pugh (1993), and Zumbo (2003). In these studies, the impact of scaling on measurement invariance has not been examined. Hence, it is important for the current study to investigate to what extent the number of scale points effects the tests of measurement invariance hypotheses in multi-group confirmatory factor analysis.

Scale-level Analyses

There are several expositions and reviews of single-group and multi-group confirmatory factor analysis (e.g., Byrne, 1998; Steenkamp & Baumgartner, 1998; Vandenberg

Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as member of the Department of Statistics and the Institute of Applied Mathematics at the University of British Columbia, Canada Email: bruno.zumbo@ubc.ca. Kim H. Koh is Assistant Professor, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore. Email: khkoh@nie.edu.sg. An earlier version of this article was presented at the 2003 National Council on Measurement in Education (NCME) conference, Chicago Illinois. We would like to thank Professor Greg Hancock for his comments on an earlier draft of this article.

& Lance, 2000); therefore this review will be very brief. In describing multi-group confirmatory factor analysis, consider a one-factor model: one latent variable and ten items all loading on that one latent variable. There are two sets of parameters of interest in this model: (1) the factor loadings corresponding to the paths from the latent variable to each of the items, and (2) the error variances, one for each of the items. The purpose of the multi-group confirmatory factor analysis is to investigate to what extent each, or both; of the two sets of model parameters (factor loadings and error variances) are invariant in the two groups.

As Byrne (1998) noted, there are various hypotheses of measurement invariance that can be tested, from weak to strict invariance. That is, one can test whether the model in its entirety is completely invariant, i.e., the measurement model as specified in one group is completely reproduced in the other, including the magnitude of the loadings and error variances. At the other end of the extreme is an invariance in which the only thing shared between the groups is overall pattern of the model but neither the magnitudes of the loadings nor of the error variances are the same for the two groups, i.e., the test has the same dimensionality, or configuration, but not the same magnitudes for the parameters.

Item-level Analyses

In item-level analyses, the framework is different than at the scale-level. At the item level, measurement specialists typically consider (a) one item at a time, and (b) a unidimensional statistical model that incorporates one or more thresholds for an item response. That is, the response to an item is governed by referring the latent variable score to the threshold(s) and from this comparison the item response is determined.

Consider the following example of a four-point Likert item, "How much do you like learning about mathematics?" The item responses are scored on a 4-point scale such as (1) Dislike a lot, (2) Dislike, (3) Like, and (4) Like a lot. This item, along with other items, serve as a set of observed ordinal variables, *x*'s, to measure the latent continuous variable *x**, namely attitudes toward learning mathematics. For each observed ordinal variable *x*, there is an underlying continuous variable *x**. If *x* has *m* ordered categories, *x* is connected to *x** through the non-linear step function: $x = i$ if

$$\tau_{i-1} < x^* \leq \tau_{i,} \quad i = 1,2,3,...,m,$$

where

$$\tau_0 = -\infty, \tau_1 < \tau_2 < \tau_3 < ...\tau_{m-1},$$

and $\tau_m = +\infty$

are parameters called threshold values. For a variable *x* with *m* categories, there are *m-1* unknown thresholds. Given that the above item has four response categories, there are three thresholds with the latent continuous variable. If one approaches the item level analyses from a scale-level perspective, the item responding process is akin to the thresholds one invokes in computing a polychoric correlation matrix (Jöreskog & Sörbom, 1996).

In an item-level analysis measurement specialists often focus on differences in thresholds across the groups. That is, the focus is on determining if the thresholds are the same for the two groups. If studying an achievement or knowledge test, it should be asked if the items are equally difficult for the two groups, with the thresholds being used as measures of item difficulty (i.e., an item with a higher threshold is more difficult). These differences in thresholds are investigated by methods collectively called "methods for detecting differential item functioning (DIF)". In common measurement practice this sort of measurement invariance is examined, for each item, one item at a time, using a DIF detection method such as the Mantel-Haenszel (MH) test or logistic regression (conditioning on the observed scores), or methods based on item response theory (IRT).

The IRT methods investigate the thresholds directly whereas the non-IRT methods test the difference in thresholds indirectly by studying the observed response option proportions by using categorical data analysis methods such as the MH or logistic regression methods (see Zumbo & Hubley, 2003 for a review).

Although both item- and scale-level methods are becoming popular in educational and psychosocial measurement, many researchers are still recommending and using only scale-level methods such as multi-group confirmatory factor analysis (for example, see, Byrne, 1998; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). There are, of course, scale-level methods that allow one to incorporate and test for item threshold differences in multi-group confirmatory factor analysis; however, these methods are not yet widely used. Instead, the popular texts on structural equation modeling by Byrne as well as the widely cited articles by Steenkaump and Baumgartner, and Vandenberg and Lance focus on and instruct users of structural equation modeling on the use of Pearson covariance matrices and the Chi-squared tests for model comparison based on maximum likelihood estimation (For an example see Byrne, 1998, Chapter 8 on a description of multi-group methods and p. 239 of her text for a recommendation on using ML estimation with the type of data we are describing above).

The question that this article addresses is reflected in the title: Do Differences in Item-Level Characteristics Manifest Themselves in Scale-Level Measurement Invariance Tests of Multi-Group Confirmatory Factor Analyses? That is, if a researcher applies the conventional tests of scale-level measurement invariance through multi-group confirmatory factor analysis of a Pearson covariance matrix and maximum likelihood estimation to test hypotheses of strong and full measurement invariance when the researcher has the ordinal (often called Likert) response format described above, do these scale-level analyses reflect (or ignore) differences in item threshold characteristics? If one were a measurement specialist focusing on item-level analyses (e.g., an IRT specialist), another way of asking this question is: Does DIF, or other forms of lack of item parameter invariance such as item drift, manifest itself in construct comparability across groups?

The present study is an extension of Zumbo (2003). A limitation of his earlier work is that it focused on the population analogue and did not investigate, as in this, the pattern and characteristics of the statistical decisions over the long run; i.e., over many replications. We study the rejection rates for a test of the statistical hypotheses in multi-group confirmatory factor analysis.

Methodology

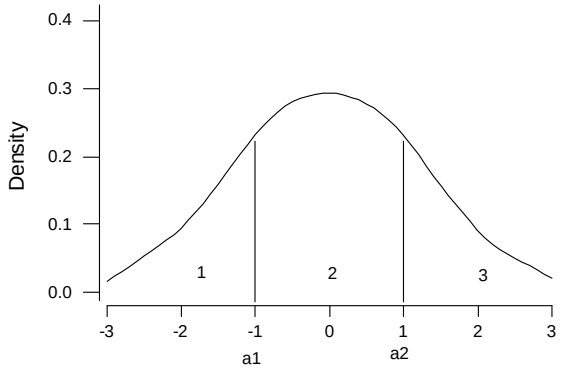
A computer simulation was conducted to investigate whether item-level differences in thresholds manifest themselves in the tests of strong and full measurement invariance hypotheses in multi-group CFA of a Pearson covariance matrix with maximum likelihood estimation.

Simulated was a one-factor model with 38 items. Obtained was a population covariance matrix based on the data reported in Zumbo (2000, 2003) that were based on the item characteristics of a sub-section of the TOEFL. Based on this covariance matrix, 100,000 simulees were generated on these 38 items with a multivariate normal distribution with marginal (univariate) means of zero and standard deviations of one. The simulation was restricted to a one-factor model because item-level methods (wherein differences in item thresholds, called DIF in that literature, is widely discussed) predominantly assume unidimensionality of their items, for example, IRT, MH, or logistic regression DIF methods.

The same item thresholds were used as those used by Bollen and Barb (1981) in their study of ordinal variables and Pearson correlation. In short, this method partitions the continuum ranging from –3 to +3. The thresholds are those values that divide the continuum into equal parts. The example in Figure 1 is a three-point scale using the notation described above for the x^* and x. Item thresholds were applied to these 38 normally distributed item vectors to obtain the ordinal item responses.

The simulation design involved two completed crossed factors: (i) number of scale points ranging from three to seven, and (ii) the percentage of items with different thresholds (i.e., percentage of DIF items) ranging from zero to 42.1 (1, 4, 8 and 16 items out of the total of 38).

Figure 1. A Three Category, Two Threshold x and its corresponding x*.



Note: Number of categories for x: 3 (values 1, 2, 3). Item thresholds for x*: a1, a2 (values of –1 and 1).

Three to seven item scale points were chosen because in order to only deal with those scale points for which Byrne (1998) and others suggest the use of Pearson covariance matrices with maximum likelihood estimation for ordinal item data. The resulting simulation design is a five by five completely crossed design.

The differences in thresholds were modeled based on suggestions from the item response theory (IRT) DIF literature for binary items. That is, the IRT DIF literature (e.g., Zumbo, 2003; Zwick & Ercikan, 1989) suggests that an item threshold difference of 0.50 standard deviations is a moderate DIF. This idea was extended and applied to each of the thresholds for the DIF item(s). For example, for a three-point item response scale group one would have thresholds of -1.0 and 1.0 whereas group two would have thresholds of –0.5 and 1.5. Note that for both groups the latent variables are simulated with a mean of zero and standard deviation of one. The same principle applies for the four to seven point scales.

Given that both groups have the same latent variable mean and standard deviation, the difference thresholds for the two groups (i.e., the DIF) would imply that the item(s) that is (are) performing differently across the two groups would have different item response distributions. It should be noted that the Bollen and Barb methodology results in symmetric Likert item responses that are normally distributed. The results in Table 1 allow one to compare the effect of having different thresholds in terms of the skewness and kurtosis.

The descriptive statistics reported in Table 1 were computed from a simulated sample of 100,000 continuous normal scores that were transformed with our methodology. For a continuous normal distribution the skewness and kurtosis statistics reported would both be zero. Focusing first on the skewness, it can be see in Table 1 that they range from -0.008 to 0.011 (with a common standard error of 0.008) indicating that, as expected, the Likert responses were originally near symmetrical. Applying the

Table 1. Descriptive Statistics of the Items without and without Different Thresholds.

# of Scale Points	Skewness		Kurtosis	
	Original	Different Thresholds	Original	Different Thresholds
3	-0.001	-0.004	0.144	-0.364
4	-0.008	0.125	-0.268	-0.294
5	0.011	0.105	-0.211	-0.277
6	-0.005	0.084	-0.185	-0.261
7	-0.003	0.082	-0.169	-0.238

Note: These statistics were computed from a sample of 100,000 responses using SPSS 11.5. In all cases, standard errors of the skewness and kurtosis were 0.008 and 0.015, respectively.

threshold difference, as described above, resulted in item responses that were nearly symmetrical for three, six, and seven scale points, and only small positive skew (0.125 and 0.105) for the four and five scale points. In terms of kurtosis, there is very little change with the different thresholds, except for the three-point scale that resulted in the response distribution being more platykurtic with the different thresholds.

The items on which the differences in thresholds were modeled were selected randomly. Thus in the four item condition, the item from the one-item condition was included and an additional three items were randomly selected. In the eight-item condition, the four items were included an additional four items were randomly selected, and so on.

The sample size for the multi-group CFA was three hundred per group, a sample size that is commonly see in practice. The number of replications for each cell in the simulation design was 100. The nominal alpha was set at .05 for each invariance hypothesis test. It is important to note that the rejection rates reported in this paper are, technically, Type I error rates only for the "no DIF" conditions. In the other cases, when DIF is present, the rejection rates represent the likelihood of rejecting the null hypothesis (for each of the full and strong measurement invariance hypotheses) when the null is true at the unobserved latent variable level, but not necessarily true in the manifest variables because the thresholds are different across the groups.

For each replication the strong and full measurement invariance hypotheses were tested. These hypotheses were tested by comparing the baseline model (with no between group constraints) to each of the strong and full measurement invariance models. That is, strong measurement invariance is the equality of item loadings – Lambda X, and the full measurement invariance is the equality of both item loadings and uniquenesses, Lambda X and Theta-Delta, across groups. For each cell, we searched the LISREL output for the 100 replications for warning or error messages.

A one-tailed 95% confidence interval was computed for each empirical error rate. The confidence interval is particularly useful in this context because we have only 100 replications so we want to take into account sampling variability of the empirical error rate. The upper confidence bound was compared to Bradley's (1978) criterion of liberal robustness of error. If the upper confidence interval was .075 or less it met the liberal criterion.

Table 2. Rejection Rates for the Full and Strong Measurement Invariance Hypotheses, with and without DIF Present.

Percentage of items having different thresholds across the two groups (% of DIF items)	**Number of scale points for the item response format**				
	3 pt.	**4pt.**	**5pt.**	**6pt.**	**7pt.**
0 (no DIF items)	FI .07 (.074) SI .03 (.033)	FI .01 (.012) SI .03 (.033)	FI .01 (.012) SI .04 (.043)	FI .05 (.054) SI .03 (.033)	FI .02 (.022) SI .06 (.064)
2.9 (1 item)	**FI .09 (.095) ⇑** SI .07 (.074)	FI .02 (.022) SI .02 (.022)	FI .01 (.012) SI .01 (.012)	FI .00 (.000) SI .03 (.033)	FI .02 (.022) SI .03 (.033)
10.5 (4 items)	FI .04 (.043) SI .06 (.064)	FI .03 (.033) SI .02 (.022)	FI .03 (.033) SI .04 (.043)	FI .03 (.033) SI .06 (.064)	FI .03 (.033) SI .07 (.074)
21.1 (8 items)	**FI .08 (.084) ⇑** SI .04 (.043)	FI .00 (.000) SI .00 (.000)	FI .04 (.043) SI .04 (.043)	FI .02 (.022) SI .01 (.012)	FI .02 (.022) SI .07 (.074)
42.1 (16 items)	FI .07 (.074) SI .04 (.043)	FI .02 (.022) SI .02 (.022)	FI .02 (.022) SI .06 (.064)	FI .02 (.022) SI .05 (.054)	FI .02 (.022) SI .02 (.022)

Note. The upper confidence bound is provided in parentheses next to the empirical error rate. The empirical error rates in the range of Bradley's liberal criterion are indicated in plain text type whereas empirical error rates that do not even satisfy the liberal criterion are identified with symbol ⇑ and in **bold font**.

Results

To determine whether the tests of strong and full measurement invariance (using the Chi-squared difference tests arising from using a Pearson Covariance matrix and maximum likelihood estimation in, for example, LISREL) are affected by differences in item thresholds we examined the level of error rates in each of the conditions of the simulation design. Table 2 lists the results of the simulation study. Each tabled valued is the empirical error rate over the 100 replications with 300 respondents per group (upon searching the output for errors and warnings produced by LISREL, one case was found of a non-positive definite theta-delta (TD) matrix for the study cells involving three scale points for the 2.9 and 21.1 percent of DIF items. The one replication with this warning was excluded from the calculation of the error rate and upper 95% bound for those two cells, therefore the cell statistics were calculated for 99 replications for those two cases). The values in the range of Bradley's liberal criterion are indicated in plain text type. Values that do not even satisfy the liberal criterion are identified with symbol ⇑.

The results show that almost all of the empirical error rates are within the range of Bradley's liberal criterion. Only two cells have empirical error rates that exceed the upper confidence interval of .075. These two cells are for the three-scale-point condition. This suggests that the differences of item thresholds may have an impact on the full measurement invariance hypotheses in some conditions for measures with a three-point item response format, although this finding is seen in only two of the four conditions involving differences in thresholds. For scale points ranging from four to seven, the empirical error rates are either at or near the nominal error. Interestingly, the empirical error rates of the three scale points are

slightly inflated when a measure has 10.5 and 21.1 percent (moderate amount) of DIF items.

Conclusion

The conclusion from this study is that when one is comparing groups' responses to items that have a rating scale format in a multi-group confirmatory factor analysis of measurement invariance by using maximum likelihood estimation and a Pearson correlation matrix, one should ensure measurement equivalence by investigating item-level differences in thresholds. In addition, giving consideration only to the results of scale-level methods as evidence may be misleading because item-level differences may not manifest themselves in scale-level analyses of this sort.

Of course, the conclusions of this study apply to any situation in which one is (a) using rating scale (sometimes called Likert) items, and comparing two or more groups of respondents in terms of their measurement equivalence, however, it also provides further empirical support for the recommendation found in the International Test Commission Guidelines for Adapting Educational and Psychological Tests that researchers carry out empirical studies to demonstrate factorial equivalence of their test across groups *and* to identify any item-level DIF that may be present (see Hambleton & Patsula, 1999; van de Vijver & Hambleton, 1996) and is an extension of previous studies by Zumbo (2000; 2003) comparing and item- and scale-level methods.

Overall, the results demonstrate the inadequacy of judging the suitability of a measurement instrument across groups by only investigating the factor structure of the measure for the different groups with a Pearson covariance matrix and maximum likelihood estimation. It has been common to assume that if the factor structure of a test remains the same in a second group, then the measure functions the same and measurement equivalence is achieved. Evidence is provided that item level bias can still be present when a CFA of the two different groups reveals an equivalent factorial structure of rating scale items using a Pearson covariance matrix and maximum likelihood estimation. Since it is the scores from a test or instrument that are ultimately used to achieve the intended purpose, the scores may be contaminated by item level bias and, ultimately, valid inferences from the test scores become problematic.

References

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289-311.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS*. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adaptive tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1-11.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago, IL.: Scientific Software International.

Luecht, R. (1996). MIRTGEN 1.0 [Computer Software].

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Zumbo, B. D. (2000, April). *The effect of DIF and impact on classical test statistics: undetected DIF and impact, and the reliability and interpretability of scores from a language proficiency test*. Paper presented at the *National Council on Measurement in Education (NCME)*, New Orleans, LA.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses?: Implications for translating language tests. *Language Testing, 20*, 136-147.

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment* (p. 505-509). Sage Press, Thousand Oaks, CA.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 283-287

1538 – 9472/05/$95.00

Brief Report
Exploratory Factor Analysis in Two Measurement Journals: Hegemony by Default

J. Thomas Kellow
College of Education
University of South Florida-Saint Petersburg

Exploratory factor analysis studies in two prominent measurement journals were explored. Issues addressed were: (a) factor extraction methods, (b) factor retention rules, (c) factor rotation strategies, and (d) saliency criteria for including variables. Many authors continue to use principal components extraction, orthogonal (varimax) rotation, and retain factors with eigenvalues greater than 1.0.

Key words: Factor analysis, principal components, current practice

Introduction

Factor analysis has often been described as both an art and a science. This is particularly true of exploratory factor analysis (EFA), where researchers follow a series of analytic steps involving judgments more reminiscent of qualitative inquiry, an irony given the mathematical sophistication underlying EFA models.

A number of issues must be considered before invoking EFA, such as sample size and the relationships between measured variables (see Tabachnick & Fidell, 2001, for an overview). Once EFA is determined to be appropriate, researchers must consider carefully decisions related to: (a) factor extraction methods, (b) rules for retaining factors, (c) factor rotation strategies, and (d) saliency criteria for including variables. There is considerable latitude regarding which methods may be appropriate or desirable in a particular analytic scenario (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

J. Thomas Kellow is an Assistant Professor of Measurement and Research in the College of Education at the University of South Florida-Saint Petersburg. Email him at: kellow@stpt.usf.edu

Factor Extraction Methods

There are numerous methods for initially deriving factors, or components in the case of principal component (PC) extraction. Although some authors (Snook & Gorsuch, 1989) have demonstrated that certain conditions involving the number of variables factored and initial communalities lead to essentially the same conclusions, the unthinking use of PC as an extraction mode may lead to a distortion of results. Stevens (1992) summarizes the views of prominent researchers, stating that:

> When the number of variables is moderately large (say > 30), and the analysis contains virtually no variables expected to have low communalities (e.g., .4), then practically any of the factor procedures will lead to the same interpretations. Differences can occur when the number of variables is fairly small (< 20), and some communalities are low. (p. 400)

Factor Retention Rules

Several methods have been proposed to evaluate the number of factors to retain in EFA. Although the dominant method seems to be to retain factors with eigenvalues greater than 1.0, this approach has been questioned by numerous authors (Zwick & Velicer, 1986; Thompson & Daniel, 1996). Empirical evidence suggests that, while under-factoring is probably the greater

danger, sole reliance on the eigenvalues greater than 1.0 criterion may result in retaining factors of trivial importance (Stevens, 1992). Other methods for retaining factors may be more defensible and perhaps meaningful in interpreting the data. Indeed, after reviewing empirical findings on its utility, Preacher and McCallum (2003) reported that "the general conclusion is that there is little justification for using the Kaiser criterion to decide how many factors to retain" (p. 23).

Factor Rotation Strategies

Once a decision has been made to retain a certain number of factors, these are often rotated in a geometric space to increase interpretability. Two broad options are available, one (orthogonal) assuming the factors are uncorrelated, and the second (oblique) allowing for correlations between the factors. Although the principal of parsimony may tempt the researcher to assume, for the sake of ease of interpretability, uncorrelated factors, Pedhazur and Shmelkin (1991) argued that both solutions should be considered. Indeed, it might be argued that it rarely is tenable to assume that multidimensional constructs, such as self-concept, are comprised of dimensions that are completely independent of one another. Although interpretation of factor structure is somewhat more complicated when using oblique rotations, these methods may better honor the reality of the phenomenon being investigated.

Saliency Criteria for Including Variables

Many researchers regard a factor loading (more aptly described as a pattern or structure coefficient) of $|.3|$ or above as worthy of inclusion in interpreting factors (Nunnally, 1978). This rationale is predicated on a rather arbitrary decision rule that 9% of variance accounted for makes a variable noteworthy. In a similar vein, Stevens (1992) offered $|.4|$ as a minimum for variable inclusion as this means the variable shares at least 15% of its variance with a factor. Others (Cliff & Hamburger, 1967) argue for the statistical significance of a variable as an appropriate criterion for inclusion. As Hogarty, Kromrey, Ferron, and Hines (in press) noted, "although a variety of rules of thumb of this nature are venerable, they are often ad hoc and ill advised."

Purpose of the Study

This article does not attempt to provide an introduction to the statistical and conceptual intricacies of EFA techniques, as numerous excellent resources are available that address these topics (e.g., Gorsuch, 1983; Stevens, 1992; Tabachnick & Fidell, 2001; Thompson, 2004). Rather, the focus is on the practices of EFA authors with respect to the above issues. Three of the four important EFA analytic decisions described above are treated by default in SPSS and SAS. These programs are the most widely used analytic platforms in psychology. When conducting EFA in either program, one is guided to (a) use PC as the *extraction method* of choice, (b) use eigenvalues greater than 1.0 to *retain factors*, and (c) use orthogonal (varimax) procedures for *rotation of factors*. Only the fourth decision, *variable retention*, is left solely to the preference of the investigator.

EFA practices in two prominent psychological measurement journals were examined: *Educational and Psychological Measurement* (*EPM*) and *Personality and Individual Differences (PID)* over a six-year period. These journals were chosen because of their prominence in the field of measurement and the prolific presence of EFA articles within their pages. In addition, *EPM* is known for publishing factor analytic studies across a diverse array of specialization areas in education and psychology. While *PID* is concerned primarily with the study of personality, it publishes a great deal of international studies from diverse institutions. These features strengthen the external validity of the present findings.

Methodology

An electronic search was conducted using the PsycInfo database for *EPM* and *PID* studies published from January of 1998 to October 2003 that contained the key word 'factor analysis.' After screening out studies that employed only confirmatory factor analysis or examined the statistical properties of EFA or CFA approaches using simulated data sets, a total of 184 articles

were identified. In some instances the authors conducted two or more EFA analyses on split samples. For the present purposes these were coded as separate studies. This resulted in 212 studies that invoked EFA models. Variables extracted from the EFA articles were:

a) factor extraction methods;
b) factor retention rules;
c) factor rotation strategies; and
d) saliency criteria for including variables.

Results

Factor Extraction Methods

The most common extraction method employed (64%) was principal components (PC). The next most popular choice was principal axis (PA) factoring (27%). Techniques such as maximum likelihood were infrequently invoked (6%). A modest percentage of authors (8%) conducted both PC and PA methods on their data and compared the results for similar structure.

Factor Extraction Rules

The most popular method used for deciding the number of factors to retain was the Kaiser criterion of eigenvalues greater than 1.0. Over 45% of authors used this method. Close behind in frequency of usage was the scree test (42%). Use of other methods, such as percent of variance explained logics and parallel analysis, was comparatively infrequent (about 8% each). Many authors (41%) explored multiple criteria for factor retention. Among these authors, the most popular choice was a combination of the eigenvalues greater than 1.0 and scree methods (67%).

Factor Rotation Strategies

Virtually all of the EFA studies identified (96%) invoked some form of factor rotation solution. Varimax rotation was most often employed (47%), with Oblimin being the next most common (38%). Promax rotation also was used with a modest degree of frequency (11%). A number of authors (18%) employed both Varimax and Oblimin solutions to examine the influence of correlated factors on the resulting factor pattern/structure matrices.

Saliency Criteria for Including Variables

Thirty-one percent of EFA authors did not articulate a specific criterion for interpreting salient pattern/structure coefficients, preferring instead to examine the matrix in a logical fashion, considering not only the size of the pattern/structure coefficient, but also the discrepancy between coefficients for the same variable across different factors (components) and the logical "fit" of the variable with a particular factor.

Of the 69% of authors who identified an *a priori* criterion as an absolute cutoff, 27% opted to interpret coefficients with a value of $|.3|$ or higher, while 24% chose the $|.4|$ value. Other criteria chosen with modest frequency (both about 6%) included $|.35|$ and $|.5|$ as absolute cutoff values. For the remaining authors who invoked an absolute criterion, values ranged from $|.25|$ to $|.8|$. A few (3%) of these values were determined based on the statistical significance of the pattern/structure coefficient.

Conclusion

Not surprisingly, the hegemony of default settings in major statistical packages continues to dominate the pages of *EPM* and *PID*. The Little Jiffy model espoused by Kaiser (1970), wherein principal components are rotated to the varimax criterion and all components with eigenvalues greater than 1.0 is alive and well. It should be noted that this situation is almost certainly not unique to *EPM* or *PID* authors. An informal perusal of a wide variety of educational and psychological journals that occasionally publish EFA results easily confirms the status of current practice.

The rampant use of PC as an extraction method is not surprising given its status as the default in major statistical packages. Gorsuch (1983) has pointed out that, with respect to extraction methods, PC and factor models such as PA often yield comparable results when the number of variables is large and communalities (h^2) also are large. Although comforting, authors are well advised to consider alternative extraction methods with their data even when these assumptions are met. When these assumptions are not met, such as "when the rank of the factored matrix is small, there is

considerable measurement error, measurement error is not homogeneous across variables, and sampling error is small due to larger sample size, *other extraction methods have more appeal*" (Thompson & Daniel, 1996, p. 202, italics added).

The eigenvalues greater than 1.0 criterion was the most popular option for EFA analysts. A number of researchers, however, combined both the eigenvalues greater than 1.0 criterion and the scree test in combination, which is interesting inasmuch as both methods consult eigenvalues, only in different ways. A likely explanation is that both can be readily obtained in common statistical packages.

Other approaches to ascertaining the appropriate number of factors (components) such as parallel analysis (Horn, 1965) and the bootstrap (Thompson, 1988) are available, as are methods based on standard error scree (Zoski & Jurs, 1996). Each of these methods, however, requires additional effort on the part of the researcher. However, EFA authors should consider alternatives for factor retention in much the same way that CFA authors consult the myriad fit indices available in model assessment. As Thompson and Daniel noted, "The simultaneous use of multiple decision rules is appropriate and often desirable" (p. 200).

For authors invoking an absolute criterion for retaining variables, the $|.3|$ level and the $|.4|$ were by far the most popular. Researchers who feel compelled to set such arbitrary criteria often look to textbook authors to guide their choice. The latter criterion can be traced to Stevens (1992), who stated that "It would seem that one would want in general a variable to share *at least* 15% of its variance with the construct (factor) it is going to be used to help name. This means only using loadings (sic) which are about .4 or greater for interpretation purposes" (p. 384). The former rule appears to be attributable to Nunnally (1982), who claimed that "It is doubtful that loadings (sic) of any smaller size should be taken seriously, because they represent less than 10 percent of the variance" (p. 423).

One-third of EFA authors chose not to adhere to a strict, and ultimately arbitrary, criterion for variable inclusion. Rather, these researchers considered the pattern/structure coefficients within the context of the entire matrix, applying various logics such as simple structure and *a priori* inclusion of variables. A (very) few authors considered the statistical significance of the coefficients in their interpretation of salient variables.

Two problems with this approach are that (a) with very large samples even trivial coefficients will be statistically significant, and (b) variables that are meaningfully influenced by a factor may be disregarded because of a small sample size. The issue of determining the salience of variables based on their contribution to a model mirrors that of the debate over statistical significance and effect size. If standards are invoked based solely on the statistical significance of a coefficient, or alternatively, are set based on a strict criterion related to the absolute size of a coefficient related to its variance contribution, it would seem that we would "merely be being stupid in another metric" (Thompson, 2002, p. 30).

Despite criticisms that the technique is often employed in a senseless fashion (e.g., Preacher & MacCallum, 2003), EFA provides researchers with a valuable inductive tool for exploring the dimensionality of data provided it is used thoughtfully. The old adage that factor analysis is as much an art as a science is no doubt true. But few artists rely on unbending rules to create their work, and authors who employ EFA should be mindful of this fact.

References

Cliff, N., & Hamburger, C. D. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin, 68*, 430-445.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Earlbaum.

Hogarty, K. Y., Kromrey, J. D., Ferron, J. M., & Hines, C. V. (in press). Selection of variables in exploratory factor analysis: An empirical comparison of a stepwise and traditional approach. *Psychometrika*.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35*, 401–415.

Nunnally, J. C. (1978). *Psychomteric theory* (2nd ed.). New York: McGraw Hill.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis*. Hillsdale, NJ: Erlabaum.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electronic factoring machine. *Understanding Statistics, 2*, 13-43.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin*. *Personality and Social Psychology Bulletin, 28*, 1629-1646.

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106*, 148-154.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th Ed.). Needham Heights, MA: Allyn & Bacon.

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. *Educational and Psychological Measurement, 48*, 681-686.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Zoski, K. W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement, 56*, 443-451.

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 288-299

1538 – 9472/05/$95.00

Multiple Imputation For Missing Ordinal Data

Ling Chen
University of Arizona

Mariana Toma-Drane
University of South Carolina

Robert F. Valois
University of South Carolina

J. Wanzer Drane
University of South Carolina

Simulations were used to compare complete case analysis of ordinal data with including multivariate normal imputations. MVN methods of imputation were not as good as using only complete cases. Bias and standard errors were measured against coefficients estimated from logistic regression and a standard data set.

Key words: complete case analysis, missing data mechanism, multiple logistic regression

Introduction

Surveys are important sources of information in epidemiologic studies and other research as well, but often encounter missing data (Patricia, 2002). Ordinal variables are very common in survey research; however, they challenge primary data collectors who might need to impute missing values of these variables due to their hierarchical nature but with unequal intervals.

The traditional approach, complete case analysis (CC), excludes from the analysis observations with any missing value among variables of interest (Yuan, 2000). CC remains the most common method in the absence of readily available alternatives in software packages. However, using only complete cases could result in losing information about incomplete cases, thus biasing parameter estimates, and compromising statistical power (Patricia, 2002). Multiple imputation (MI) procedure replaces each missing value with *m* plausible values generated under an appropriate model. These *m* multiply imputed datasets are then analyzed separately by using procedures for complete data to obtain desired parameter estimates and standard errors. Results from the *m* analyses are then combined for inferences by computing the mean of the *m* parameter estimates and a variance estimate that include both a within-imputation and a between-imputation component (Rubin, 1987).

MI has some desirable features, such as introducing appropriate random error into the imputation process and making it possible to obtain unbiased estimates of all parameters; allowing use of complete-data methods for data analysis; producing more reasonable estimates of standard errors and thereby increasing efficiencies of estimates (Rubin, 1987). In addition, MI can be used with any kind of data and any kind of analysis without specialized software (Allison, 2000). MI appears to be a more attractive method handling missing data in multivariate analysis compared to CC (King et al., 2001; Little & Rubin, 1989).

However, certain requirements should be met to have its attractive properties. First, the data must be missing at random (MAR). Second, the model used to generate the imputed values must be correct in some sense. Third, the model used for the analysis must catch up, in some sense, with the model used in the imputation

Ling Chen is a doctoral student in the Department of Statistics at the University of Missouri, Columbia. Mariana Toma-Drane, is a doctoral student at Norman J. Arnold, School of Public Health, Department of Health Promotion Education and Behavior. John Wanzer Drane is Professor of Biostatistics at USC and Fellow of the American Academy of Health Behavior. Robert F. Valois is Professor Health Promotion, Education and Behavior at USC and a Fellow of the American Academy of Health Behavior.

(Allison, 2000). All these conditions have been rigorously described by Rubin (1987) and Schafer (1997). The problem is that it is easy to violate these conditions in practice.

The purpose of this study was to investigate how well multivariate normal (MVN) based MI deals with non-normal missing ordinal covariates in multiple logistic regression, while there is definite violation against the distributional assumptions of the missing covariates for the imputation model.

Simulated scenarios were created for the comparison assuming various missing rates for the covariates (5%, 15% and 30%) and different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing nonignorable (NI). The performance of MVN based MI was compared to CC in each scenario.

Methodology

The mechanism that leads to values of certain variables being missing is a key element in choosing an appropriate analysis and interpreting the results (Little & Rubin, 1987).

In sample survey context, let Y denote an $n \times p$ matrix of multivariate data, which is not fully observed. Let Y_{obs} denote the set of fully observed values of Y and Y_{mis} denote the set containing missing values of Y, i.e., $Y = (Y_{obs}, Y_{mis})$.

Rubin (1976) introduced a missing data indicator matrix R. The (i, j)th element $R_{ij} = 1$ if Y_{ij} is observed; and $R_{ij} = 0$ if Y_{ij} is missing. The notation of missing data mechanisms was formalized in terms of a model for the conditional distribution $P(R \mid Y, \zeta)$ of R given Y according to whether the probability of response depends on Y_{obs} or Y_{mis} or both, where ζ is an unkown parameter.

Data are MCAR, if the distribution of R does not depend on Y_{obs} or Y_{mis}; that is $P(R \mid Y, \zeta) = P(R \mid \zeta)$ for all Y. In this case, the observed values of Y form a random subset of all the sampled values of Y. Data are MAR if the distribution of R depends on the data Y only through the observed values Y_{obs}; that is, $P(R|Y, \zeta) = P(R|Y_{obs}, \zeta)$ for all Y_{mis}. MAR implies missing depends on observed covariates and outcomes, or missingness can be predicted by observed information. MCAR is a special case of MAR. The missing data mechanism is ignorable for likelihood-based inferences for both MCAR and MAR (Little & Rubin, 1987). Missing NI occurs when the probability of response of Y depends on the value of Y_{mis} and possibly the value of Y_{obs} as well.

The data used in this investigation are from the 1997 South Carolina Youth Risk Behavior Survey (SCYRBS). The total number of complete and partial questionnaires collected is 5545. The survey employed a two-stage cluster sampling with derived weightings designed to obtain a representative sample of all South Carolina public high school students in grades 9-12, with the exception of those in special education schools. The survey ran from March until June 1997.

The questionnaire covers six categories of priority health-risk behaviors required by the Center for Disease Control and Prevention, and locally, two additional psychological categories of questions were added that include quality of life and life satisfaction (Valois, Zulling, Huebner & Drane, 2001). The six categories of priority health-risk behaviors among youth and young adults are those that contribute to unintentional and intentional injuries; tobacco use; alcohol and other drug use; sexual behaviors; dietary behaviors and physical inactivity (Kolbe, 1990).

The items on self-report youth risk behaviors are Q10 through Q20. The six life-satisfaction variables, Q99 through Q104, are based on six domains: family, friends, school, self, living environment and overall life satisfaction. Each of the questions has seven response options based on the Multidimensional Students' Life Satisfaction Scale (Seligson, Huebner & Valois, 2003). The response options are from the Terrible-to-Delighted Scale: 1 - terrible; 2 - unhappy; 3 - mostly dissatisfied; 4 - equally satisfied and dissatisfied; 5 - mostly satisfied; 6 - pleased; and 7 - delighted (10).

The four race-gender groups: White Females (WF, 26.7%), White Males (WM, 26.0%), Black Females (BF, 26.0%) and Black Males (BM, 21.3%) accounted for almost equal percentage in the sample. The sample was due to the belief that the relationship between life satisfaction and youth risk behaviors varies

across different race-gender groups, as demonstrated in previous research (Valois, Zulling, Huebner & Drane, 2001).

Multiple Logistic Regression Analysis

Exploring the relationship between life satisfaction and youth risk behaviors powered this study. Three covariates in ordinal scale were selected from the 1997 SCYRBS Questionnaire (see the Appendix for details). They were dichotomized as Q10: DRKPASS (Riding with a drunk driver); Q14: GUNSCHL (Carrying a gun or other weapon on school property) and Q18: FIGHTIN (Physical fighting), respectively. Each of them was coded "1" for "never" (0 time) and "2" for "ever" (equal to or greater than 1 time), with "1" as the referent level. All the six ordinal variables of life satisfaction (Q99 ~ Q104) were pooled for each participant to form a pseudo-continuous dependent variable ranging in score from 6 to 42, i.e., "Lifesat = Q99 + Q100 + Q101 + Q102 + Q103 + Q104". The score was expressed as Satisfaction Score (SS) with lower scores indicative of reduced satisfaction with life (Valois, Zulling, Huebner & Drane, 2001). SS ranging from 6 to 27 was categorized as dissatisfied. For the dichotomized outcome variable D2, the students in dissatisfied group (D2 = 1) served as the risk group and the others as the referent group (D2 = 0).

As defined, all the four variables used in logistic regression were dichotomized. DRKPASS, GUNSCHL and FIGHTIN were used as predictor variables while D2 was chosen as the response or criterion variable. The three predictor variables are each independently associated with life dissatisfaction with odds ratios (OR) ranging from 1.42 to 2.27; they are also associated with each other with odds ratios ranging from 2.22 to 4.52.

To use the sampling design in multiple logistic regression analysis, dichotomous logistic regression (PROC MULTILOG) was conducted using SAS-callable Survey Data Analysis (SUDAAN) for weighted data at an alpha level of 0.05 (Shah, Barnwell & Bieler, 1997) (See Appendix.). The analyses were done separately for the four race-gender groups, and the regression coefficient (β) and the standard error of the regression coefficient (Se (β)) for each covariate were obtained.

Simulations

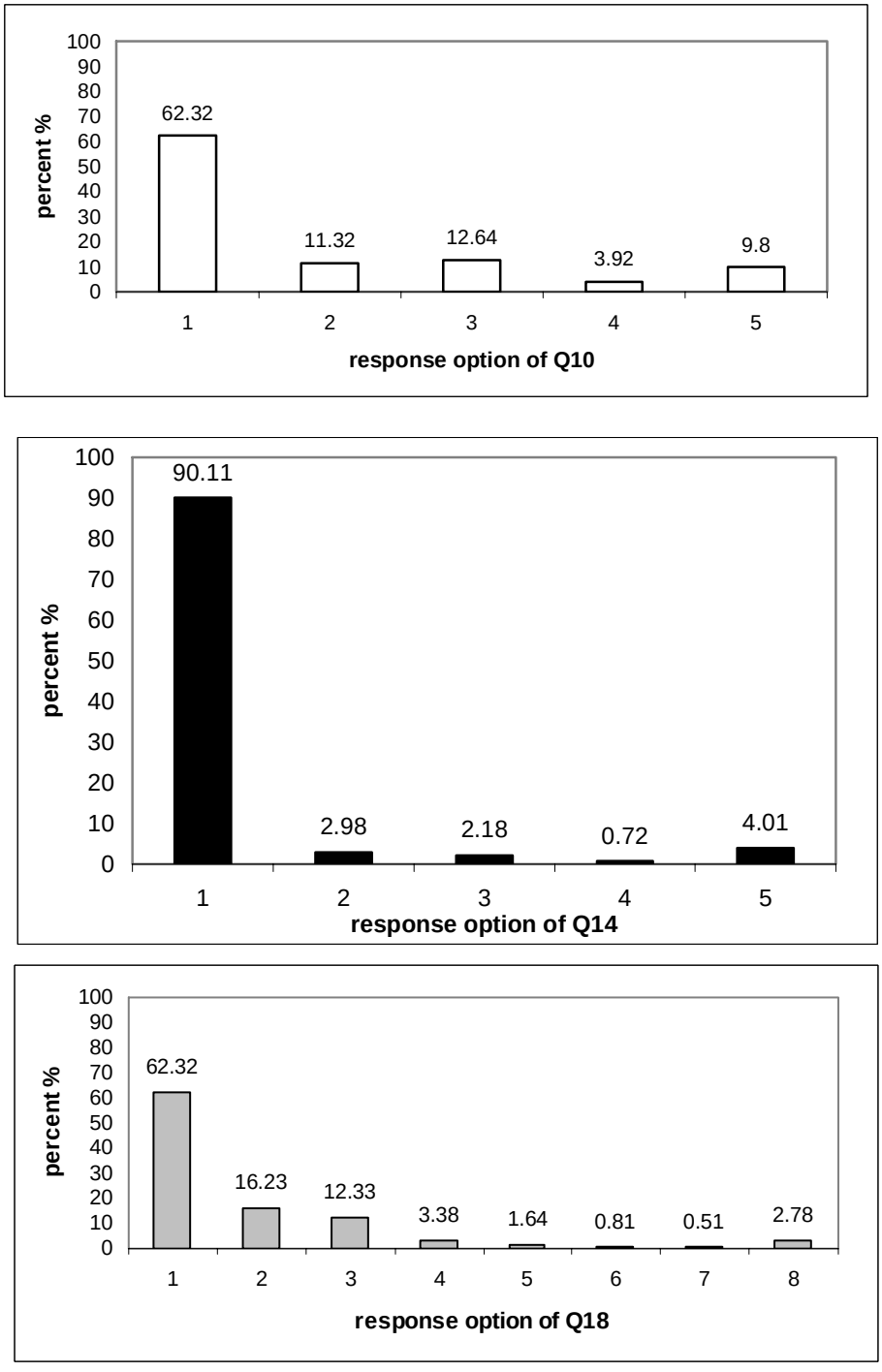
Simulations were applied to compare the performance of CC and MI in estimating regression. Create a complete standard dataset. The SAS MI procedure was used to impute the very few missing values in the youth risk behavior variables (Q10 through Q20) and the six life-satisfaction variables (Q99 through Q104) in the 1997 SCYRBS Dataset *once*, because missing percentages of these variables are very low, ranging from 0.13% to 4.11%. The resulting dataset was regarded as the Complete Standard Dataset in the simulations. This dataset was considered the true gold standard and some values of the three variables related to the three predictors in logistic regression were set to be missing. The PROC MI code (see Appendix) used to create the Complete Standard Dataset was the same as that used to impute values for missing covariates except that missing values were imputed five times in the simulations. The distributions of the three ordinal covariates in the Complete Standard Dataset were also examined. The three covariates are all highly skewed instead of being approximately normal (Figure 1).

Simulating datasets with missing covariates

Three missing data mechanisms were simulated: MCAR, MAR and NI. For the case of MCAR, each simulated sample began by *randomly* deleting a certain percentage of the values of Q10, Q14 and Q18 from the Complete Standard Dataset such that the three covariates were missing at the same rate (5%, 15% and 30%).

For MAR, a certain percentage of values of Q10 were removed from the Complete Standard Dataset with a probability related to the outcome variable (D2) and the other two variables Q14 and Q18. For the NI condition, a certain percentage of values of Q10 were removed such that the larger values of Q10 were more likely to be missing, as in real datasets some covariates corresponding to sensitive matters, whether large or small, their responses are often more likely to be missing (Wu & Wu, 2001). For all the scenarios assuming MAR and NI, Q14 and Q18 were randomly removed assuming MCAR at the same rate as Q10.

Figure 1. Distribution of the three covariates in the Complete Standard Dataset.



Nine scenarios were created where the covariates Q10 (DRKPASS), Q14 (GUNSCHL) and Q18 (FIGHTIN) were missing at the same rate (5%, 15% and 30%), the life-satisfaction variables (Q99 ~ Q104) were complete as in the Complete Standard Dataset, however. In each scenario 500 datasets with missing covariates were generated. Table 1 lists the missing data mechanisms for the covariates, and the average percentage of complete cases (all the three covariates complete) in the 500 datasets for each scenario. All the simulations were performed using SAS version 8.2 (2002).

Multiple Imputation

The missing covariates in each simulated dataset were then imputed five times using the SAS MI procedure (see Appendix). First, initial parameter estimates were obtained by running the Expectation-Maximization (EM) algorithm until convergence up to a maximum of 1000 iterations. Using the EM estimates as starting values, 500 cycles were ran of Markov Chain Monte Carlo (MCMC) full-data augmentation under a ridge prior with the hyperparameter set to 0.75 to generate five imputations. A multivariate normal model was applied to the data augmentation for the non-normal ordinal data without trying to meet the distributional assumptions of the imputation model.

Three auxiliary variables (Q11, Q13 and Q19) as well as the outcome variable D2 were entered into the imputation model as if they were jointly normal, to increase the accuracy of the imputed values of Q10, Q14 and Q18 (Allison, 2000; Schafer, 1997 & 1998; Rubin, 1996).

The maximum and minimum values for the imputed values were specified, which were based on the scale of the response options for the 1997 SCYRBS questions. These specifications were necessary so that the imputations were not made outside of the range of the original variables. The continuously distributed imputes for Q10, Q14 and Q18 were rounded to the nearest category using a cutoff value of 0.5.

Inferences from CC and MI

For inference from CC, multiple logistic regression analysis was performed for each of the 500 datasets with missing covariates. The estimates for β and Se (β) for CC in each scenario were the average of the 500 estimates from the 500 incomplete datasets, respectively. For inference from MI, The point estimate of β was first obtained from the five imputed dataset estimates; and Se (β) was obtained by combining the within-imputation variance and between-imputation variance from the five repeated imputations (Rubin, 1987; SAS Institute, 2002). The estimates for β and Se (β) for MI in each scenario were the average of the 500 point estimates of β and the 500 combined Se (β), respectively.

Comparison of complete case and multiple imputation model results

To compare the performance of CC and MI, biases and standard errors of point estimates were mainly considered. Each regression coefficient calculated from the Complete Standard Dataset was taken as the true coefficient and those from CC and MI in each scenario were compared to the true ones. Bias is expressed as estimate from CC or MI minus the estimate from the Complete Standard Dataset, i.e., estimated $\beta - \beta_{\text{true value}}$. The average absolute value of bias (AVB) of β for each covariate was compared between the two methods for the same race-gender group.

Results

The missing values in the risk behavior and life-satisfaction variables were imputed, and the resulting dataset was defined as the Complete Standard Dataset as if it was originally complete. Table 2 contains the estimates and standard errors of the regression coefficients from the 1997 SCYRBS dataset together with those from the Complete Standard Dataset. Given the low percentages of missing variables in 1997 SCYRBS dataset and thus the few cases omitted from the CC, the results from the two datasets are very similar.

Table 1. Simulated scenarios for datasets with missing covariates.

Scenario	Missing percentage of each covariate	Average percentage of complete cases	Missing data mechanism for each covariate		
			Q10 (DRKPASS)	Q14 (GUNSCHL)	Q18 (FIGHTIN)
1	5%	85.73%	MCAR	MCAR	MCAR
2	5%	85.42%	MAR	MCAR	MCAR
3	5%	85.55%	NI	MCAR	MCAR
4	15%	61.34%	MCAR	MCAR	MCAR
5	15%	61.19%	MAR	MCAR	MCAR
6	15%	62.54%	NI	MCAR	MCAR
7	30%	34.22%	MCAR	MCAR	MCAR
8	30%	34.30%	MAR	MCAR	MCAR
9	30%	34.10%	NI	MCAR	MCAR

Table 2. Logistic regression coefficients and standard error estimates in the 1997 SCYRBS Dataset and the Complete Standard Dataset.

Group	DRKPASS		GUNSCHL		FIGHTIN	
	β *	Se(β) †	β	Se(β)	β	Se(β)
White female	0.14	0.10	0.99	0.21	0.88	0.16
N=1359 (1361) ‡	(0.16)	(0.11)	(0.94)	(0.23)	(0.84)	(0.16)
Black female	0.03	0.14	0.69	0.28	0.36	0.15
N=1335 (1336)	(0.02)	(0.14)	(0.63)	(0.24)	(0.45)	(0.16)
White male	0.32	0.17	0.10	0.17	0.43	0.13
N=1338 (1340)	(0.25)	(0.16)	(0.32)	(0.15)	(0.53)	(0.11)
Black male	0.43	0.16	0.95	0.20	0.32	0.11
N=1119 (1119)	(0.35)	(0.14)	(0.94)	(0.23)	(0.52)	(0.11)

* β, logistic regression coefficient.

† Se (β), standard error of logistic regression coefficient.

‡ Numbers in parentheses, sample size, logistic regression coefficient and standard error of logistic regression coefficient from the Complete Standard Dataset.

An example is presented from comparing CC and MI across the nine scenarios among White Females in table 3. The histogram of the average AVB of β for each covariate in this example is shown in figure 2. To evaluate the the imputation procedure, the absolute value of bias in point estimates and coverage probability were mainly considered. The coverage probability is defined as the possibility of the true regression coefficient β being covered by the actual 95 percent confidence interval. Further, the percent AVB of β for each covariate, calculated by dividing AVB by the corresponding true β, better compares the two methods with regard to bias. Greater or equal to10% of bias is beyond acceptance.

Both CC and MI produced biased estimates of β in all the scenarios. CC showed little or no bias for all the scenarios under MCAR. The AVB of β for each covariate is consistently less than 0.05 for all the three covariates even with about 34% complete cases (30% missing for each covariate). However, CC showed larger AVB's of β in the scenarios under MAR and NI than in those under MCAR with the same missing covariate rates. Further, MI was generally less successful than CC because MI showed larger AVB's of β than CC in most of the scenarios regardless of missing data mechanism and missing covariate rate. (Results for the other three race-gender groups not shown here.)

Figure 2. Average AVB's (absolute value of bias) of logistic regression coefficients across the nine scenarios among White Females. S1 ~ S9 represent Scenario1 ~ Scenario 9, respectively.

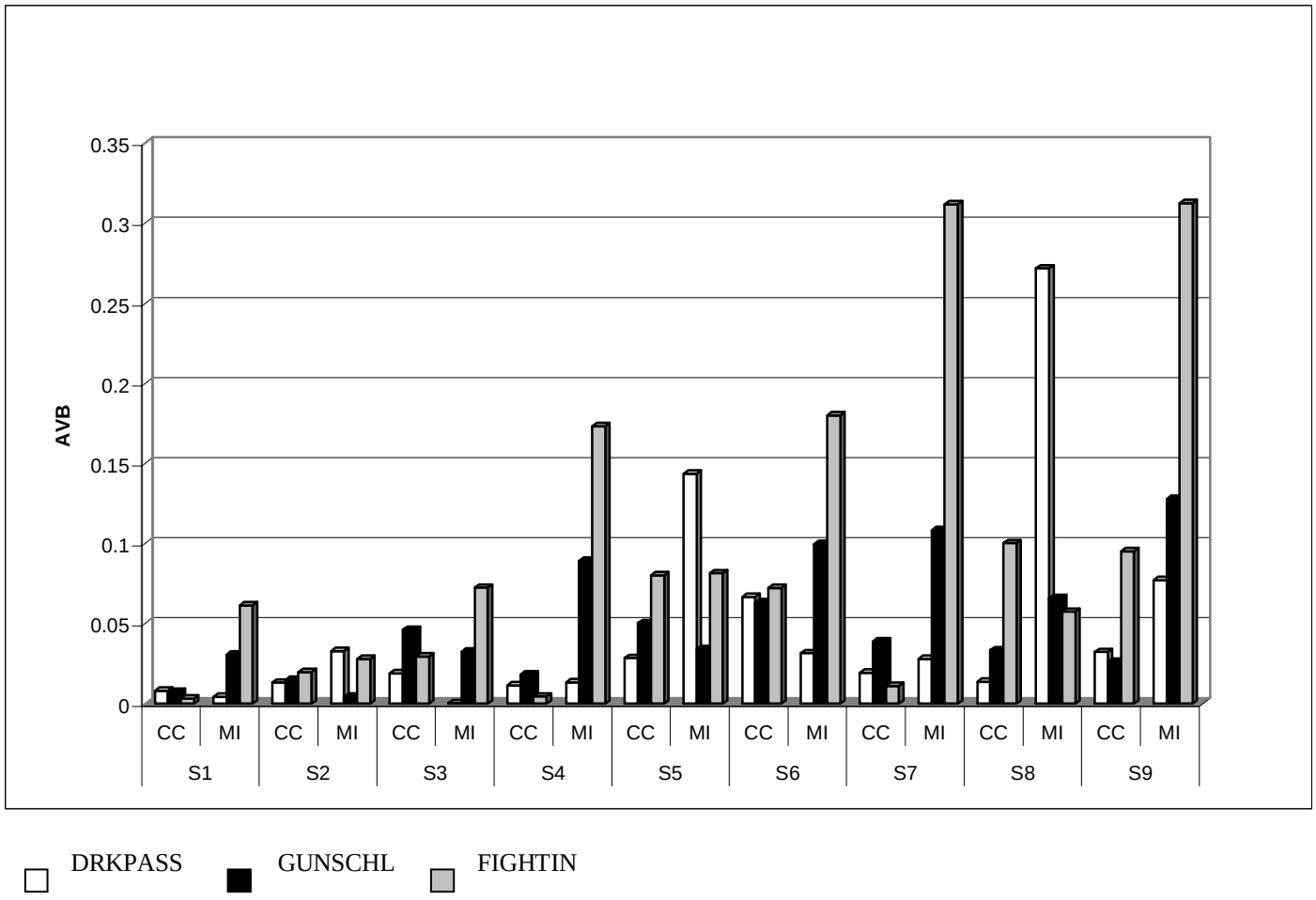


Table 3**.** Comparison of complete case and multiple imputation model results across the nine scenarios among White Females.

		DRKPASS β *$_{\text{true value}}$ = 0.16 Se (β) $_{\text{true value}}$ = 0.11		GUNSCHL β $_{\text{true value}}$ = 0.94 Se (β) $_{\text{true value}}$ = 0.23		FIGHTIIN β $_{\text{true value}}$ = 0.84 Se (β) $_{\text{true value}}$ = 0.16	
		AVB ‡	Se(β) †	AVB	Se(β)	AVB	Se(β)
Scenario 1	CC	0.0082	0.1178	0.0075	0.2574	0.0033	0.1740
	MI	0.0043	0.1088	0.0306	0.2414	0.0613	0.1627
Scenario 2	CC	0.0132	0.1190	0.0148	0.2725	0.0198	0.1768
	MI	0.0329	0.1094	0.0044	0.2448	0.0280	0.1602
Scenario 3	CC	0.0189	0.1162	0.0462	0.2712	0.0295	0.1759
	MI	0.0004	0.1061	0.0324	0.2470	0.0725	0.1593
Scenario 4	CC	0.0116	0.1467	0.0182	0.3302	0.0046	0.1964
	MI	0.0133	0.1151	0.0893	0.2591	0.1732	0.1574
Scenario 5	CC	0.0286	0.1521	0.0504	0.3666	0.0802	0.2039
	MI	0.1437	0.1166	0.0339	0.2610	0.0815	0.1555
Scenario 6	CC	0.0667	0.1451	0.0633	0.3517	0.0724	0.2105
	MI	0.0315	0.1137	0.0996	0.2628	0.1800	0.1556
Scenario 7	CC	0.0194	0.2097	0.0390	0.4840	0.0111	0.2478
	MI	0.0279	0.1237	0.1083	0.2704	0.3118	0.1523
Scenario 8	CC	0.0138	0.1991	0.0335	0.4312	0.1002	0.2347
	MI	0.2718	0.1227	0.0660	0.2738	0.0575	0.1500
Scenario 9	CC	0.0323	0.2227	0.0261	0.5611	0.0951	0.2661
	MI	0.0771	0.1344	0.1278	0.2828	0.3126	0.1506

* β, logistic regression coefficient.

† Se (β), standard error of logistic regression coefficient.

‡ AVB, absolute value of bias (| estimated $\beta - \beta_{\text{true value}}$ |).

Table 4. Coverage probability in Scenarios 2 and 8 for White Females.

		DRKPASS (%)	GUNSCHL (%)	FIGHTIN (%)
Scenario 2	CC	96.8	96.4	95.0
	MI	94.2	99.0	93.8
Scenario 8	CC	95.0	94.0	87.0
	MI	77.0	90.4	88.2

Table 5. Average Correct Imputation Rate for the three covariates.

Transformation *	Scenario	Original scale (%)			Recoded (%)		
		Q10	Q14	Q18	DRKPASS	GUNSCHL	FIGHTIN
Without	2	15.94	83.20	31.40	47.81	86.77	49.20
	8	21.25	83.22	29.21	41.05	86.75	47.39
With	2	40.04	89.47	50.80	65.14	92.11	66.00
	8	52.40	89.54	50.75	65.52	91.81	63.22

Also, in most scenarios the percent AVB of β from MI is far greater than that from CC and is greater than 10% of acceptance level. This discrepancy was especially obvious for all the scenarios under MCAR (Scenarios 1, 4 and 7). Moreover, the AVB's and percent AVB's from MI increase substantially as larger proportions of the covariates were missing. Interestingly, MI showed consistently decreased Se (β) for each covariate in all the scenarios, which is not surprising, because the standard error of MI is based on full datasets (Allison, 2001).

Table 4 lists the coverage probabilities in Scenarios 2 and 8 among White Females as an example. In both scenarios, the coverage probabilities from MI are not all better than those from CC.

Clearly, the current MVN based multiple imputation did not perform as well as CC in generating unbiased regression estimates. To investigate how well the present MI actually imputed the missing non-normal ordinal covariates, Scenarios 2 and 8 were used to check the imputation efficiency, as the two scenarios have the same setting for missing data mechanism but different missing covariate rates. The Average Correct Imputation Rate is calculated as the average proportion of correctly imputed observations among the missing covariates. Correct imputation occurs when the imputed value is identical to its true value in the Complete Standard Dataset. Table 5 displays the Average Correct Imputation Rates for the three covariates in both original scales (Q10, Q14 and Q18) and recoded scales (DRKPASS, GUNSCHL and FIGHTIN).

The Average Correct Imputation Rates for Q10 and Q18 are lower than 32% in both scenarios. Recoding helped to improve imputation efficiency for all the three covariates, this can be explained by the loss of precision after recoding. Surprisingly, the Average Correct Imputation Rates for Q14 (GUNSCHL) are very close in the two scenarios. In addition, they are consistently and considerably higher than those for the other two covariates. This may be explained by the fact that a vast majority of its observations fall into one category (figure 1).

Natural logarithmic transformation on the three covariates was also attempted before multiple imputation to approximate normal variables and to fit the distributional assumptions of the imputation model. The Average Correct Imputation Rates for Q10 and Q18 in original and recoded scales in both scenarios improved as compared to before the transformation, but still not satisfactory (below 53%). Nevertheless, the majority of Q14 (above 89%) in both scales have been correctly imputed.

Also examined was the effect of rounding on imputation efficiency, because the continuously distributed imputes have been rounded to the nearest category using a cutoff value 0.5 to preserve their ordinal property. For illustration an example is presented using a random dataset with missing covariates created in Scenario 8. The 50^{th} ~ 65^{th} observations of Q10 in this dataset are listed in table 6 along with their five imputed values in the same manner as in the simulations but without rounding the continuous imputed values. A large proportion (34 out of 50) of the imputed values is in different categories from their true values after being rounded using the cutoff value 0.5.

The prevalence of dissatisfaction, D2 = 1, ranges from 0.58% to 6.95% among the four race-gender groups. Interestingly, even with such low frequencies of the outcome (D2 = 1), all the covariates are significantly related to the outcome with odds of dissatisfaction with the trait present ranges from 1.42 to 2.27 times the same odds when the trait is absent. The three traits DRKPASS, GUNSCHL and FIGHTIN are strongly associated with each other with odds ratios between the traits ranging from 2.22 to 4.52. The significant associations between the four variables support these four variables as objects of our study of imputations on their values and whether imputation removes biases under these conditions.

In this study, CC showed smaller bias in the scenarios assuming MCAR for each covariate than in those with MAR and NI, regardless of proportions of missing covariates. This is consistent with the study by Allison (2000). The finding that the scenarios under NI showed relatively large biases in CC as compared to the MCAR conditions is also in accordance with King et al. (2001).

Table 6. Five Imputations for missing Q10 without rounding on imputed values from one random dataset in Scenario 8.

Obs.	Q10	True value	Imputation number 1	2	3	4	5
50	2	2					
51	.	1	1.7823 *	1.6022 *	1.7633 *	1.8180 *	1.3918
52	2	2					
53	.	1	1.3587	2.0277	1.8274 *	1.6079 *	1.4763
54	.	2	2.1264	2.4809	2.3249	2.0099	2.0358
55	.	1	1.7062 *	1.6104 *	1.6476 *	1.5978 *	1.6790 *
56	1	1					
57	.	1	1.4641	1.8700 *	1.5210 *	1.2140	1.5401 *
58	.	1	1.9022 *	1.8579 *	1.6802 *	1.7611 *	1.5634 *
59	1	1					
60	1	1					
61	.	1	1.5195 *	1.6148 *	1.5551 *	1.9029 *	1.5423 *
62	.	1	1.6313 *	1.6186 *	1.7034 *	1.4602	1.8294 *
63	.	1	1.6788 *	1.6553 *	1.6355 *	1.6657 *	1.5695 *
64	.	2	1.7022	2.0307	1.7366	1.4447 *	1.8448
65	1	1					

Accumulating evidence suggests that MI is usually better than, and almost always not worse than CC (Wu & Wu, 2001; Schafer, 1998; Allison, 2001; Little, 1992). Evidence provided by Schafer (1997, 2000) demonstrated that incomplete categorical (ordinal) data can often be imputed reasonably from algorithms based on a MVN model. However, our study did not show consistent results with the findings from Schafer, this is mainly due to ignorance of assumption of normality.

It is known that sensitivity to model assumptions is an important issue regarding the consistency and efficiency of normal maximum likelihood method applied to incomplete data. The improved, though unsatisfactory, imputation after natural logarithmic transformation presented a good demonstration of the importance of sensitivity to normal model assumption.

Moreover, normal ML methods do not guarantee consistent estimates, and they are certainly not necessarily efficient when the data are non-normal (Little, 1992). The MVN based MI procedure not specifically tailored to highly skewed ordinal data may have seriously distorted the ordinal variables' distributions or their relationship with other variables in our study, and therefore is not reliable when imputing highly skewed ordinal data.

It was suggested that and highly skewed variables may well be transformed to approximate normality (Tabachnick & Fidell, 2000). Nevertheless, highly skewed ordinal variables with only four or five values can hardly be transformed to nearly normal variables as shown by the unsatisfactory imputation efficiencies after natural logarithmic transformation. This study gives a warning that doing imputation without checking distributional assumptions of imputation model can lead to worse trouble than not imputing at all.

In addition, rounding after MI should be further explored in terms of appropriate cutoff values. One is cautioned that rounding could also bring its own bias into regression analysis in multiple imputations of categorical variables.

Conclusion

Applied researchers can be reasonably confident in utilizing CC to generate unbiased regression estimates even when large proportions of data missing completely at random. For ordinal variables with highly skewed distributions, MVN based MI cannot be expected to be superior to CC in generating unbiased regression estimates. It is cautionary that researchers doing imputation without checking distributional assumptions of imputation model can get into worse trouble than not imputing at all.

References

Allison P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2000). Multiple imputation for missing data: a cautionary tale. *Sociological Methods & Research, 28*, 301-9.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95* (1), 49-69.

Kolbe, L. J. (1990). An epidemiologic surveillance system to monitor the prevalence of youth behaviors that most affect health. *Journal of Health Education, 21*(6), 44-48.

Little R. J. A. (1992). Regression with missing X's: a review. *Journal of American Statistical Association*, 87, 1227-37.

Little, R. J. A., & Rubin D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18*, 292-326.

Little, R. J. A., & Rubin D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Patrician, P. A. (2002). Focus on research methods multiple imputation for missing data. *Research in Nursing & Health, 25*, 76-84.

Rubin D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91 (434), 473-89.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-92.

SAS/STAT Software (2002), Changes and Enhancements, Release 8.2, Cary, NC: SAS Institute Inc.

Schafer J. L., & Olsen M.K. (2000). Modeling and imputation of semicontinuous survey variables. Federal Committee on Statistical Methodology Research Conference: Complete Proceedings, 2000.

Schafer J. L. (1998). *The practice of multiple imputation*. Presented at the meeting of the 17. Methodology Center, Pennsylvania State University, University Park, PA.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, New York: Chapman & Hall.

Seligson J. L., Huebner E. S., & Valois R. F. (2003) Preliminary validation of the Brief Multidimensional Student's Life Satisfaction Scale (BMSLSS). *Social Indicators Research, 61*, 121-45.

Shah B. V., Barnwell G. B., & Bieler G. S. (1997). *SUDAAN*, software for the statistical analysis of correlated data, User's Manual. Release 7.5 ed. Research Triangle Park, NC: Research Triangle Institute.

Tabachnick B.G., & Fidell L.S. (2000). *Using multivariate statistics* (4th ed.). New York: HarperCollins College Publishers.

Valois, R. F., Zulling K. J., Huebner E. S., & Drane J. W. (2001). Relationship between life satisfaction and violent behaviors among adolescents. *American Journal of Health Behavior, 25*(4), 353-66.

Wu H., & Wu L. (2001). A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. *Statistics in Medicine*, 20, 1755-69.

Yuan, Y. C. (2000). Multiple imputation for missing data: concepts and new developments. *SAS Institute Technical Report*, 267-78.

Appendix A: 1997 SCYRBS Questionnaire items associated with the three covariates in regression analysis

Question 10 (Q10). During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?
1. 0 times
2. 1 time
3. 2 or 3 times
4. 4 or 5 times
5. 6 or more times

Question 14 (Q14). During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club on school property?
1. 0 days
2. 1 day
3. 2 or 3 days
4. 4 or 5 days
5. 6 or more days

Question 18 (Q18). During the past 12 months, how many times were you in a physical fight?
1. 0 times
2. 1 time
3. 2 or 3 times
4. 4 or 5 times
5. 6 or 7 times
6. 8 or 9 times
7. 10 or 11 times
8. 12 or more times

Appendix B: SAS Code

SAS PROC MI code for multiple imputation

```
proc mi data=first.c&I out=outmi&I seed=6666 nimpute=5
minimum=1 1 1 1 1 1 0  maximum=5 5 5 5 8 5 1 round=1 noprint;
em maxiter=1000 converge=1E-10;
mcmc impute=full initial=em prior=ridge=0.75 niter=500 nbiter=500;
freq weight;
var Q10 Q11 Q13 Q14 Q18 Q19 D2;
run;
```

Appendix C: SUDAAN Code

SUDAAN PROC MULTILOG code for multiple logistic regression analysis

```
Proc multilog data=stand filetype=sas design=wr noprint;
nest stratum psu;
weight weight;
subpopn sexrace=1  / name=”white female”;
subgroup D2 drkpass gunschl fightin;
levels 2 2 2 2;
reflevel drkpass=1 gunschl=1 fightin=1;
model D2 = drkpass gunschl fightin;
output beta sebeta/filename=junk_2 filetype=sas;
run;
```

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 3, No. 1, 300-311

1538 – 9472/05/$95.00

JMASM Algorithms and Code

JMASM16: Pseudo-Random Number Generation In R For Some Univariate Distributions

Hakan Demirtas
School of Public Health
University of Illinois at Chicago

An increasing number of practitioners and applied researchers started using the R programming system in recent years for their computing and data analysis needs. As far as pseudo-random number generation is concerned, the built-in generator in R does not contain some important univariate distributions. In this article, complementary R routines that could potentially be useful for simulation and computation purposes are provided.

Key words: Simulation; computation; pseudo-random numbers

Introduction

Following upon the work of Demirtas (2004), pseudo-random generation functions written in R for some univariate distributions are presented. The built-in pseudo-random number generator in R does not have routines for some important univariate distributions. Built-in codes are available only for the following univariate distributions: uniform, normal, chi-square, *t*, F, lognormal, exponential, gamma, Weibull, Cauchy, beta, logistic, stable, binomial, negative binomial, Poisson, geometric, hypergeometric and Wilcoxon.

The purpose of this article is to provide complementary R routines for generating pseudo-random numbers from some univariate distributions. In the next section, eighteen R functions of which the first thirteen correspond to the distributions that are not contained in the generator (Codes 1-13) are presented. The quality of the resulting variates have not been tested in the computer science sense. However, the first three moments for each distribution were rigorously tested. For the purposes of most applications, fulfillment of this criterion should be a reasonable approximation to reality. The last 5 functions (Codes 14-18) address already available univariate distributions; the reason for their inclusion is that variates generated with these routines are of a slightly better quality than those generated by the built-in code in terms of above-mentioned criterion.

Hakan Demirtas is an Assistant Professor of Biostatistics at the University of Illinois at Chicago. His research interests are the analysis of incomplete longitudinal data, multiple imputation and Bayesian computing. E-mail address: demirtas@uic.edu.

Functions for random number generation

The following abbreviations are used: *PDF* stands for the probability density function; *PMF* stands for the probability mass function; *CDF* stands for the cumulative distribution function; *GA* stands for the generation algorithm and *EAA* stands for an example of application areas; *nrep* stands for the number of identically and independently distributed random variates. The formal arguments other than *nrep* reflect the parameters in *PDF* or *PMF*. $E(X)$ and $V(X)$ denote the expectation and the variance of the random variable X, respectively.

Left truncated normal distribution

PDF: $$f(x \mid \mu,\sigma,\tau) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma(1-\Phi(\frac{\tau-\mu}{\sigma}))}$$

for $\tau \le x < \infty$ where $\Phi()$ is the standard normal *CDF*, μ, σ and τ are the mean, standard deviation and left truncation point, respectively. *EAA*: Modeling the tail behavior in simulation studies. *GA*: Robert's (1995) acceptance/ rejection algorithm with a shifted exponential as the majorizing density. For $\mu=0$ and $\sigma=1$, $E(X) = \frac{e^{-\tau^2/2}}{\sqrt{2\pi}(1-\Phi(\tau))}$, $V(X)$ is a complicated function of τ (see Code 1).

Left truncated gamma distribution

PDF:

$$f(x \mid \alpha,\beta) = \frac{1}{(\Gamma(\alpha)-\Gamma_{\tau/\beta}(\alpha))\beta^{\alpha}} x^{\alpha-1}e^{-x/\beta}$$

for $\tau \le x < \infty$, $\alpha>1$ and $min(\tau,\beta)>0$ where α and β are the shape and scale parameters, respectively, τ is the cutoff point at which truncation occurs and $\Gamma_{\tau/\beta}$ is the incomplete gamma function. *EAA*: Modeling left-censored data. *GA*: An acceptance/rejection algorithm (Dagpunar, 1978) where the majorizing density is chosen to be a truncated exponential.

$$E(X) = \beta[\frac{\Gamma(\alpha+1)-\Gamma_{\tau}(\alpha+1)}{\Gamma(\alpha)-\Gamma_{\tau}(\alpha)}],$$

$$V(X) = \beta^2[\frac{\Gamma(\alpha+2)-\Gamma_{\tau}(\alpha+2)}{\Gamma(\alpha)-\Gamma_{\tau}(\alpha)}] - E(X)^2.$$

The procedure works best when τ is small (see Code 2).

Code 1. Left truncated normal distribution:

```
draw.left.truncated.normal<-function(nrep,mu,sigma,tau){
if (sigma<=0){
stop("Standard deviation must be positive!\n")}
lambda.star<-(tau+sqrt(tau^2+4))/2
accept<-numeric(nrep) ; for (i in 1:nrep){
sumw<-0 ; while (sumw<1){
y<-rexp(1,lambda.star)+tau
gy<-lambda.star*exp(lambda.star*tau)*exp(-lambda.star*y)
fx<-exp(-(y-mu)^2/(2*sigma^2))/(sqrt(2*pi)*sigma*(1-pnorm((tau-
mu)/sigma)))
ratio1<-fx/gy ; ratio<-ratio1/max(ratio1)
u<-runif(1); w<-(u<=ratio) ; accept[i]<-y[w]; sumw<-sum(w)}}
accept}
```

Code 2: Left truncated gamma distribution

```
draw.left.truncated.gamma<-function(nrep,alpha,beta,tau){
if (tau<0){stop("Cutoff point must be positive!\n")}
if ((alpha<=1)){stop("Shape parameter must be greater than 1!\n")}
if ((beta<=0)){stop("Scale parameter must be positive!\n")}
y<-numeric(nrep); for (i in 1:nrep){
index<-0 ; scaled.tau<-tau/beta
lambda<-(scaled.tau-alpha+sqrt((scaled.tau-
alpha)^2+4*scaled.tau))/(2*scaled.tau)
while (index<1){
u<-runif(1); u1<-runif(1) ; y[i]<-(-log(u1)/lambda)+tau
w<-((1-lambda)*y[i]-(alpha-1)*(1+log(y[i])+log((1-lambda)/(alpha-
1)))<=-log(u))
index<-sum(w)}} ; y<-y*beta
y}
```

Laplace (double exponential) distribution

PDF: $f(x)=\frac{\lambda}{2}e^{-\lambda|x-\alpha|}$ for $\lambda>0$, where α and λ are the location and scale parameters, respectively. *EAA*: Monte Carlo studies of robust procedures, because it has a heavier tail than the normal distribution. *GA*: A sample from an exponential distribution with mean λ is generated, then the sign is changed with 1/2 probability and the resulting variates get shifted by α. $E(X)=\alpha$, $V(X)=2/\lambda^2$ (see Code 3).

Inverse Gaussian distribution

PDF:

$$f(x\mid\mu,\lambda)=\left(\frac{\lambda}{2\pi}\right)^{1/2}x^{-3/2}e^{-\frac{\lambda(x-\mu)^2}{2\mu^2x}}$$ for $x\geq 0$, $\mu>0$, $\lambda>0$, where μ and λ are the location and scale parameters, respectively. *EAA*: Reliability studies. *GA*: An acceptance/rejection algorithm developed by Michael et al. (1976). $E(X)=\mu$, $V(X)=\mu^3/\lambda$ (see Code 4).

Von Mises distribution

PDF: $f(x\mid K)=\frac{1}{2\pi I_0(K)}e^{K\cos(x)}$ for $-\pi\leq x\leq\pi$ and $K>0$, where $I_0(K)$ is a modified Bessel function of the first kind of order 0. *EAA*: Modeling directional data. *GA*: Acceptance/rejection method of Best and Fisher (1979) that uses a transformed folded Cauchy distribution as the majorizing density. $E(X)=0$ (see Code 5).

Zeta (Zipf) distribution

PDF: $f(x\mid\alpha)=\frac{1}{\zeta(\alpha)x^{\alpha}}$ for $x=1,2,3,\ldots$ and $\alpha>1$, where $\zeta(\alpha)=\sum_{x=1}^{\infty}x^{-\alpha}$ (Riemann zeta function). *EAA*: Modeling the frequency of random processes. *GA*: Acceptance/rejection algorithm of Devroye (1986).

$$E(X)=\frac{\zeta(\alpha-1)}{\zeta(\alpha)},$$

$$V(X)=\frac{\zeta(a)\zeta(a-2)-(\zeta(a-1))^2}{(\zeta(a))^2},$$

(see Code 6).

Code 3. Laplace (double exponential) distribution:

```
draw.laplace<-function(nrep, alpha, lambda){
if (lambda<=0){stop("Scale parameter must be positive!\n")}
y<-rexp(nrep,lambda)
change.sign<-sample(c(0,1), nrep, replace = TRUE)
y[change.sign==0]<--y[change.sign==0] ; laplace<-y+alpha
laplace}
```

Code 4. Inverse Gaussian distribution:

```
draw.inverse.gaussian<-function(nrep,mu,lambda){
if (mu<=0){stop("Location parameter must be positive!\n")}
if (lambda<=0){stop("Scale parameter must be positive!\n")}
inv.gaus<-numeric(nrep); for (i in 1:nrep){
v<-rnorm(1) ; y<-v^2
x1<-mu+(mu^2*y/(2*lambda))-(mu/(2*lambda))*(sqrt(4*mu*lambda*y+mu^2*y^2))
u<-runif(1) ; inv.gaus[i]<-x1
w<-(u>(mu/(mu+x1))) ; inv.gaus[i][w]<-mu^2/x1}
inv.gaus}
```

Code 5. Von Mises distribution:

```
draw.von.mises<-function(nrep,K){
if (K<=0){stop("K must be positive!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
u1<-runif(1) ; u2<-runif(1); u3<-runif(1)
tau<-1+(1+4*K^2)^0.5 ; rho<-(tau-(2*tau)^0.5)/(2*K)
r<-(1+rho^2)/(2*rho) ; z<-cos(pi*u1)
f<-(1+r*z)/(r+z) ; c<-K*(r-f)
w1<-(c*(2-c)-u2>0) ; w2<-(log(c/u2)+1-c>=0)
y<-sign(u3-0.5)*acos(f) ; x[i][w1|w2]<-y
index<-1*(w1|w2)}}
x}
```

Code 6. Zeta (Zipf) distribution

```
draw.zeta<-function(nrep,alpha){
if (alpha<=1){stop("alpha must be greater than 1!\n")}
zeta<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
u1<-runif(1) ; u2<-runif(1)
x<-floor(u1^(-1/(alpha-1))) ; t<-(1+1/x)^(alpha-1)
w<-x<(t/(t-1))*(2^(alpha-1)-1)/(2^(alpha-1)*u2)
zeta[i]<-x ; index<-sum(w)}}
zeta}
```

Logarithmic distribution

PMF: $f(x|\theta) = -\frac{\theta^x}{x log(1-\theta)}$ for x=1,2,3,... and 0<θ<1. *EAA*: Modeling the number of items processed in a given period of time. *GA*: The chop-down search method of Kemp (1981). $E(X) = \frac{-\theta}{(1-\theta)} \frac{1}{log(1-\theta)}$, $V(X) = \frac{-\theta log(1-\theta) - \theta^2}{(1-\theta)^2 (log(1-\theta))^2}$ (see Code 7).

Beta-binomial distribution

PMF: $f(x|n,\alpha,\beta) = \frac{n!}{x!(n-x)!B(\alpha,\beta)} \int_0^1 \pi^{\alpha-1+x} (1-\pi)^{n+\beta-1-x} d\pi$ for x=0,1,2,..., α>0 and β>0, where n is the sample size, α and β are the shape parameters and B(α,β) is the complete beta function. *EAA*: Modeling overdispersion or extravariation in applications where clusters of separate binomial distributions. *GA*: First π is generated as the appropriate beta and then it is used as the success probability in binomial. $E(X) = \frac{n\alpha}{\alpha+\beta}$, $V(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$ (see Code 8).

Code 7. Logarithmic distribution:

```
draw.logarithmic<-function(nrep,theta){
if ((theta<=0)|(theta>1)){stop("theta must be between 0 and 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; x0<-1 ; u<-runif(1)
while (index<1){t<--(theta^x0)/(x0*log(1-theta))
px<-t ; w<-(u<=px) ; x[i]<-x0 ; u<-u-px
index<-sum(w) ; x0<-x0+1}}
x}
```

Code 8. Beta-binomial distribution:

```
draw.beta.binomial<-function(nrep,alpha,beta,n){
if ((alpha<=0)|(beta<=0)){stop("alpha and beta must be positive!\n")}
if (floor(n)!=n){stop("Size must be an integer!\n")}
if (floor(n)<2){stop("Size must be greater than 2!\n")}
beta.variates<-numeric(nrep) ; beta.binom<-numeric(nrep)
for (i in 1:nrep){
beta.variates[i]<-rbeta(1,alpha,beta)
beta.binom[i]<-rbinom(1,n,beta.variates[i])}
beta.binom}
```

Rayleigh distribution

$$PDF: \quad f(x\,|\,\sigma)=\frac{x}{\sigma^2}e^{-x^2/2\sigma^2} \quad \text{for } x\geq 0$$

and $\sigma>0$, where σ is the scale parameter. *EAA*: Modeling spatial patterns. *GA*: The inverse *CDF* method. $E(X)=\sigma\sqrt{\pi/2}$, $V(X)=\sigma^2(4-\pi)/2$ (see Code 9).

Pareto distribution

$$PDF: \quad f(x|a,b)=\frac{ab^a}{x^{a+1}} \quad \text{for}$$

$0<b\leq x<\infty$ and $a>0$, where a and b are the shape and location parameters, respectively. *EAA*: Gene filtering in microarray experiments. *GA*: The inverse *CDF* method. $E(X)=\frac{ab}{a-1}$, $V(X)=\frac{ab^2}{(a-2)(a-1)^2}$. The procedure works best when a and b are not too small (see Code 10).

Non-central *t* distribution

Describes the ratio $\frac{Y}{\sqrt{U/\nu}}$ where U is a central chi-square random variable with ν degrees of freedom and Y is an independent normally distributed random variable with variance 1 and mean λ. *EAA*: Thermodynamic stability scores. *GA*: Based on arithmetic functions of normal and χ^2 variates. $E(X)=\lambda\sqrt{\nu/2}\,\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}$, $V(X)=(1+\lambda^2)\nu-\mathrm{E}(X^2)$ (see Code 11).

Code 9. Rayleigh distribution:

```
draw.rayleigh<-function(nrep,sigma){
if (sigma<=0){stop("Standard deviation must be positive!\n")}
u<-runif(nrep); rayl<-sigma*sqrt(-2*log(u))
rayl}
```

Code 10: Pareto distribution:

```
draw.pareto<-function(nrep,shape,location){
if (shape<=0){stop("Shape parameter must be positive!\n")}
if (location<=0){stop("Location parameter must be positive!\n")}
u<-runif(nrep) ; pareto<-location/(u^(1/shape))
pareto}
```

Code 11. Non-central *t* distribution:

```
draw.noncentral.t<-function(nrep,nu,lambda){
if (nu<=1){stop("Degrees of freedom must be greater than 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
x[i]<-rt(1,nu)+(lambda/sqrt(rchisq(1,nu)/nu))}
x}
```

Non-central chi-squared distribution

PDF:

$$f(x\mid\lambda,\nu)=\frac{e^{-(x+\lambda)/2}x^{\nu/2-1}}{2^{\nu/2}}\sum_{k=0}^{\infty}\frac{(\lambda x)^k}{4^k\,k!\,\Gamma(k+\nu/2)}$$

for $0\le x\le\infty$, $\lambda>0$ and $\nu>1$, where λ is the non-centrality parameter and ν is degrees of freedom. Both λ and ν can be non-integers. *EAA*: Wavelets in biomedical imaging. *GA*: Based on the sum of squared standard normal deviates. $E(X)=\lambda+\nu$, $V(X)=4\lambda+2\nu$ (see Code 12).

Doubly non-central *F* distribution

Describes the ratio of two scaled non-central χ^2 variables; that is, $F=\frac{X_1^2/n}{X_2^2/m}$ for $X_1^2\sim\chi^2(n,\lambda_1)$ and $X_2^2\sim\chi^2(m,\lambda_2)$, where n and m are numerator and denumerator degrees of freedom, respectively; λ_1 and λ_2 are the numerator and denumerator non-centrality parameters, respectively. *EAA*: Biomedical microarray studies. *GA*: Simple ratio of non-central χ^2 variables adjusted by corresponding degrees of freedom. $E(X)$ and $V(X)$ are too complicated to include here (see Code 13).

Code 12. Non-central chi-squared distribution:

```
draw.noncentral.chisquared<-function(nrep,df,ncp){
if (ncp<0){stop("Non-Centrality parameter must be non-negative!\n")}
if (df<=1){stop("Degrees of freedom must be greater than 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
df.int<-floor(df) ; df.frac<-df-df.int
mui<-sqrt(ncp/df.int) ; jitter<-0
if (df.frac!=0){jitter<-rchisq(1,df.frac)}
x[i]<-sum((rnorm(df.int)+mui)^2)+jitter}
x}
```

Code 13. Doubly non-central F distribution:

```
draw.noncentral.F<-function(nrep,df1,df2,ncp1,ncp2){
if (ncp1<0){stop("Numerator non-centrality parameter must be non-
negative!\n")}
if (ncp2<0){stop("Denominator non-centrality parameter must be non-
negative!\n")}
if (df1<=1){stop("Numerator degrees of freedom must be greater than
1!\n")}
if (df2<=1){
stop("Denominator degrees of freedom must be greater than 1!\n")}
x<-draw.noncentral.chisquared(nrep,df1,ncp1)/
draw.noncentral.chisquared(nrep,df2,ncp2)
x}
```

Standard *t* distribution

PDF:

$$f(x\mid\nu)=\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}(1+\frac{x^2}{\nu})^{-(\nu+1)/2} \text{ for}$$

$-\infty<x<\infty$, where ν is the degrees of freedom and Γ() is the complete gamma function. *GA*: A rejection polar method developed by Bailey (1994). $E(X)=0$, $V(X)=\frac{\nu}{\nu-2}$, (see Code 14).

Weibull distribution

PDF: $f(x\mid\alpha,\beta)=\frac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}$ for $0\le x<\infty$ and $min(\alpha,\beta)>0$, where α and β are the shape and scale parameters, respectively. *EAA*: Modeling lifetime data. *GA*: The inverse *CDF* method.

$E(X)=(1+1/\alpha)\beta$,

$V(X)=\left[\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)\right]\beta^2$,

(see Code 15).

Code 14. Standard **t** distribution:

```
draw.t<-function(nrep,df){
if (df<=1){stop("Degrees of freedom must be greater than 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
v1<-runif(1,-1,1) ; v2<-runif(1,-1,1); r2<-v1^2+v2^2
r<-sqrt(r2) ; w<-(r2<1)
x[i]<-v1*sqrt(abs((df*(r^(-4/df)-1)/r2)))
index<-sum(w)}}
x}
```

Code 15. Weibull distribution:

```
draw.weibull<-function(nrep, alpha, beta){
if ((alpha<=0)|(beta<=0)){
stop("alpha and beta must be positive!\n")}
u<-runif(nrep) ; weibull<-beta*((-log(u))^(1/alpha))
weibull}
```

Gamma distribution when $\alpha<1$

$$PDF: \quad f(x|\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta}$$

for $0\le x<\infty$, $min(\alpha,\beta)>0$, where α and β are the shape and scale parameters, respectively. *EAA*: Bioinformatics. *GA*: An acceptance/rejection algorithm developed by Ahrens and Dieter (1974) and Best (1983). It works when $\alpha<1$. $E(X)=\alpha\beta$, $V(X)=\alpha\beta^2$ (see Code 16).

Gamma distribution when $\alpha>1$

PDF: Same as before. *EAA*: Bioinformatics. *GA*: A ratio of uniforms method introduced by Cheng and Feast (1979). It works when $\alpha>1$. $E(X)=\alpha\beta$, $V(X)=\alpha\beta^2$ (see Code 17).

Code 16. Gamma distribution when $\alpha<1$

```
draw.gamma.alpha.less.than.one<-function(nrep,alpha,beta){
if (beta<=0){stop("Scale parameter must be positive!\n")}
if ((alpha<=0)|(alpha>=1)){
stop("Shape parameter must be between 0 and 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
u1<-runif(1) ; u2<-runif(1)
t<-0.07+0.75*sqrt(1-alpha) ; b<-1+exp(-t)*alpha/t
v<-b*u1 ; w1<-(v<=1) ; w2<-(v>1)
x1<-t*(v^(1/alpha)) ; w11<-(u2<=(2-x1)/(2+x1))
w12<-(u2<=exp(-x1)) ; x[i][w1&w11]<-x1[w1&w11]
x[i][w1&!w11&w12]<-x1[w1&!w11&w12]
x2=-log(t*(b-v)/alpha) ; y<-x2/t
w21<-(u2*(alpha+y*(1-alpha))<=1)
w22<-(u2<=y^(alpha-1)) ; x[i][w2&w21]<-x2[w2&w21]
x[i][w2&!w21&w22]<-x2[w2&!w21&w22]
index<-1*(w1&w11)+1*(w1&!w11&w12)+1*(w2&w21)+1*(w2&!w21&w22)}}
x<-beta*x
x}
```

Code 17. Gamma distribution when $\alpha>1$:

```
draw.gamma.alpha.greater.than.one<-function(nrep,alpha,beta){
if (beta<=0){stop("Scale parameter must be positive!\n")}
if (alpha<=1){stop("Shape parameter must be greater than 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
u1<-runif(1); u2<-runif(1)
v<-(alpha-1/(6*alpha))*u1/((alpha-1)*u2)
w1<-((2*(u2-1)/(alpha-1))+v+(1/v)<=2)
w2<-((2*log(u2)/(alpha-1))-log(v)+v<=1)
x[i][w1]<-(alpha-1)*v ; x[i][!w1&w2]<-(alpha-1)*v
index<-1*w1+1*(!w1&w2)}}
x<-x*beta
x}
```

Beta distribution when *max*(α,β)<1

PDF:

$$f(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$ for

$0 \le x \le 1$, $0 \le \alpha < 1$ and $0 \le \beta < 1$, where α and β are the shape parameters and $B(\alpha,\beta)$ is the complete beta function. *EAA*: Analysis of biomedical signals. *GA*: An acceptance/rejection algorithm developed by Johnk (1964). It works when both parameters are less than 1. $E(X)=\frac{\alpha}{\alpha+\beta}$,

$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ (see Code 18).

found to be negligible, suggesting that random number generation routines presented are accurate. These routines could be a handy addition to a practitioner's set of tools given the growing interest in R. However, the reader is invited to be cautious about the following issues: 1) It is not postulated that algorithms presented are the most efficient. Furthermore, implementation of a given algorithm may not be optimal. Given sufficient time and resources, one can write more efficient routines. 2) Quality of every random number generation process depends on the uniform number generator.

Code 18. Beta distribution when *max*(α,β)<1:

```
draw.beta.alphabeta.less.than.one<-function(nrep,alpha,beta){
if ((alpha>=1)|(alpha<=0)|(beta>=1)|(beta<=0)) {
stop ("Both shape parameters must be between 0 and 1!\n")}
x<-numeric(nrep) ; for (i in 1:nrep){
index<-0 ; while (index<1){
u1<-runif(1) ; u2<-runif(1)
v1<-u1^(1/alpha) ; v2<-u2^(1/beta)
summ<-v1+v2 ; w<-(summ<=1)
x[i]<-v1/summ ; index<-sum(w)}}
x}
```

Results for arbitrarily chosen parameter values

For each distribution, the parameters can take infinitely many values and first two moments virtually fluctuate on the entire real line. The quality of random variates was tested by a broad range of simulations to see any potential aberrances and abnormalities in some subset of the parameter domains and to avoid any selection biases. The empirical and theoretical moments for arbitrarily chosen parameter values are reported in Table 1 and 2.

Table 1 tabulates the theoretical and empirical means for each distribution for arbitrary values. Throughout the table, the number of replications (*nrep*) is chosen to be 10,000. A similar comparison is made for the variances, as shown in Table 2. In both tables, the deviations from the expected moments are

McCullough (1999) raised some questions about the quality of Splus generator. At the time of this writing, a source that tested the R generator is unknown to the author. In addition, the differences between empirical and distributional moments have merely been examined for each distribution. More comprehensive and computer science-minded tests are needed possibly using DIEHARD suite (Marsaglia, 1995) or other well-regarded test suites.

Table 1: Comparison of theoretical and empirical means for arbitrarily chosen parameter values.

Distribution	*Parameter(s)*	*Theoretical mean*	*Empirical mean*
Left truncated normal	$\mu=0$, $\sigma=1$, $\tau=0.5$	1.141078	1.143811
Left truncated gamma	$\alpha=4$, $\beta=2$, $\tau=0.5$	8.002279	8.005993
Laplace	$\alpha=4$, $\lambda=2$	4	3.999658
Inverse Gaussian	$\mu=1$, $\lambda=1$	1	1.001874
Von Mises	$K=10$	0	0.002232
Zeta (Zipf)	$\alpha=4$	1.110626	1.109341
Logarithmic	$\theta=0.6$	1.637035	1.637142
Beta-binomial	$\alpha=2$, $\beta=3$, $n=10$	4	4.016863
Rayleigh	$\sigma=4$	5.013257	5.018006
Pareto	$a=5$, $b=5$	6.25	6.248316
Non-central t	$\nu=5$, $\lambda=1$	1.189416	1.191058
Non-central Chi-squared	$\nu=5$, $\lambda=2$	7	7.004277
Doubly non-central F	$n=5$, $m=10$, $\lambda_1=2$, $\lambda_2=3$	0.667381	0.666293
Standard t	$\nu=5$	0	0.001263
Weibull	$\alpha=5$, $\beta=5$	4.590844	4.587294
Gamma with $\alpha<1$	$\alpha=0.3$, $\beta=0.4$	0.12	0.118875
Gamma with $\alpha>1$	$\alpha=3$, $\beta=0.4$	1.2	1.200645
Beta with $\alpha<1$ and $\beta<1$	$\alpha=0.7$, $\beta=0.4$	0.636363	0.636384

Table 2: Comparison of theoretical and empirical variances for arbitrarily chosen parameter values.

Distribution	*Parameter(s)*	*Theoretical variance*	*Empirical variance*
Left truncated normal	$\mu=0$, $\sigma=1$, $\tau=0.5$	0.603826	0.602914
Left truncated gamma	$\alpha=4$, $\beta=2$, $\tau=0.5$	15.98689	15.86869
Laplace	$\alpha=4$, $\lambda=2$	0.5	0.502019
Inverse Gaussian	$\mu=1$, $\lambda=1$	1	0.997419
Zeta (Zipf)	$\alpha=4$	0.545778	0.556655
Logarithmic	$\theta=0.6$	1.412704	1.4131545
Beta-binomial	$\alpha=2$, $\beta=3$, $n=10$	6	6.001696
Rayleigh	$\sigma=4$	6.867259	6.854438
Pareto	$a=5$, $b=5$	2.604167	2.604605
Non-central t	$\nu=5$, $\lambda=1$	1.918623	1.903359
Non-central Chi-squared	$\nu=5$, $\lambda=2$	18	18.09787
Doubly non-central F	$n=5$, $m=10$, $\lambda_1=2$, $\lambda_2=3$	0.348817	0.346233
Standard t	$\nu=5$	1.666667	1.661135
Weibull	$\alpha=5$, $\beta=5$	1.105749	1.098443
Gamma with $\alpha<1$	$\alpha=0.3$, $\beta=0.4$	0.048	0.047921
Gamma with $\alpha>1$	$\alpha=3$, $\beta=0.4$	0.48	0.481972
Beta with $\alpha<1$ and $\beta<1$	$\alpha=0.7$, $\beta=0.4$	0.110193	0.110126

References

Ahrens, J. H., & Dieter, U. (1974). Computer methods for sampling from gamma, beta, poisson and binomial distributions. Computing, 1, 223-246.

Bailey, R. W. (1994). Polar generation of random variates with the t-distribution. *Mathematics of Computation, 62*, 779-781.

Best, A. W. (1983). A note on gamma variate generators with shape parameter less than unity. *Computing, 30*, 185-188.

Best, D. J., & Fisher, N. I. (1979). Efficient simulation of the von mises distribution. *Applied Statistics, 28*, 152-157.

Cheng, R. C. H., & Feast, G. M. (1979). *Some simple gamma variate generation*. *Applied Statistics, 28*, 290-295.

Dagpunar, J. S. (1978). Sampling of variates from a truncated gamma distribution. *Journal of Statistical Computation and Simulation, 8*, 59-64.

Demirtas, H. (2004). Pseudo-random number generation in r for commonly used multivariate distributions. *Journal of Modern Applied Statistical Methods*, 3, 385-497.

Devroye, L. (1986). *Non-Uniform random variate generation*. New York: Springer-Verlag.

Jhonk, M. D. (1964). Erzeugung von betaverteilter und gammaverteilter zufallszahlen. *Metrika, 8*, 5-15.

Kemp, A. W. Efficient generation of logarithmically distributed pseudo-random variables. *Applied Statistics, 30*, 249-253.

Marsaglia, G. (1995). *The marsaglia random number cdrom, including the diehard battery of tests of randomness*. Department of Statistics, Florida State University, Tallahassee, Florida.

McCullough, B. D. (1999). Assessing the reliability of statistical software: Part 2. *The American Statistician, 53*, 149-159.

Michael, J. R., William, R. S., & Haas, R. W. (1976). Generating random variates using transformations with multiple roots. The American Statistician, 30, 88-90.

Robert, C. P. (1995). Simulation of truncated random variables. *Statistics and Computing, 13*, 169-184.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 312-318

1538 – 9472/05/$95.00

JMASM17: An Algorithm And Code For Computing Exact Critical Values For Friedman's Nonparametric ANOVA

Sikha Bagui Subhash Bagui
The University of West Florida, Pensacola

Provided in this article is an algorithm and code for computing exact critical values (or percentiles) for Friedman's nonparametric rank test for k related treatment populations using Visual Basic (VB.NET). This program has the ability to calculate critical values for any number of treatment populations (k) and block sizes (b) at any significance level (α). We developed an exact critical value table for $k = 2(1)5$ and $b = 2(1)15$. This table will be useful to practitioners since it is not available in standard nonparametric statistics texts. The program can also be used to compute any other critical values.

Key words: Friedman's test, randomized block designs (RBD), ANOVA, Visual Basic

Introduction

While experimenting (or dealing) with randomized block designs (RBDs) (or one-way repeated measures designs), if the normality of treatment populations or the assumptions of equal variances are not met, or the data are in ranks, it is recommended that Friedman's rank-based nonparametric test be used as an alternative to the conventional F test for the RBD (or one-way repeated measures analysis of variance) for k related treatment populations. This test was developed by Friedman (1937), and was designed to test the null hypothesis that all the k treatment populations are identical versus the alternative that at least two of the treatment populations differ in location. This test is based on a statistic that is a rank analogue of SST (total sum of squares) for the RBD and is computed in the following manner. After the data from a RBD are obtained, the observed values in each block b are ranked from 1 (the smallest in the block) to k (the largest in the block). Let R_i denote the sum of the ranks of the values corresponding to treatment population i, $i = 1, 2, \cdots, k$. Then the Friedman's test statistic is given by

$$F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} R_i^2 - 3b(k+1).$$

If the null hypothesis is true, it is expected that the rankings be randomly distributed within each block. If that is the case, the sum of the rankings in each treatment population will be approximately equal, and the resulting value of F_r will be small.

If the alternative hypothesis is true, the expectation is that this will to lead to differences among the R_i values and obtain correspondingly large values of F_r. Thus, the null hypothesis is rejected in favor of the alternative hypothesis for large values of F_r. Exact null sampling distribution of F_r is not known. But, as with the Kruskal-Wallis (1952) statistic, the null distribution of the Friedman's F_r can be approximated by a chi-square (χ^2) distribution

Sikha Bagui is an Assistant Professor in the Department of Computer Science. Her areas of research are database and database design, data mining, pattern recognition, and statistical computing. Email: bagui@uwf.edu. Subhash Bagui is a Professor in the Department of Mathematics and Statistics. His areas of research are statistical classification and pattern recognition, bio-statistics, construction of designs, tolerance regions, statistical computing and reliability. Email: sbagui@uwf.edu.

with $(k-1)$ degrees of freedom as long as b is large. Empirical evidence indicates that the approximation is adequate if the number of blocks b and /or the number of treatment populations k exceeds 5.

Again, in small sample situations, the chi-square approximation will not be adequate. Common statistics books with a chapter on nonparametric statistics do not provide exact critical values for Friedman's F_r test. Conover (1999) did not provide exact critical values for the Friedman's F_r test, but Hollander and Wolf (1973) and Lehmann (1998) provided a partial exact critical values table for Friedman's F_r test. Most commonly used statistical software such as MINITAB and SPSS provide only the asymptotic P-value for Friedman's F_r statistic. In view of this, in this article, we provide a VB.NET program that computes the exact critical values of Friedman's F_r statistic for any number of blocks b and any number of treatment populations k, at any significance level (α).

Also provided is an exact critical values table for the Friedman's F_r test for various combinations of (small) block sizes b and (small) treatment population sizes k. Headrick (2003) wrote an article for generating exact critical values for the Kruskal-Wallis nonparametric test using Fortran 77. We used his idea to generate exact critical values for Friedman's F_r test using VB.NET. VB.NET is user friendly and more accessible. Our VB.NET program works well with reasonable values of b and k.

Methodology

In order to generate the critical values of Friedman's F_r statistic, we need to have the null distribution for F_r. In Friedman's test, the null hypothesis is that the effect of all k treatment populations are identical. Thus it is reasonable to use such types of null distributions for the F_r statistic that are derived under the assumption that all observations for treatment populations are from the same population.

Therefore, to find the null distribution of the F_r statistic, first, generate b uniform pseudo-random numbers from the interval (0,1) for each of the k treatment populations. Assume that the probability of a tie is zero. Then random variates within each block are ranked from 1 to k. The program then calculates rank sums of each treatment population, R_i, and computes the value of the F_r statistic

$$F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} R_i^2 - 3b(k+1).$$

This process is replicated a sufficient number of times until the null distribution of the F_r statistic is modeled adequately. Then the program returns a critical value that is associated with a percentile fraction of 0.90, 0.95, 0.975, or 0.99 (or equivalently a significance level alpha of 0.10, 0.05, 0.025, or 0.01). In some cases returned values may be true for a range of P-values.

With adequate number of runs, this VB.NET program yields the same values reported by Lehmann (1998) in Table M. In Table 1 below, we provide critical values for the F_r test for $b = 2(1)15$, $k = 2(1)5$ and $\alpha = 0.1$, 0.05, 0.025, 0.01. The notation $F_{1-\alpha}$ in the Table 1 means $(1-\alpha)100\%$ percentile of the F_r statistic which is equivalent to α level critical value of the F_r statistic. This table will be useful to the practitioners since it is not available in standard statistics texts with a chapter on nonparametric statistics. The critical values in Table 1 are generated using 1 million replications in each case.

Table 1. Critical values for Friedman's F_r test.

Rows (b)	Columns (k)	$F_{0.90}$	$F_{0.95}$	$F_{0.975}$	$F_{0.99}$
2	2	2.0000	2.0000	2.0000	2.0000
3	2	3.0000	3.0000	3.0000	3.0000
4	2	4.0000	4.0000	4.0000	4.0000
5	2	1.800	5.0000	5.0000	5.0000
6	2	2.6667	2.6667	2.6667	2.6667
7	2	3.5714	3.5714	3.5714	7.0000
8	2	2.0000	4.5000	4.5000	4.5000
9	2	2.7778	2.7778	5.4444	5.4444
10	2	3.6000	3.6000	3.6000	6.4000
11	2	2.2727	4.4545	4.4545	7.3636
12	2	3.000	3.0000	5.3333	5.3333
13	2	1.9231	3.7692	3.7692	6.2308
14	2	2.5714	4.5714	4.5714	7.1429
15	2	3.2667	3.2667	5.4000	5.4000
2	3	4.0000	4.0000	4.0000	4.0000
3	3	4.6667	4.6667	6.0000	6.0000
4	3	4.5000	6.0000	6.5000	6.5000
5	3	4.8000	5.2000	6.4000	7.6000
6	3	4.3333	6.3333	7.0000	8.3333
7	3	4.5714	6.0000	7.1429	8.0000
8	3	4.7500	5.2500	7.0000	7.7500
9	3	4.6667	6.0000	6.8889	8.6667
10	3	4.2000	5.6000	7.4000	8.6000
11	3	4.9091	5.6364	7.0909	8.9091
12	3	4.6667	6.1667	7.1667	8.6667
13	3	4.5841	5.9469	7.2920	9.0796
14	3	4.4286	5.5714	7.0000	9.0000
15	3	4.8000	5.7333	6.9333	8.5333
2	4	5.4000	5.4000	6.0000	6.0000
3	4	5.8000	7.0000	7.4000	8.2000
4	4	6.0000	7.5000	8.1000	9.3000
5	4	6.1200	7.3200	8.2800	9.7200
6	4	6.2000	7.4000	8.6000	10.0000
7	4	6.2571	7.6286	8.6571	10.3714
8	4	6.1500	7.5000	8.8500	10.3500
9	4	6.0667	7.5333	8.7333	10.4667
10	4	6.2400	7.5600	8.8800	10.6800
11	4	6.1636	7.5818	8.8909	10.6364
12	4	6.1000	7.6000	9.0000	10.7000
13	4	6.0462	7.6154	9.0000	10.7539
14	4	6.2571	7.6286	9.0000	10.8857
15	4	6.2000	7.5600	9.0800	10.7600
2	5	6.8000	7.2000	7.6000	7.6000
3	5	7.2000	8.2667	9.3333	9.8667
4	5	7.4000	8.6000	9.6000	11.0000
5	5	7.5200	8.8000	10.0800	11.5200
6	5	7.6000	8.9333	10.2667	11.8667
7	5	7.6571	9.0286	10.5143	12.0043
8	5	7.7000	9.2000	10.6400	12.2000
9	5	7.6444	9.1556	10.5778	12.3556
10	5	7.6800	9.2000	10.6400	12.4000
11	5	7.7091	9.2363	10.7636	12.5818
12	5	7.6667	9.2667	10.7333	12.5333
13	5	7.6923	9.2923	10.7692	12.7385
14	5	7.7143	9.3143	10.8000	12.7429
15	5	7.6800	9.3333	10.8267	12.7467

Conclusion

In case of large values of b and k, the program needs a large number of replications in order to adequately model the null distribution for Friedman's F_r statistic. The replication numbers should be in increasing order such as 10,000, 50,000, 100,000, 500,000, and 1,000,000 etc. and the process stopped once two consecutive values are almost the same. If there are b blocks and k treatment populations, then at least $(k!)^b$ are necessary for a near fit for the F_r statistic.

For a good fit of F_r, one needs many more replications than $(k!)^b$. The VB.NET program is given in the Appendix. This program allows the user to provide values for replication numbers, block sizes, treatment population numbers, and the percentile fractions. Based on this, the program will return a critical value. Also, remember that distribution of Friedman's F_r statistic is discrete, so it not possible to achieve the exact level of significance. Thus the critical values obtained here correspond to approximate level of significance 0.01, 0.025, 0.05, and 0.10.

References

Conover, W. J. (1999). *Practical Nonparametric Statistics*. (3rd ed.). New York: Wiley.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32,* 675-701.

Headrick, T. C. (2003). An algorithm for generating exact critical values for the Kruskal-Wallis One-way ANOVA. *Journal of Modern Applied Statistical Methods*, 2, 268-271.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion analysis of variance. *Journal of American Statistical Association*, 47, 583-621.

Lehmann, E. L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*. New Jersey: Prentice Hall.

Minitab. (2000). *Minitab for Windows, release* 13.3, Minitab, Inc., State College, PA.

SPSS. (2002). *SPSS for Windows, version* 11.0, SPSS, Inc., Chicago, IL.

Appendix:

```
Imports System.Windows.Forms

Public Class Form1
    Inherits System.Windows.Forms.Form

Dim sum = 0, squared = 0, square_sum = 0, m = 0, n = 0, i = 0, j = 0, k = 0,
l = 0, p = 0, q = 0, r As Integer
    Dim count = 0, v = 0, z As Integer
    Dim num As Single
    Dim percentile As Single
    Dim f As Single
    Dim file1 As System.IO.StreamWriter
```

```
Dim array1(,) As Single = New Single(,) {}
    Dim array6(,) As Single = New Single(,) {}
    Dim array3() As Single = New Single() {}
    Dim array4() As Single = New Single() {}
    Dim array5() As Integer = New Integer() {}
    Dim array7() As Integer = New Integer() {}
    Dim array8() As Single = New Single() {}
    Dim array9() As Single = New Single() {}

    Private Sub Form1_Load(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles MyBase.Load
        'Calling the random number generator
        Randomize()
    End Sub

    Private Sub Button1_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Button1.Click
        m = Val(TextBox1.Text) 'm is the number of rows(blocks)
        n = Val(TextBox2.Text)  'n is the number of columns
        z = Val(TextBox3.Text)  'z is the number of runs
        percentile = Val(TextBox4.Text)  'percentile value

        'Defining the arrays
        Dim array1(m, n) As Single
        Dim array6(m, n) As Single
        Dim array3(n) As Single
        Dim array4(n) As Single
        Dim array5(n) As Integer
        Dim array7(n) As Integer
        Dim array8(z) As Single
        Dim array9(z) As Single

        Dim row, col As Integer
        Dim output As String

        For p = 1 To z
            output = " "
            'creating initial m x n random array
            For row = 1 To m
                For col = 1 To n
                    array1(row, col) = Rnd()
                    output &= array1(row, col) & " "
                Next
                output &= vbCrLf
            Next

            j = 1
            k = 1
            For r = 1 To array1.GetUpperBound(0)

                'pulling out one row
                For col = 1 To n 'array1.GetUpperBound(0)
                    num = array1(j, col)
                    array3(col) = num
                    output &= array3(col) & " "
                Next
```

```
    j = j + 1

    'copying one row into new array
    For row = 1 To array3.GetUpperBound(0)
        array4(row) = array3(row)
    Next

    'sorting one row
    Array.Sort(array3)  'Array3 - is sorted array

    'ranking row
    For row = 1 To array4.GetUpperBound(0)
        For i = 1 To array3.GetUpperBound(0)
            If array4(row) = array3(i) Then
                array5(row) = i
            End If
        Next
    Next

    'putting row back into two dimensional array
    For row = 1 To array5.GetUpperBound(0)
        output &= array5(row) & " "
        array6(k, row) = array5(row)
    Next
    output &= vbCrLf
    k = k + 1
Next
output = " "

'displaying two dimensional array
For row = 1 To array6.GetUpperBound(0)
    For col = 1 To n 'array6.GetUpperBound(0)
        output &= array6(row, col) & " "
    Next
    output &= vbCrLf
Next

'summing columns in two dimensional array
l = 1
sum = 0
square_sum = 0
For col = 1 To n 'array6.GetUpperBound(0)
    For row = 1 To array6.GetUpperBound(0)
        sum += array6(row, l)
    Next
    output = sum
    square_sum = sum * sum
    array7(l) = square_sum
    output &= vbCrLf
    output &= array7(l) & " "
    l = l + 1
    sum = 0
    square_sum = 0
```

```
    Next

            f = 0
            squared = 0

= 1 To array7.GetUpperBound(0)
                squared += array7(row)
            Next
            output = squared
            output = " "
            f = Convert.ToSingle(12 / (m * n * (n + 1)) * squared - 3 * m *
            (n + 1))
            array8(p) = f
            output &= array8(p) & " "
            f = 0
            squared = 0
        Next
        output = " "

        For row = 1 To array8.GetUpperBound(0)
            output &= array8(row) & " "
        Next

        For row = 1 To array8.GetUpperBound(0)
            array9(row) = array8(row)
            output &= array9(row) & " "
        Next

        Array.Sort(array9) 'Array9 - sorted F values

        For row = 1 To array9.GetUpperBound(0)
            output &= array9(row) & " "
        Next

        output = " "

        count = 0
        For row = 1 To array9.GetUpperBound(0)
            count += 1
        Next

        output = count

        v = percentile * count

        output = " "

        output = array9(v)

        MessageBox.Show(output, "95% percentile value")

    End Sub

End Class
```

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 319-332

1538 – 9472/05/$95.00

An Algorithm For Generating Unconditional Exact Permutation Distribution For A Two-Sample Experiment

J. I. Odiase S. M. Ogbonmwan
Department of Mathematics
University of Benin, Nigeria

An Algorithm that generates the unconditional exact permutation distribution of a 2 x n experiment is presented. The algorithm is able to handle ranks as well as actual observations. It makes it possible to obtain exact p-values for several statistics, especially when sample sizes are small and the application of large sample approximation is unreliable. An illustrative implementation is achieved and leads to the computation of exact p-values for the Mood test when the sample size is small.

Key words: permutation test, Monte Carlo test, p-value, rank order statistic, Mood test

Introduction

An important part of Statistical Inference is the representation of observed data in terms of a p-value. In fact, the p-value plays a major role in determining whether to accept or reject the null hypothesis. The p-value assists in establishing whether the observed data are statistically significant and so, any statistical approach that will guarantee its proper computation should be developed and employed in inferential statistics so that the probability of making a type I error is exactly α.

In practice, data are usually collected under varied conditions with some distributional assumptions such as that the data came from a normal distribution. It is advisable to avoid as much as possible making so many distributional assumptions because data are usually never collected under ideal or perfect conditions, that is, do not conform perfectly to an assumed distribution or model being employed in its analysis. The p-value obtained through the permutation approach turns out to be the most reliable because it is exact, see Agresti (1992) and Good (2000).

If the experiment to be analyzed is made up of small or sparse data, large sample procedures for statistical inference are not appropriate (Senchaudhuri et al., 1995; Siegel & Castellan, 1988). In this article, consideration is given to the special case of 2 x n tables with row and column totals allowed to vary with each permutation – this seems more natural than fixing the row and column totals. This is the unconditional exact permutation approach which is all-inclusive rather than the constrained or conditional exact permutation approach of fixing row and column totals. This later approach mainly addresses contingency tables (Agresti, 1992).

Several approaches have been suggested as alternatives to the computationally intensive unconditional exact permutation see Fisher (1935) and Agresti (1992) for a discussion on exact conditional permutation distribution. Also see Efron (1979), Hall and Tajvidi (2002), Efron and Tibshirani (1993), Opdyke (2003) for Monte Carlo approaches. Other approaches like the Bayesian and the likelihood have also been found useful in obtaining exact permutation

J. I. Odiase (justiceodiase@yahoo.com) is a Lecturer in the Department of Mathematics, University of Benin, Nigeria. His research interests include statistical computing and nonparametric statistics. S. M. Ogbonmwan (ogbonmwasmaltra@yahoo.co.uk) is an Associate Professor of Statistics, Department of Mathematics, University of Benin, Nigeria. His research interests include statistical computing and nonparametric statistics.

distribution (Bayarri & Berger, 2004; Spiegelhalter, 2004).

Large sample approximations are commonly adopted in several nonparametric tests as alternatives to tabulated exact critical values. The basic assumption required for such approximations to be reliable alternatives is that the sample size should be sufficiently large. However, there is no generally agreed upon definition of what constitutes a large sample size (Fahoome, 2002).

Available software for exact inference is expensive, with varied restrictions in the implementation of exact permutation procedures in the software. Computational time is highly prohibitive even with very fast processor speed of available personal computers. R. A. Fisher compiled by hand 32,768 permutations of Charles Darwin's data on the height of cross-fertilized and self-fertilized zea mays plants. The enormity of this task possibly discouraged Fisher from probing further into exact permutation tests (Ludbrook & Dudley, 1998).

Permutation tests provide exact results, especially when complete enumeration is feasible. A comprehensive documentation of the properties of permutation tests can be found in Pesarin (2001). The problem with permutation tests has been high computational demands, viz space and time complexities. Sampling from the permutation sample space rather than carrying out complete enumeration of all possible distinct rearrangements is what most of the available permutation procedures do, see Opdyke (2003) for a detailed listing of widely available permutation sampling procedures.

Opdyke (2003) however observed that most of the existing procedures can perform Monte Carlo sampling without replacement within a sample, but none can avoid the possibility of drawing the same sample more than once, thereby reducing the power of the permutation test.

The purpose of this article is to fashion out a sure and efficient way of obtaining unconditional exact permutation distribution by ensuring that a complete enumeration of all the distinct permutations of any 2-sample experiment is achieved. This will produce exact p-values and therefore ensure that the probability of making a type I error is exactly α. This article also provides computer algorithms for achieving complete enumeration.

Methodology

Good (2000) considered the tails of permutation distribution in order to arrive at p-values, though he never carried out complete enumeration required for a permutation test. This approach has no precise model for the tail of the distribution from which data are drawn, (Hall & Weissman, 1997). The five steps for a permutation test presented in Good (2000) can be summarized thus:

1. Analyze the problem.
2. Choose a statistic and establish a rejection rule that will distinguish the hypothesis from the alternative.
3. Compute the test statistic for the original observations.
4. Rearrange the observations, compute the test statistic for every new arrangement and repeat this process until all permutations are obtained.
5. Construct the distribution for the test statistic based on Step 4.

Step 4 is where the difficulty in permutation test lies because a complete enumeration of all distinct permutations of the experiment is required. A 2-sample experiment with 15 variates in each sample requires 155,117,520 permutations. Clearly, the enumeration cannot be done manually, even if the computer produces 1000 permutations in a second, over 43 hours will be required for a complete enumeration. When this is achieved, p-values can be computed. Good (2000) identified the sufficient condition for a permutation test to be exact and unbiased against shifts in the direction of higher values as the exchangeability of the observations in the combined sample.

Let $X_i = \left(x_{i1}, x_{i2}, \cdots, x_{in_i}\right)^T$, i = 1, 2 and n_i is the i^{th} sample size. Also, let $X_N = (X_1, X_2)$, where $N = n_1 + n_2$. X_N is composed of N independent and identically distributed random variables. We have $\frac{(n_1 + n_2)!}{n_1! n_2!} = \frac{N!}{n_1! n_2!}$

possible permutations of the N variates of the 2 samples of size *n*, i = 1, 2 which are equally likely, each having the probability $\left(\frac{N!}{n_1! n_2!}\right)^{-1}$.

For equal sample sizes, $n = n_1 = n_2$, the number of permutations = $\frac{(2n)!}{(n!)^2}$ or $\frac{N!}{(n!)^2}$ and the probability of each permutation = $\frac{(n!)^2}{N!}$.

For all possible permutations of the N variates, systematically develop a pattern necessary for the algorithm required for the generation of all the distinct permutations. The presentation of the systematic generation of all the possible permutations of the N variates now follows.

Examine an experiment of two samples (treatments), each with two variates, i.e., $\begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{pmatrix}$, where x_{11}, x_{12}, x_{21} and x_{22} represent sample values. Number of distinct arrangements = $\frac{4!}{2!\,2!} = 6$ (permutations) as listed in Table 1.

Table 1: Permutations of a 2 x 2 Experiment.

1	2	3	4	5	6
x_{11} x_{21} x_{12} x_{22}	x_{21} x_{11} x_{12} x_{22}	x_{22} x_{21} x_{12} x_{11}	x_{11} x_{12} x_{21} x_{22}	x_{11} x_{21} x_{22} x_{12}	x_{21} x_{11} x_{22} x_{12}

Numbers 1 – 6 on top of the permutations represent the permutation numbers

The actual process of permuting the variates of the experiment reveals the following.

$\begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{pmatrix}$ original arrangement of the experiment 1 permutation

$x_{11} \leftarrow x_{2i}$, i = 1, 2 2 permutations

$x_{12} \leftarrow x_{2i}$, i = 1, 2 2 permutations

$\begin{pmatrix} x_{21} & x_{11} \\ x_{22} & x_{12} \end{pmatrix}$ exchange the samples (columns) 1 permutation

In an attempt to offer a mathematical explanation for the method of exchanges of variates leading to the algorithm, observe that

$\binom{2}{0}\binom{2}{0} = 1 \text{ x } 1 = 1$ Permutation (original arrangement of the experiment)

$\binom{2}{1}\binom{2}{1} = 2 \text{ x } 2 = 4$ Permutations (using one variate from first sample)

$\binom{2}{2}\binom{2}{2} = 1 \text{ x } 1 = 1$ Permutation (exchange the samples, i.e., 2 variates)

Total = 1 + 4 + 1 = 6

Observe that permutation (1) is the original arrangement, permutations (2) to (5) are obtained by using the elements of the first column to interchange the elements of the second column, one at a time. Permutation (6) is obtained by interchanging the columns of the original arrangement of the experiment, making use of the two elements in the first column.

Examine a 2-sample experiment, where each sample has 3 variates, i.e. $\begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{pmatrix}$.

The expectation is to have $\frac{6!}{3!\,3!} = 20$ permutations, which are given in Table 2.

The process of permuting the variates reveal the following:

$\begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{pmatrix}$ original arrangement of the experiment 1 permutation

$x_{11} \leftarrow x_{2i}$, $i = 1, 2, 3$ 3 permutations
$x_{12} \leftarrow x_{2i}$, $i = 1, 2, 3$ 3 permutations
$x_{13} \leftarrow x_{2i}$, $i = 1, 2, 3$ 3 permutations

$\begin{pmatrix} x_{1s} \\ x_{1t} \end{pmatrix} \leftarrow \begin{pmatrix} x_{2i} \\ x_{2j} \end{pmatrix}$;
$s \neq t$, $i \neq j$ (3 x 3) 9 permutations

$\begin{pmatrix} x_{21} & x_{11} \\ x_{22} & x_{12} \\ x_{23} & x_{13} \end{pmatrix}$

exchange the samples (columns) 1 permutation

Again, observe that

$\binom{3}{0}\binom{3}{0} = 1 \text{ x } 1 = 1$ Permutation
(original arrangement of the experiment)

$\binom{3}{1}\binom{3}{1} = 3 \text{ x } 3 = 9$ Permutations
(using one variate from first sample)

$\binom{3}{2}\binom{3}{2} = 3 \text{ x } 3 = 9$ Permutations
(using two variates from first sample)

$\binom{3}{3}\binom{3}{3} = 1 \text{ x } 1 = 1$ Permutation
(exchange samples, i.e., three variates)

Total = 1 + 9 + 9 + 1 = 20

Similarly, observe that permutation (1) is the original matrix, permutations (2) to (10) are obtained by using the elements of the first column to interchange the elements of the second column, one at a time. Permutations (11) to (19) are obtained by using 2 elements of the first column to interchange the elements of the second column, and permutation (20) is obtained by interchanging the columns of the original arrangement of the experiment.

Continuing in the above fashion, clearly, the number of permutations for any 2-sample experiment can be written as

$$\sum_{i=0}^{n}\binom{n}{i}\binom{n}{i} = \sum_{i=0}^{n}\binom{n}{i}^2 = \binom{2n}{n}$$

for equal sample sizes. An adjustment for unequal sample sizes yields $\sum_{i=0}^{\min(n_1,n_2)}\binom{n_1}{i}\binom{n_2}{i}$ permutations, because $\binom{a}{b} = 0$ for $b > a$.

After obtaining all the distinct permutations from a complete enumeration, the statistic of interest is computed for each permutation. Each value of the statistic obtained from a complete enumeration occurs with probability $\left(\frac{N!}{n_1!\,n_2!}\right)^{-1}$ for sample sizes, n_1 and n_2, $N = n_1 + n_2$, this translates to $\frac{(n!)^2}{N!}$ for $n = n_1 = n_2$. The distribution of the statistic is

Table 2: Permutations of a 2 x 3 Experiment.

1	2	3	4	5
x_{11} x_{21}	x_{21} x_{11}	x_{22} x_{21}	x_{23} x_{21}	x_{11} x_{12}
x_{12} x_{22}	x_{12} x_{22}	x_{12} x_{11}	x_{12} x_{22}	x_{21} x_{22}
x_{13} x_{23}	x_{13} x_{23}	x_{13} x_{23}	x_{13} x_{11}	x_{13} x_{23}
6	7	8	9	10
x_{11} x_{21}	x_{11} x_{21}	x_{11} x_{13}	x_{11} x_{21}	x_{11} x_{21}
x_{22} x_{12}	x_{23} x_{22}	x_{12} x_{22}	x_{12} x_{13}	x_{12} x_{22}
x_{13} x_{23}	x_{13} x_{12}	x_{21} x_{23}	x_{22} x_{23}	x_{23} x_{13}
11	12	13	14	15
x_{21} x_{11}	x_{21} x_{11}	x_{22} x_{21}	x_{21} x_{11}	x_{21} x_{11}
x_{22} x_{12}	x_{23} x_{22}	x_{23} x_{11}	x_{12} x_{13}	x_{12} x_{22}
x_{13} x_{23}	x_{13} x_{12}	x_{13} x_{12}	x_{22} x_{23}	x_{23} x_{13}
16	17	18	19	20
x_{22} x_{21}	x_{11} x_{12}	x_{11} x_{12}	x_{11} x_{21}	x_{21} x_{11}
x_{12} x_{11}	x_{21} x_{13}	x_{21} x_{22}	x_{22} x_{12}	x_{22} x_{12}
x_{23} x_{13}	x_{22} x_{23}	x_{23} x_{13}	x_{23} x_{13}	x_{23} x_{13}

Numbers 1 – 20 on top of the permutations represent the permutation numbers.

thereafter obtained by simply tabulating the distinct values of the statistic against their probabilities of occurrence in the complete enumeration.

This method of obtaining unconditional exact permutation distribution also suffices when ranks of observations of an experiment are used instead of the actual observations. In handling ranks with this approach, tied observations do not pose any problems because the permutation process will be implemented as if the tied observations or ranks are distinct.

Given an n x p experiment,

$$X_N = \begin{pmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{pmatrix}, \quad N = np$$

with x_{ij} as actual observations, i = 1, 2, …, p, j = 1, 2, …, n for some rank order statistic, replace these observations with ranks. In order to achieve this, do a combined ranking from the smallest to the largest observation. For equal sample sizes, this yields an n x p matrix of ranks represented as follows:

$$R_N = \begin{pmatrix} R_1^{(1)} & \cdots & R_1^{(p)} \\ \vdots & \vdots & \vdots \\ R_n^{(1)} & \cdots & R_n^{(p)} \end{pmatrix},$$

N = np and $R_i^{(j)}$ is the ith rank for sample j, see Sen and Puri (1967) for an expository discussion of rank order statistics. At this stage, the method can now be applied to this matrix of ranks. Note that any rearrangement or permutation of this matrix of ranks can be used in generating all the other distinct permutations.

The first step in developing the algorithm is to formulate the matrix of ranks, by adopting the trivial permutation, because it does not matter what rearrangement of the actual matrix of ranks is used in initiating the process of permutation, that is,

$$\begin{pmatrix} 1 & n_1+1 \\ 2 & n_1+2 \\ 3 & n_1+3 \\ \vdots & \vdots \\ n_1 & n_1+n_2 \end{pmatrix}$$

and in the case of $n_1 = n_2 = n$, $\begin{pmatrix} 1 & n+1 \\ 2 & n+2 \\ 3 & n+3 \\ \vdots & \vdots \\ n & 2n \end{pmatrix}$.

For the above matrix of ranks, ensure that ties are taken care of, by replacing ranks of tied observations with the mean of their ranks.

In designing the computer algorithm for the method of complete enumeration via permutation described so far, it is intended that all statements should be read like sentences or as a sequence of commands. We write Set T ← 1, where Set is part of the statement language and T is a variable. Words that form the statement language required for this work include: do, od, else, for, if, fi, set, then, through, to, as used in Goodman and Hedetniemi (1977). To distinguish variable names from words in the statement language, variable names appear in full capital letters.

As a way of illustration, in formulating the computer algorithm for unconditional exact permutation distribution, a consideration is given to rank order statistic. The computer algorithm for the generation of the "trivial" matrix of ranks is presented in the next session for equal sample sizes.

Results

Algorithm (RANK) Generation of the trivial matrix of ranks

Step 1. Set P ← number of treatments;
K ← Number of variates
Step 2. For I ← 1 to P do through Step 4
Step 3. For J ← 1 to K do through Step 4
Step 4. [X is the matrix of ranks]
Set X(J, I) ← (I – 1)K + J od

For all possible permutations of the N samples of p subsets of size n, the model of the number of permutations required for the computer algorithm for an experiment of two samples is:

$$\sum_{i=0}^{n} \binom{n}{i}\binom{n}{i} \quad \text{permutations}$$

where n is the number of variates in each sample (column) i.e., the balanced case. The computer algorithm now follows.

Algorithm (PERMUTATION)

Step 1 For J1 ← 1 to K do through Step 5
Step 2 Set TEMP ← X(J1, P - 1), I1 ← P
Step 3 For J2 ← 1 to K do Step 5
Step 4 Set X(J1, P - 1) ← X(J2, I1), X(J2, I1) ← TEMP
Step 5 [Compute statistic and restore original values of X] od
Step 6 For I ← 1 to K – 1 do through Step 16
Step 7 Set TEMP1 ← X(I, P - 1)
Step 8 For J ← I + 1 to K do through Step 16
Step 9 Set TEMP2 ← X(J, P - 1)
Step 10 For L ← P to P do through Step 16
Step 11 For I1 ← 1 to K do through Step 16
Step 12 For L1 ← L to P do through Step 16
Step 13 If L ← L1 then Set T ← I1 + 1
else Set T ← 1 fi
Step 14 For J1 ← T to K do Step 16
Step 15 Set X(I, P - 1)←X(I1, L), X(I1, L)←TEMP1,
X(J, P - 1) ← X(J1, L1), X(J1, L1) ← TEMP2
Step 16 [Compute statistic and restore original values of X] od
Step 17 For I ← 1 to K – 2 do through Step 32
Step 18 Set TEMP1 ← X(I, P - 1)
Step 19 For J ← I + 1 to K – 1 do through Step 32
Step 20 Set TEMP2 ← X(J, P - 1)
Step 21 For M ← J + 1 to K do through Step 32
Step 22 Set TEMP3 ← X(M, P - 1)
Step 23 For L ← P to P do through Step 32
Step 24 For I1 ← 1 to K do through Step 32
Step 25 For L1 ← L to P do through Step 32
Step 26 If L ← L1 then Set T ← I1 + 1
else Set T ← 1 fi
Step 27 For J1 ← T to K do through Step 32
Step 28 For L2 ← L1 to P do through Step 32
Step 29 If L1 ← L2 then Set T1 ← J1 + 1
else Set T1 ← 1 fi
Step 30 For J2 ← T1 to K do Step 32
Step 31 Set X(I, P - 1)←X(I1, L), X(I1, L) ← TEMP1,
X(J, P - 1) ← X(J1, L1), X(J1, L1) ← TEMP2,
X(M, P - 1) ← X(J2, L2), X(J2, L2) ← TEMP3
Step 32 [Compute statistic and restore original values of X] od
Step 33 For I ← 1 to K – 3 do through Step 53
Step 34 Set TEMP1 ← X(I, P - 1)
Step 35 For J ← I + 1 to K – 2 do through Step 53
Step 36 Set TEMP2 ← X(J, P - 1)
Step 37 For M ← J + 1 to K – 1 do through Step 53
Step 38 Set TEMP3 ← X(M, P - 1)
Step 39 For N ← M + 1 to K do through Step 53
Step 40 Set TEMP4 ← X(N, P - 1)
Step 41 For L ← P to P do through Step 53
Step 42 For I1 ← 1 to K do through Step 53

Step 43 For L1 ← L to P do through Step 53
Step 44 If L ← L1 then Set T ← I1 + 1
else Set T ← 1 fi
Step 45 For J1 ← T to K do through Step 53
Step 46 For L2 ← L1 to P do through Step 53
Step 47 If L1 ← L2 then Set T1 ← J1 + 1
else Set T1 ← 1 fi
Step 48 For J2 ← T1 to K do through Step 53
Step 49 For L3 ← L2 to P do through Step 53
Step 50 If L2 ← L3 then Set T2 ← J2 + 1
else Set T2 ← 1 fi
Step 51 For J3 ← T2 to K do Step 53
Step 52 Set X(I, P - 1)← X(I1, L), X(I1, L)←TEMP1,
X(J, P - 1) ← X(J1, L1), X(J1, L1) ← TEMP2,
X(M, P - 1) ← X(J2, L2), X(J2, L2) ← TEMP3,
X(N, P - 1) ← X(J3, L3), X(J3, L3) ← TEMP4 od
Step 53 [Compute statistic and restore original values of X] od
Step 54 [Interchange samples and compute statistic]

The PERMUTATION algorithm was translated to FORTRAN codes and implemented in Intel Visual FORTRAN for a 2 x 5 experiment. The 252 distinct permutations generated are presented in the Appendix. The algorithm can be extended to any sample size, depending on the processor speed and memory space of the computer being used to implement the algorithm. For an optimal management of computer memory (space complexity), the permutations are not stored, they are discarded immediately the statistic of interest is computed.

By way of illustration, generate the p-values for a 2 x 5 experiment for the Mood test. Fahoome (2002) noted that when $\alpha = 0.05$, sample size should exceed 5 for the large sample approximation to be adopted for the Mood test. The unconditional permutation approach makes it possible to obtain exact p-values even for fairly large sample sizes. Given two samples, $y_{11}, y_{12}, \ldots, y_{1n}$ and $y_{21}, y_{22}, \ldots, y_{2n}$, the test statistic for the Mood test is

$$M = \sum_{i=1}^{n}\left(R_{1i} - \frac{2n+1}{2}\right)^2$$

for equal sample sizes.
R_{1i} is the rank of y_{1i}, i = 1, 2, …, n obtained after carrying out a combined ranking for the two samples. The large sample approximation for equal samples is

$$z = \frac{M - \frac{n\left(N^2-1\right)}{12}}{\sqrt{\frac{n^2(N+1)\left(N^2-4\right)}{180}}},$$

where N = 2n and M the Mood test statistic.

The p-values obtained are presented in Table 3 and the distribution of the test statistic is represented graphically in Figure 1.

Table 3. p-values for Mood Statistic.

M	p(M)	p-value	M	p(M)	p-value
11.25	0.0079	0.0079	43.25	0.0397	0.6032
15.25	0.0079	0.0159	45.25	0.0476	0.6508
17.25	0.0159	0.0317	47.25	0.0714	0.7222
21.25	0.0317	0.0635	49.25	0.0397	0.7619
23.25	0.0317	0.0952	51.25	0.0397	0.8016
25.25	0.0159	0.1111	53.25	0.0476	0.8492
27.25	0.0397	0.1508	55.25	0.0397	0.8889
29.25	0.0476	0.1984	57.25	0.0159	0.9048
31.25	0.0397	0.2381	59.25	0.0317	0.9365
33.25	0.0397	0.2778	61.25	0.0317	0.9682
35.25	0.0714	0.3492	65.25	0.0159	0.9841
37.25	0.0476	0.3968	67.25	0.0079	0.9921
39.25	0.0397	0.4365	71.25	0.0079	1.0000
41.25	0.1270	0.5635			

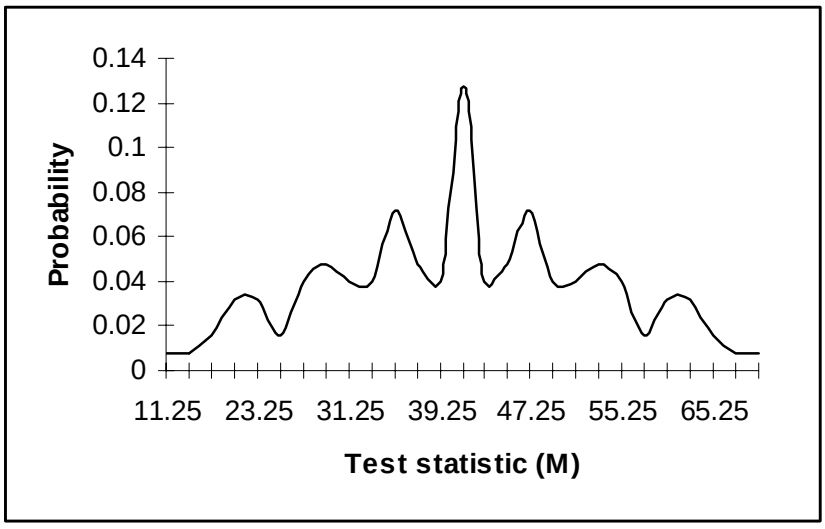


Figure 1. Exact distribution of Mood test statistic for a 2 x 5 experiment.

Clearly, results obtained from using Normal distribution, which is the large sample asymptotic distribution for the Mood test, will certainly not be exactly the same as using the exact permutation distribution, especially for small sample sizes. The permutation approach produces the exact p-values.

Example

Consider the following example on page 278 of Freund (1979) on difference of means.

Table 2: Heat-producing capacity of coal in millions of calories per tonne

Mine1	Mine2
8400	7510
8230	7690
8380	7720
7860	8070
7930	7660

Subjecting the data in Table 2 to Mood test, the test statistic (M) is 39.25 and from Table 3 containing unconditional exact permutation distribution of Mood test statistic, the corresponding p-value is 0.4365 which exceeds $\alpha = 0.05$, suggesting that we cannot reject the null hypothesis of no difference between the heat-producing capacity of coal from the two mines. Adopting the large sample Normal approximation for Mood test, z calculated is –0.17 which gives a p-value of 0.4325 and this exceeds $\alpha/2 = 0.025$, meaning that the observed data are compatible with the null hypothesis of no difference as earlier obtained from the exact permutation test.

Conclusion

Several authors have attempted to obtain exact p-values for different statistics using the permutation approach. Two things have made their attempts an uphill task. First is the speed of computer required to perform a permutation test. Until recently, the speed of available computers has been grossly inadequate to handle complete enumeration for even small sample sizes. Recent advances in computer design has drawn researchers in this area closer to the realization of complete enumeration even for fairly large sample sizes. Secondly, the intensive looping in computer programming required for complete enumeration for unconditional exact permutation test demands a good programming skill.

In this article, a straight forward but computer intensive approach has been adopted in creating an algorithm that can carryout a systematic enumeration of distinct permutations of a 2-sample experiment. With this algorithm, the p-values for statistics involving two samples can be accurately generated, thereby ensuring that the probability of making a type I error is exactly α

References

Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science, 7,* 131-177.

Bayarri, M. J. & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19,* 58-80.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7,* 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. NY: Chapman and Hall.

Fahoome, G. (2002). Twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods, 1,* 248-268.

Fisher, R. A. (1935). *Design of experiments*. Oliver and Boyd, Edinburgh.

Freund, J. E. (1979). *Modern elementary statistics* (5th edition). Englewood Cliff, N. J.: Prentice-Hall.

Good, P. (2000). *Permutation tests: a practical guide to resampling methods for testing hypotheses* (2nd edition). NY: Springer Verlag.

Goodman, S. E., & Hedetniemi, S. T. (1977). *Introduction to the design and analysis of algorithms.* London: McGraw-Hill Book Company.

Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in high dimensional settings. *Biometrika, 89,* 359-374.

Hall, P., & Weissman, I. (1997). On the estimation of extreme tail probabilities. *The Annals of Statistics, 25,* 1311-1326.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician, 52*, 127-132.

Opdyke, J. D. (2003). Fast permutation tests that maximize power under conventional Monte Carlo sampling for pairwise and multiple comparisons. *Journal of Modern Applied Statistical Methods, 2*, 27-49.

Pesarin, F. (2001). *Multivariate permutation tests.* NY: Wiley.

Sen, P. K., & Puri, M. L. (1967). On the theory of rank order tests for location in the multivariate one sample problem. *The Annals of Mathematical Statistics, 38*, 1216-1228.

Senchaudhuri, P., Mehta, C. R., & Patel, N. T. (1995). Estimating exact p-values by the method of control variates or Monte Carlo rescue. *Journal of the American Statistical Association, 90*, 640-648.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd edition). NY: McGraw-Hill.

Spiegelhalter, D. J. (2004). Incorporating Bayesian ideas into health-care evaluation. *Statistical Science, 19*, 156-174

Appendix: Permutations of a 2 x 5 Experiment.

1	2	3	4	5	6	7	8	9	10	11	12
1 6	6 1	7 6	8 6	9 6	10 6	1 2	1 6	1 6	1 6	1 6	1 3
2 7	2 7	2 1	2 7	2 7	2 7	6 7	7 2	8 7	9 7	10 7	2 7
3 8	3 8	3 8	3 1	3 8	3 8	3 8	3 8	3 2	3 8	3 8	6 8
4 9	4 9	4 9	4 9	4 1	4 9	4 9	4 9	4 9	4 2	4 9	4 9
5 10	5 10	5 10	5 10	5 10	5 1	5 10	5 10	5 10	5 10	5 2	5 10
13	14	15	16	17	18	19	20	21	22	23	24
1 6	1 6	1 6	1 6	1 4	1 6	1 6	1 6	1 6	1 5	1 6	1 6
2 3	2 7	2 7	2 7	2 7	2 4	2 7	2 7	2 7	2 7	2 5	2 7
7 8	8 3	9 8	10 8	3 8	3 8	3 4	3 8	3 8	3 8	3 8	3 5
4 9	4 9	4 3	4 9	6 9	7 9	8 9	9 4	10 9	4 9	4 9	4 9
5 10	5 10	5 10	5 3	5 10	5 10	5 10	5 10	5 4	6 10	7 10	8 10
25	26	27	28	29	30	31	32	33	34	35	36
1 6	1 6	6 1	6 1	6 1	6 1	7 6	7 6	7 6	8 6	8 6	9 6
2 7	2 7	7 2	8 7	9 7	10 7	8 1	9 1	10 1	9 7	10 7	10 7
3 8	3 8	3 8	3 2	3 8	3 8	3 2	3 8	3 8	3 1	3 1	3 8
4 5	4 9	4 9	4 9	4 2	4 9	4 9	4 2	4 9	4 2	4 9	4 1
9 10	10 5	5 10	5 10	5 10	5 2	5 10	5 10	5 2	5 10	5 2	5 2
37	38	39	40	41	42	43	44	45	46	47	48
6 1	6 1	6 1	6 1	7 6	7 6	7 6	8 6	8 6	9 6	6 1	6 1
2 3	2 7	2 7	2 7	2 1	2 1	2 1	2 7	2 7	2 7	2 4	2 7
7 8	8 3	9 8	10 8	8 3	9 8	10 8	9 1	10 1	10 8	3 8	3 4
4 9	4 9	4 3	4 9	4 9	4 3	4 9	4 3	4 9	4 1	7 9	8 9
5 10	5 10	5 10	5 3	5 10	5 10	5 3	5 10	5 3	5 3	5 10	5 10

Appendix Continued:

49	50	51	52	53	54	55	56	57	58	59	60
6 1	6 1	7 6	7 6	7 6	8 6	8 6	9 6	6 1	6 1	6 1	6 1
2 7	2 7	2 1	2 1	2 1	2 7	2 7	2 7	2 5	2 7	2 7	2 7
3 8	3 8	3 4	3 8	3 8	3 1	3 1	3 8	3 8	3 5	3 8	3 8
9 4	10 9	8 9	9 4	10 9	9 4	10 9	10 1	4 9	4 9	4 5	4 9
5 10	5 4	5 10	5 10	5 4	5 10	5 4	5 4	7 10	8 10	9 10	10 5
61	62	63	64	65	66	67	68	69	70	71	72
7 6	7 6	7 6	8 6	8 6	9 6	1 2	1 2	1 2	1 2	1 6	1 6
2 1	2 1	2 1	2 7	2 7	2 7	6 3	6 7	6 7	6 7	7 2	7 2
3 5	3 8	3 8	3 1	3 1	3 8	7 8	8 3	9 8	10 8	8 3	9 8
4 9	4 5	4 9	4 5	4 9	4 1	4 9	4 9	4 3	4 9	4 9	4 3
8 10	9 10	10 5	9 10	10 5	10 5	5 10	5 10	5 10	5 3	5 10	5 10
73	74	75	76	77	78	79	80	81	82	83	84
1 6	1 6	1 6	1 6	1 2	1 2	1 2	1 2	1 6	1 6	1 6	1 6
7 2	8 7	8 7	9 7	6 4	6 7	6 7	6 7	7 2	7 2	7 2	8 7
10 8	9 2	10 2	10 8	3 8	3 4	3 8	3 8	3 4	3 8	3 8	3 2
4 9	4 3	4 9	4 2	7 9	8 9	9 4	10 9	8 9	9 4	10 9	9 4
5 3	5 10	5 3	5 3	5 10	5 10	5 10	5 4	5 10	5 10	5 4	5 10
85	86	87	88	89	90	91	92	93	94	95	96
1 6	1 6	1 2	1 2	1 2	1 2	1 6	1 6	1 6	1 6	1 6	1 6
8 7	9 7	6 5	6 7	6 7	6 7	7 2	7 2	7 2	8 7	8 7	9 7
3 2	3 8	3 8	3 5	3 8	3 8	3 5	3 8	3 8	3 2	3 2	3 8
10 9	10 2	4 9	4 9	4 5	4 9	4 9	4 5	4 9	4 5	4 9	4 2
5 4	5 4	7 10	8 10	9 10	10 5	8 10	9 10	10 5	9 10	10 5	10 5
97	98	99	100	101	102	103	104	105	106	107	108
1 3	1 3	1 3	1 3	1 6	1 6	1 6	1 6	1 6	1 6	1 3	1 3
2 4	2 7	2 7	2 7	2 3	2 3	2 3	2 7	2 7	2 7	2 5	2 7
6 8	6 4	6 8	6 8	7 4	7 8	7 8	8 3	8 3	9 8	6 8	6 5
7 9	8 9	9 4	10 9	8 9	9 4	10 9	9 4	10 9	10 3	4 9	4 9
5 10	5 10	5 10	5 4	5 10	5 10	5 4	5 10	5 4	5 4	7 10	8 10
109	110	111	112	113	114	115	116	117	118	119	120
1 3	1 3	1 6	1 6	1 6	1 6	1 6	1 6	1 4	1 4	1 4	1 4
2 7	2 7	2 3	2 3	2 3	2 7	2 7	2 7	2 5	2 7	2 7	2 7
6 8	6 8	7 5	7 8	7 8	8 3	8 3	9 8	3 8	3 5	3 8	3 8
4 5	4 9	4 9	4 5	4 9	4 5	4 9	4 3	6 9	6 9	6 5	6 9
9 10	10 5	8 10	9 10	10 5	9 10	10 5	10 5	7 10	8 10	9 10	10 5
121	122	123	124	125	126	127	128	129	130	131	132
1 6	1 6	1 6	1 6	1 6	1 6	6 1	6 1	6 1	6 1	6 1	6 1
2 4	2 4	2 4	2 7	2 7	2 7	7 2	7 2	7 2	8 7	8 7	9 7
3 5	3 8	3 8	3 4	3 4	3 8	8 3	9 8	10 8	9 2	10 2	10 8
7 9	7 5	7 9	8 5	8 9	9 4	4 9	4 3	4 9	4 3	4 9	4 2
8 10	9 10	10 5	9 10	10 5	10 5	5 10	5 10	5 3	5 10	5 3	5 3

Appendix Continued:

133	134	135	136	137	138	139	140	141	142	143	144
7 6	7 6	7 6	8 6	6 1	6 1	6 1	6 1	6 1	6 1	7 6	7 6
8 1	8 1	9 1	9 7	7 2	7 2	7 2	8 7	8 7	9 7	8 1	8 1
9 2	10 2	10 8	10 1	3 4	3 8	3 8	3 2	3 2	3 8	3 2	3 2
4 3	4 9	4 2	4 2	8 9	9 4	10 9	9 4	10 9	10 2	9 4	10 9
5 10	5 3	5 3	5 3	5 10	5 10	5 4	5 10	5 4	5 4	5 10	5 4
145	146	147	148	149	150	151	152	153	154	155	156
7 6	8 6	6 1	6 1	6 1	6 1	6 1	6 1	7 6	7 6	7 6	8 6
9 1	9 7	7 2	7 2	7 2	8 7	8 7	9 7	8 1	8 1	9 1	9 7
3 8	3 1	3 5	3 8	3 8	3 2	3 2	3 8	3 2	3 2	3 8	3 1
10 2	10 2	4 9	4 5	4 9	4 5	4 9	4 2	4 5	4 9	4 2	4 2
5 4	5 4	8 10	9 10	10 5	9 10	10 5	10 5	9 10	10 5	10 5	10 5
157	158	159	160	161	162	163	164	165	166	167	168
6 1	6 1	6 1	6 1	6 1	6 1	7 6	7 6	7 6	8 6	6 1	6 1
2 3	2 3	2 3	2 7	2 7	2 7	2 1	2 1	2 1	2 7	2 3	2 3
7 4	7 8	7 8	8 3	8 3	9 8	8 3	8 3	9 8	9 1	7 5	7 8
8 9	9 4	10 9	9 4	10 9	10 3	9 4	10 9	10 3	10 3	4 9	4 5
5 10	5 10	5 4	5 10	5 4	5 4	5 10	5 4	5 4	5 4	8 10	9 10
169	170	171	172	173	174	175	176	177	178	179	180
6 1	6 1	6 1	6 1	7 6	7 6	7 6	8 6	6 1	6 1	6 1	6 1
2 3	2 7	2 7	2 7	2 1	2 1	2 1	2 7	2 4	2 4	2 4	2 7
7 8	8 3	8 3	9 8	8 3	8 3	9 8	9 1	3 5	3 8	3 8	3 4
4 9	4 5	4 9	4 3	4 5	4 9	4 3	4 3	7 9	7 5	7 9	8 5
10 5	9 10	10 5	10 5	9 10	10 5	10 5	10 5	8 10	9 10	10 5	9 10
181	182	183	184	185	186	187	188	189	190	191	192
6 1	6 1	7 6	7 6	7 6	8 6	1 2	1 2	1 2	1 2	1 2	1 2
2 7	2 7	2 1	2 1	2 1	2 7	6 3	6 3	6 3	6 7	6 7	6 7
3 4	3 8	3 4	3 4	3 8	3 1	7 4	7 8	7 8	8 3	8 3	9 8
8 9	9 4	8 5	8 9	9 4	9 4	8 9	9 4	10 9	9 4	10 9	10 3
10 5	10 5	9 10	10 5	10 5	10 5	5 10	5 10	5 4	5 10	5 4	5 4
193	194	195	196	197	198	199	200	201	202	203	204
1 6	1 6	1 6	1 6	1 2	1 2	1 2	1 2	1 2	1 2	1 6	1 6
7 2	7 2	7 2	8 7	6 3	6 3	6 3	6 7	6 7	6 7	7 2	7 2
8 3	8 3	9 8	9 2	7 5	7 8	7 8	8 3	8 3	9 8	8 3	8 3
9 4	10 9	10 3	10 3	4 9	4 5	4 9	4 5	4 9	4 3	4 5	4 9
5 10	5 4	5 4	5 4	8 10	9 10	10 5	9 10	10 5	10 5	9 10	10 5
205	206	207	208	209	210	211	212	213	214	215	216
1 6	1 6	1 2	1 2	1 2	1 2	1 2	1 2	1 6	1 6	1 6	1 6
7 2	8 7	6 4	6 4	6 4	6 7	6 7	6 7	7 2	7 2	7 2	8 7
9 8	9 2	3 5	3 8	3 8	3 4	3 4	3 8	3 4	3 4	3 8	3 2
4 3	4 3	7 9	7 5	7 9	8 5	8 9	9 4	8 5	8 9	9 4	9 4
10 5	10 5	8 10	9 10	10 5	9 10	10 5	10 5	9 10	10 5	10 5	10 5

Appendix Continued:

217	218	219	220	221	222	223	224	225	226	227	228
1 3	1 3	1 3	1 3	1 3	1 3	1 6	1 6	1 6	1 6	6 1	6 1
2 4	2 4	2 4	2 7	2 7	2 7	2 3	2 3	2 3	2 7	7 2	7 2
6 5	6 8	6 8	6 4	6 4	6 8	7 4	7 4	7 8	8 3	8 3	8 3
7 9	7 5	7 9	8 5	8 9	9 4	8 5	8 9	9 4	9 4	9 4	10 9
8 10	9 10	10 5	9 10	10 5	10 5	9 10	10 5	10 5	10 5	5 10	5 4
229	230	231	232	233	234	235	236	237	238	239	240
6 1	6 1	7 6	6 1	6 1	6 1	6 1	7 6	6 1	6 1	6 1	6 1
7 2	8 7	8 1	7 2	7 2	7 2	8 7	8 1	7 2	7 2	7 2	8 7
9 8	9 2	9 2	8 3	8 3	9 8	9 2	9 2	3 4	3 4	3 8	3 2
10 3	10 3	10 3	4 5	4 9	4 3	4 3	4 3	8 5	8 9	9 4	9 4
5 4	5 4	5 4	9 10	10 5	10 5	10 5	10 5	9 10	10 5	10 5	10 5
241	242	243	244	245	246	247	248	249	250	251	252
7 6	6 1	6 1	6 1	6 1	7 6	1 2	1 2	1 2	1 2	1 6	6 1
8 1	2 3	2 3	2 3	2 7	2 1	6 3	6 3	6 3	6 7	7 2	7 2
3 2	7 4	7 4	7 8	8 3	8 3	7 4	7 4	7 8	8 3	8 3	8 3
9 4	8 5	8 9	9 4	9 4	9 4	8 5	8 9	9 4	9 4	9 4	9 4
10 5	9 10	10 5	10 5	10 5	10 5	9 10	10 5	10 5	10 5	10 5	10 5

Numbers 1 – 252 on top of the permutations represent the permutation numbers

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 333-342

1538 – 9472/05/$95.00

JMASM19: A SPSS Matrix For Determining Effect Sizes From Three Categories: r And Functions Of r, Differences Between Proportions, And Standardized Differences Between Means

David A. Walker
Educational Research and Assessment Department
Northern Illinois University

The program is intended to provide editors, manuscript reviewers, students, and researchers with an SPSS matrix to determine an array of effect sizes not reported or the correctness of those reported, such as r-related indices, r-related squared indices, and measures of association, when the only data provided in the manuscript or article are the n, M, and SD (and sometimes proportions and *t* and *F* (1) values) for two-group designs. This program can create an internal matrix table to assist researchers in determining the size of an effect for commonly utilized r-related, mean difference, and difference in proportions indices when engaging in correlational and/or meta-analytic analyses.

Key words: SPSS, syntax, effect size

Introduction

Cohen (1988) defined effect size as "the *degree* to which the phenomenon is present in the population" (p. 9) or "the degree to which the null hypothesis is false" (p. 10). For many years, researchers, editorial boards, and professional organizations have called for the reporting of effect sizes with statistical significance testing (Cohen, 1965; Knapp, 1998; Levin, 1993; McLean & Ernest, 1998; Thompson, 1994; Wilkinson & The APA Task Force on Statistical Inference, 1999). However, research applied to this issue has indicated that most published studies do not supply measures of effect size with results garnered from statistical significance testing (Craig, Eison, & Metze, 1976; Henson & Smith, 2000; Vacha-Hasse, Nilsson, Reetz, Lance, & Thompson, 2000). When reported with statistically significant results, effect size can provide information pertaining to the extent of the difference between the null hypothesis and the alternative hypothesis. Furthermore, effect sizes can show the magnitude of a relationship and the proportion of the total variance of an outcome that is accounted for (Cohen, 1988; Kirk, 1996; Shaver, 1985).

Conversely, there have long been cautions affiliated with the use of effect sizes. For instance, over 20 years ago, Kraemer and Andrews (1982) pointed out that effect sizes have limitations in the sense that they can be a

> measure that clearly indicates clinical significance only in the case of normally distributed control measures and under conditions in which the treatment effect is additive and uncorrelated with pretreatment or control treatment responses. (p. 407)

Hedges (1981) examined the influence of measurement error and invalidity on effect sizes and found that both of these problems tended to underestimate the standardized mean difference effect size. In addition, Prentice and Miller (1992) ascertained that, "The statistical size of an effect is heavily dependent on the operationalization of the independent variables

David Walker is an Assistant Professor at Northern Illinois University. His research interests include structural equation modeling, effect sizes, factor analyses, predictive discriminant analysis, predictive validity, weighting, and bootstrapping. Email: dawalker@niu.edu.

and the choice of a dependent variable" (p. 160). Robinson, Whittaker, Williams, and Beretvas (2003) warned that "depending on the choice of which effect size is reported, in some cases important conclusions may be obscured rather than revealed" (p. 52). Finally, Kraemer (1983), Sawilowsky (2003), and Onwuegbuzie and Levin (2003) cautioned that effect sizes are vulnerable to various primary assumptions. Onwuegbuzie and Levin cited nine limitations affiliated with effect sizes and noted generally that these measures:

> are sensitive to a number of factors, such as: the research objective; sampling design (including the levels of the independent variable, choice of treatment alternatives, and statistical analysis employed); sample size and variability; type and range of the measure used; and score reliability. (p. 135)

Effect sizes fall into three categories: 1) product moment correlation (*r*) and functions of *r*; 2) differences between proportions; and 3) standardized differences between means (Rosenthal, 1991). The first category of effect size, the *r*-related indices, can be considered as based on the correlation between treatment and result (Levin, 1994). For this group, "Effect size is generally reported as some proportion of the total variance accounted for by a given effect" (Stewart, 2000, p. 687), or, as Cohen (1988) delineated this effect size, "Another possible useful way to understand *r* is as a proportion of common elements between variables" (p. 78). Cohen (1988) suggested that for *r*-related indices, values of .10, .30, and .50 should serve as indicators of small, medium, and large effect sizes, while for *r*-related squared indices, values of .01, .09, and .25 should serve as indicators of small, medium, and large, respectively.

The differences between proportions group is constituted in measures, for example, such as the differences between independent population proportions (i.e., Cohen's h) or the difference between a population proportion and .50 (i.e., Cohen's g) (Cohen, 1988). Finally, the standardized differences between means encompasses measures of effect size in terms of mean difference and standardized mean difference such as Cohen's d and Glass' delta. Cohen (1988) defined the values of effect sizes for both the differences between proportions and the standardized differences between means as small = .20, medium = .50, and large = .80. It should be mentioned, however, that it is at the discretion of the researcher to note the context in which small, medium, and large effects are being defined when using any effect size index. As was first discussed by Glass, McGaw, and Smith (1981), and reiterated by Cohen (1988), about these effect size target values and their importance:

> these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of magnitude. (p. 532)

The purpose of this article is to provide editors, manuscript reviewers, students, and researchers with an SPSS (Statistical Package for the Social Sciences) program to determine an array of effect sizes not reported or the correctness of those reported, such as *r*-related indices, *r*-related squared indices, and measures of association, when the only data provided in the manuscript or article are n, M, and SD (and sometimes proportions and *t* and F(1) values) for between-group designs.

Another intention is that this software will be used as an educational resource for students and researchers. That is, the user can run quickly this program and determine the size of the effect. It is not the purpose of this research to serve as an effect size primer and, thus, discuss in-depth the various indices' usage, limitations, and importance. Rather, this program can assist users who have the minimal, proper statistics present to enter into the matrix to derive an effect size index of interest.

In meta-analytic research, it is often difficult to convert study outcomes, via formulae that are accessible over a vast array of the scholarly literature, into a common metric. Thus,

yet another purpose of this program is to offer researchers software that contains many of the formulae used in meta-analyses.

Methodology

The presented SPSS program will create an internal matrix table to assist researchers and students in determining the size of an effect for commonly utilized *r*-related, mean difference, and difference in proportions indices when engaging in correlational and/or meta-analytic analyses. Currently, the program produces nearly 50 effect sizes (see appendix A for truncated results of the program's ability).

This software program employs mostly data from published articles, and some simulated data, to demonstrate its uses in terms of effect size calculations. Most of the formulae incorporated into this program come from Aaron, Kromrey, and Ferron (1998), Agresti and Finlay (1997), Cohen (1988), Cohen and Cohen (1983), Cooper and Hedges (1994), Hays (1963; 1981), Hedges (1981), Hedges and Olkin (1985), Kelley (1935), Kraemer (1983), Kraemer and Andrews (1982), McGraw and Wong (1992), Olejnik and Algina (2000), Peters and Van Voorhis (1940), Richardson (1996), Rosenthal (1991), and Rosenthal, Rosnow, and Rubin (2000).

It should be noted that with the *r*-related and the standardized differences between means effect sizes, there are numerous, algebraically-related methods concerning how to calculate these indices, of which some of been provided, but not all since the same value(s) would be repeated numerous times (see Cooper & Hedges, 1994 or Richardson, 1996 for the various formulae).

Because this matrix is meant for between-group designs, k = 2, there are some specific assumptions that should be addressed. To run the program, it is assumed that the user has access to either n, M, and, SD or *t* or F(1) values from two-group comparisons. Also, this program was intended for post-test group comparison designs and not, for example, a one-group repeated measures design, which can be found in meta-analytic data sets as well.

Certain effect sizes produced by the program that the user does not wish to view, or that may be nonsensical pertaining to the research of study, should be disregarded. As well, a few of the measures developed for very specific research conditions, such as the Common Language effect size, may not be pertinent to many research situations and should be ignored if this is the case. The Mahalanobis Generalized Distance (D^2) is an estimated effect size with p = .5 implemented as the proportion value in the formula. Some of the *r*-related squared indices may contain small values that are negative. This can occur when the MS (treatment) is < the MS (residual) (Peters & Van Voorhis, 1940), or when the t or F values used in the formulae to derive these effect size indices are < 1.00 (Hays, 1963). Finally, even with exact formulas, some of the computed values may be slightly inexact, as could the direction of a value depending on the user's definition of the experimental and control groups.

Program Description and Output

As presented in the program output found in appendix A, the reader should note that they enter the M, SD, and n for both groups in the first lines of the syntax termed 'test'. If they want to run just one set of data, they put it next to test 1. If more than one set of data are desired, they put the subsequent information in test 2 to however many tests they want to conduct.

The matrix produced will group the effect sizes by the three categories noted previously and also related to an appropriate level of measurement. In parenthesis, after an effect size is displayed in the matrix, is a general explanation of that particular measure and any notes that should be mentioned such as used when there are ESS (equal sample sizes) or PEES (populations are of essentially equal size), yields a PRE (proportional reduction in error) interpretation, or examines the number of CP (concordant pairs) and DP (discordant pairs).

Further, the matrix generates power values, based on calculations of alpha set at the .05 level, related to indices such as Cohen's d, Glass' delta, and Hedges' g. Finally, because some of the standardized differences between means indices produce biased values under various conditions; numerous measures of effect for this group are provided for the user to obtain the proper measure(s) pertaining to specific

circumstances within the research context. The accuracy of the program was checked by an independent source whose hand calculations verified the formulas utilized throughout the program via various situations employing two-group n, M, SD. Appendix B provides the full syntax for this program. To obtain an SPSS copy of the syntax, send an e-mail to the author.

Reference

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed). Upper Saddle River, NJ: Prentice Hall.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2. *Bulletin of the Psychonomic Society, 7*, 280-282.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart & Winston.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. *Journal of Research and Development in Education, 33*, 285-296.

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences, 21*, 554-559.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. *Research in Schools, 5*, 39-41.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8*, 93-101.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91*(2), 404-412.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *The Journal of Experimental Education, 61*, 378-382.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review, 6*, 231-243.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 3* (2), 361-365.

McLean, J. E, & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in Schools, 5*, 15-23.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286.

Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods, 2*(1), 133-151.

Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases.* New York: McGraw-Hill.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160-164.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28*(1), 12-22.

Robinson, D. H., Whittaker, T. A., Williams, N. J., & Beretvas, S. N. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *The Journal of Experimental Education, 72*(1), 51-64.

Rosenthal, R. (1991). (Series Ed.), *Meta-analytic procedures for social research.* Newbury Park, CA: Sage Publications.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes inbehavioral research: A correlational approach*. Cambridge, England: Cambridge University Press.

Sawilowsky, S. S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods, 2*, 467-474

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan, 67*, 57-60.

Stewart, D. W. (2000). Testing statistical significance testing: Some observations of an agnostic. *Educational and Psychological Measurement, 60*, 685-690.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Vacha-Hasse, T., Nilsson, J, E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*, 413-425.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Appendix A: A Sample of the Program Output.

Descriptive Statistics

Test	M1	SD1	n1	M2	SD2	n2
1	9.160	3.450	31	5.350	3.090	31
2	15.950	3.470	20	13.050	3.270	20
3	31.150	10.830	27	30.370	9.410	27
4	105.000	15.000	24	95.000	15.000	24

Appendix A: Continued

Standardized Differences Between Means, % of Nonoverlap (with d), and Power

Glass Delta (Used When There are Unequal Variances and Calculated with the Control Group SD)	Cohens d (Using M & SD Pooled)	Cohens d (Using t Value n1=n2)	Hedges g (Used When There are Small Sample Sizes)	Hedges g (Using t Value n1=n2)	Hedges g (Using Cohens d)	U % of Nonoverlap	Power
1.2330	1.1634	1.1826	1.1488	1.1634	1.1445	61.0362	.9945
.8869	.8602	.8825	.8431	.8602	.8384	49.9468	.7552
.0829	.0769	.0784	.0758	.0769	.0754	5.9506	.0589
.6667	.6667	.6810	.6557	.6667	.6526	41.4105	.6183

Proportion of Variance-Accounted-For Effect Sizes: 2x2 Dichotomous/Nominal

Phi (The Mean Percent Difference Between Two Variables with Either Considered Causing the Other)	Tetrachoric Correlation (Estimation of Pearsons r for Continuous Variables Reduced to Dichotomies)	Pearsons Coefficient of Contingency (C) (A Nominal Approximation of the Pearsonian correlation r)	Sakodas Adjusted Pearsons C (Association Between Two Variables as a Percentage of Their Maximum Possible Variation)
.4492	.4492	.4098	.5795
.3674	.3674	.3449	.4878
.0384	.0384	.0384	.0542
.3015	.3015	.2887	.4082

Proportion of Variance-Accounted-For Effect Sizes: Measures of Relationship(PEES)

Point Biserial r (Pearsons r for Dichotomous and Continuous Variables)	Biserial r (r for Interval and Dichotomous Variables)	Pearsons r (Using Cohens d with Equal n)	Pearsons r (Using Cohens d with Unequal n)	Pearsons r (If no t Value and for Equal n; Corrected for Bias in Formula)	Pearsons r (Using t Value and for Equal n; Corrected for Bias)	Pearsons r (Using Hedges g with Unequal n)
.5028	.6300	.5028	.5028	.5090	.5090	.5042
.3951	.4950	.3951	.3951	.4037	.4037	.3970
.0384	.0481	.0384	.0384	.0391	.0391	.0386
.3162	.3962	.3162	.3162	.3223	.3223	.3176

Appendix A: Continued

Proportion of Variance-Accounted-For Effect Sizes: Univariate Analyses (k=2, ESS)

R Square (d Value)	R Square (If no t Value and for Unequal n Corrected for Bias in Formula)	R Square (Using t Value and for Unequal n Corrected for Bias)	Adjusted R Square (d Value)	Adjusted R Square (Using t Value and for Unequal n)
.2528	.2591	.2591	.2275	.2339
.1561	.1630	.1630	.1105	.1177
.0015	.0015	.0015	-.0377	-.0376
.1000	.1039	.1039	.0600	.0641

Proportion of Variance-Accounted-For Effect Sizes: Univariate Analyses (k=2, ESS)

Eta Square (Squared Correlation Ratio or the Percentage of Variation Effects Uncorrected for a Sample)	Eta Square (Calculated with F Value)	Omega Square (Corrected Estimates for the Population Effect)	Estimated Omega Square	Epsilon Square (Percentage of Variation Effects Uncorrected for a Sample	Epsilon Square (Calculated with F Value)
.2528	.2591	.2528	.2437	.2404	.2467
.1561	.1630	.1561	.1379	.1339	.1409
.0015	.0015	.0015	-.0173	-.0177	-.0177
.1000	.1039	.1000	.0828	.0804	.0844

Appendix B: Program Syntax

```
* Data enter *.
data list list /testno(f8.0) exprmean exprsd(2f9.3) exprn(f8.0)contmean contsd(2
f9.3) contn(f8.0).
* Put the M, SD, n for the Experimental Group followed by the Control Group.
Begin data

1      9.16   3.45   31           5.35   3.09   31
2      15.95  3.47   20           13.05  3.27   20
3      31.15  10.83  27           30.37  9.41   27
4      105    15     24           95     15     24
end data.
*************************************************************************
Example References
1   Example of t and Cohen's d          JEE (2002), 70(4),356-357
2   Example of F, Cohen's d, and Eta2          JEE (2002), 70(3),235
3   Example of t and Eta2               JEE (2002), 70(4),305-306
4   Example of d, r, r2, and CL         Psych Bulletin (1992), 111(2),363
*************************************************************************.
```

Appendix B: Continued

```
compute poold = ((exprn-1)*(exprsd**2)+(contn-1)*(contsd**2))/((exprn+contn)-2) .
compute glassdel = (exprmean-contmean)/contsd.
compute cohend = (exprmean-contmean)/sqrt(poold).
compute clz = (exprmean-contmean)/sqrt(exprsd**2 + contsd**2).
compute cl = CDFNORM(clz)*100.
compute akf1 = (exprn+contn)**2.
compute akf2 = 2*(exprn+contn).
compute akf3 = akf1-akf2.
compute akf4 = (akf3)/(exprn*contn).
compute r2akf = (cohend**2)/(cohend**2+akf4).
compute rakf = SQRT (r2akf).
compute hedgesg = cohend*(1-(3/(4*(exprn+contn)-9))).
compute ub = CDF.NORMAL((ABS(cohend)/2),0,1).
compute U = (2*ub-1)/ub*100.
compute critical = 0.05.
compute h = (2*exprn*contn)/(exprn+contn).
compute ncp = ABS((cohend*SQRT(h))/SQRT(2)).
compute alpha = IDF.T(1-critical/2,exprn+contn-2).
compute power1 = 1-NCDF.T(alpha,exprn+contn-2,NCP).
compute power2 = 1-NCDF.T(alpha,exprn+contn-2,-NCP).
compute B = power1 + power2.
compute f2 = cohend ** 2 / 4 .
compute f = ABS(cohend/2).
compute eta2 = (f2) / (1 + f2) .
compute eta = SQRT(eta2).
compute epsilon2 = 1-(1-eta2) * (exprn  +  contn-1) / (exprn  +  contn-2).
compute ttest = cohend  * SQRT((exprn  *  contn) /( exprn  +  contn)).
compute cohenda = 2*ttest/SQRT(exprn  +  contn-2).
compute hedgesa = 2*ttest/SQRT(exprn  +  contn).
compute hedgesb = cohend*SQRT((exprn  +  contn-2)/(exprn + contn)).
compute hedgesn = (exprn  +  contn)/(2).
compute hedgesnh = 1/(.5*((1/exprn) + (1/contn))).
compute hedgesnn = sqrt(hedgesn/hedgesnh).
compute r1= ttest/SQRT((ttest**2)+ exprn  +  contn-2).
compute r = cohend/SQRT(cohend ** 2 + 4) .
compute rd = cohend/SQRT((cohend ** 2 + 4*(hedgesnn))).
compute rg = hedgesg/SQRT((hedgesg ** 2 + 4*(hedgesnn)*((exprn  +  contn-2)/(exprn + contn)))).
compute phi = (r **2/(1+r **2)) **.5.
compute phi2 = phi **2.
compute taub = SQRT(phi **2).
compute gktau = phi **2.
compute zr = .5 * LN((1 + r) / (1 - r)) .
compute zrbias = r/(2*(exprn + contn-1)).
compute zrcor = zr - zrbias.
compute rsquare = r **2 .
compute rsquare1 = r1**2.
compute adjr2 = rsquare - ((1-rsquare)*(2/(exprn  +  contn -3)))  .
compute adjr2a = rsquare1 - ((1-rsquare1)*(2/(exprn  +  contn -3)))  .
compute adjr2akf = r2akf - ((1-r2akf)*(2/(exprn  +  contn -3)))  .
compute k = SQRT(1-r **2).
compute k2 = k **2.
compute lambda = 1-rsquare.
compute rpbs = SQRT(eta2).
compute rbs = rpbs*1.253.
compute rpbs2 = rpbs **2.
compute ftest = ttest **2.
compute omega2 = ftest / ((exprn  +  contn) + ftest).
compute estomega = (ttest**2-1)/(ttest**2 + exprn + contn -1).
```

Appendix B: Continued

```
compute eta2f = (ftest)/(ftest + exprn + contn -2).
compute esticc = (ftest-1)/(ftest + exprn + contn -2).
compute c = SQRT(chi/ (exprn + contn+chi)).
compute adjc = c/SQRT(.5).
compute cramer = SQRT(chi/ (exprn + contn*1)).
compute cramer2 = cramer **2.
compute t = SQRT(chi/ (exprn + contn*1)).
compute t2 = cramer **2.
compute d2 = r **2/(r **2+1).
compute w = SQRT (c **2/(1-c **2)).
compute w2 = w **2.
compute percenta = exprmean/(exprmean+contmean).
compute percentb = exprsd/(exprsd+contsd).
compute percentd = percenta-percentb.
compute p = (exprmean*contsd)-(exprsd*contmean).
compute q = (exprmean*contsd)+(exprsd*contmean).
compute yulesq = p/q.
compute taua = ((p-q)/((exprn+contn)*(exprn + contn-1)/2)).
compute rr = (exprmean/(exprmean+contmean))/(exprsd/(exprsd+contsd)).
compute rrr = 1-rr.
compute odds = (exprmean/contmean)/(exprsd/contsd).
compute tauc = 4*((p-q)/((exprn+contn)*(exprn+contn))).
compute zb = SQRT(chi).
compute coheng = exprsd - .50.
compute cohenh = 2 * ARSIN(SQRT(.651)) - 2 * ARSIN(SQRT(.414)).
compute cohenq = .55-zr.
execute.

* FINAL REPORTS *.
FORMAT poold to cohenq (f9.4).
VARIABLE LABELS testno 'Test'/ exprmean 'M1'/ exprsd 'SD1'/ exprn 'n1'/contmean 'M2'/ contsd 'SD2'/contn 'n2'
/glassdel 'Glass Delta'/ cohend 'Cohens d (Using M & SD)'/ U 'U % of Nonoverlap'/ B 'Power'/ hedgesg 'Hedges g'
/cohenda 'Cohens d (Using t Value n1=n2)'/hedgesa 'Hedges g (Using t Value n1=n2)'/hedgesb 'Hedges g (Using Cohens
d)'/rd 'Pearsons r (Using Cohens d with Unequal n)'/ rg 'Pearsons r (Using Hedges g with Unequal n)'/ f2 'f Square (Proportion
of Variance Accounted for by Difference in Population Membership)' /r2akf 'R Square (If no t Value and for Unequal n
Corrected for Bias in Formula)'/eta2 'Eta Square (Squared Correlation Ratio or the Percentage of Variation Effects
Uncorrected for a Sample)' /epsilon2 'Epsilon Square (Percentage of Variation Effects Uncorrected for a Sample' / omega2
'Omega Square (Corrected Estimates for the Population Effect)' /r 'Pearsons r (Using Cohens d with Equal n)' /r1 'Pearsons r
(Using t Value and for Equal n; Corrected for Bias)' /rakf 'Pearsons r (If no t Value and for Equal n; Corrected for Bias in
Formula)' /phi 'Phi (The Mean Percent Difference Between Two Variables with Either Considered Causing the Other)' /phi2
'Phi Coefficient Square (Proportion of Variance Shared by Two Dichotomies)' /zr 'Fishers Z (r is Transformed to be Distributed
More Normally)'/w2 'w Square (Proportion of Variance Shared by Two Dichotomies)' /coheng 'Cohens g (Difference Between a
Proportion and .50)' /cohenh 'Cohens h (Differences Between Proportions)' /cohenq 'Cohens q (One Case & Theoretical Value
of r)' /rsquare 'R Square (d Value)' /rsquare1 'R Square (Using t Value and for Unequal n Corrected for Bias)'/adjr2 'Adjusted R
Square (d Value)'/adjr2a 'Adjusted R Square (Using t Value and for Unequal n)'/adjr2akf 'Adjusted R Square (Unequal n and
Corrected for Bias)'/ lambda 'Wilks Lambda (Small Values Imply Strong Association)' / t2 'T Square (Measure of Average
Effect within an Association)' /d2 'D2 Mahalanobis Generalized Distance (Estimated with p = .5 as the Proportion of Combined
Populations)' /rpbs 'Point Biserial r (Pearsons r for Dichotomous and Continuous Variables)' /rbs 'Biserial r (r for Interval and
Dichotomous Variables)'/rpbs2 'r2 Point-Biserial (Proportion of Variance Accounted for by Classifying on a Dichotomous
Variable Special Case Related to R2 and Eta2)' / f 'f (Non-negative and Non-directional and Related to d as an SD of
Standardized Means when k=2 and n=n)' /k2 'k2 (r2/k2: Ratio of Signal to Noise Squared Indices)' / k 'Coefficient of Alienation
(Degree of Non-Correlation: Together r/k are the Ratio of Signal to Noise)' /c 'Pearsons Coefficient of Contingency (C) (A
Nominal Approximation of the Pearsonian correlation r)' /adjc 'Sakodas Adjusted Pearsons C (Association Between Two
Variables as a Percentage of Their Maximum Possible Variation)' /cramer 'Cramers V (Association Between Two Variables as
a Percentage of Their Maximum Possible Variation)'/odds 'Odds Ratio (The Chance of Faultering after Treatment or the Ratio
of the Odds of Suffering Some Fate)'/ rrr 'Relative Risk Reduction (Amount that the Treatment Reduces Risk)'/ rr 'Relative Risk
Coefficient (The Treatment Groups Amount of the Risk of the Control Group)'/ percentd 'Percent Difference'/ yulesq 'Yules Q
(The Proportion of Concordances to the Total Number of Relations)'/ t 'Tshuprows T (Similar to Cramers V)' /w 'w (Amount of
Departure from No Association)' /chi 'Chi Square(1)(Found from Known Proportions)' /eta 'Correlation Ratio (Eta or the Degree
of Association Between 2 Variables)'/eta2f 'Eta Square (Calculated with F Value)'/epsilonf 'Epsilon Square (Calculated with F
Value)'/esticc 'Estimated Population Intraclass Correlation Coefficient'/estomega 'Estimated Omega Square'/zrcor 'Fishers Z
```

Appendix B: Continued

```
Corrected for Bias (When n is Small)'/cl 'Common Language (Out of 100 Randomly Sampled Subjects (RSS) from Group 1 will
have Score > RSS from Group 2)'/ taua 'Kendalls Tau a (The Proportion of the Number of CP and DP Compared to the Total
Number of Pairs)'/ tetra 'Tetrachoric Correlation (Estimation of Pearsons r for Continuous Variables Reduced to
Dichotomies)'/taub 'Kendalls Tau b (PRE Interpretations)'/ gktau 'Goodman Kruskal Tau (Amount of Error in Predicting an
Outcome Utilizing Data from a Second Variable)'/cramer2 'Cramers V Square'/ tauc 'Kendalls Tau c (AKA Stuarts Tau c or a
Variant of Tau b for Larger Tables)'/.

REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=testno exprmean exprsd exprn contmean contsd contn
  /TITLE "Descriptive Statistics".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=glassdel cohend cohenda hedgesg hedgesa hedgesb U B
  /TITLE "Standardized Differences Between Means, % of Nonoverlap (with d), and Power".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= percentd yulesq
/TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 Dichotomous Associations".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= rr rrr odds
/TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 Dichotomous Associations".
REPORT FORMAT=LIST AUTOMATIC ALIGN (LEFT)
MARGINS (*,90)
  /VARIABLES= chi phi tetra c adjc
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 Dichotomous/Nominal".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= cramer w t
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 Dichotomous/Nominal".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= taub tauc taua
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 Ordinal Associations".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=gktau
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes: 2x2 PRE Measures".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= phi2 cramer2 w2 t2
  /TITLE"Proportion of Variance-Accounted-For Effect Sizes: Squared Associations".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=coheng cohenh cohenq
  /TITLE "Differences Between Proportions".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= f zr zrcor eta esticc
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes:Measures of Relationship(PEES)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= rpbs rbs r rd rakf r1 rg
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes:Measures of Relationship(PEES)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= k cl
  /TITLE "Proportion of Variance-Accounted-For Effect Sizes:Measures of Relationship(PEES)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=rsquare r2akf rsquare1 adjr2 adjr2a
  /TITLE"Proportion of Variance-Accounted-For Effect Sizes:Univariate Analyses (k=2, ESS)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=eta2 eta2f omega2 estomega epsilon2 epsilonf
  /TITLE"Proportion of Variance-Accounted-For Effect Sizes:Univariate Analyses (k=2, ESS)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= rpbs2 k2
  /TITLE"Proportion of Variance-Accounted-For Effect Sizes:Univariate Analyses (k=2, ESS)".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES=f2 lambda d2
  /TITLE"Proportion of Variance-Accounted-For Effect Sizes:Multivariate Analyses(k=2,ESS)".
```

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 343-351

1538 – 9472/05/$95.00

A Comparison of Nonlinear Regression Codes

Paul Fredrick Mondragon
United States Navy
China Lake, California

Brian Borchers
Department of Mathematics
New Mexico Tech

Five readily available software packages were tested on nonlinear regression test problems from the NIST Statistical Reference Datasets. None of the packages was consistently able to obtain solutions accurate to at least three digits. However, two of the packages were somewhat more reliable than the others.

Key words: nonlinear regression, Levenberg – Marquardt, NIST StRD

Introduction

The goal of this study is to compare the nonlinear regression capabilities of several software packages using the nonlinear regression datasets available from the National Institute of Standards and Technology (NIST) Statistical Reference Datasets (National Institute of Standards and Technology [NIST], 2000).

The nonlinear regression problems were solved by the NIST using quadruple precision (128 bits) and two public domain programs with different algorithms and different implementations; the convergence criterion was residual sum of squares (RSS) and the tolerance was 1E-36. Certified values were obtained by rounding the final solutions to 11 significant digits. Each of the two public domain programs, using only double precision, could achieve 10 digits of accuracy for every problem. (McCullough, 1998).

Paul Mondragon is an Operations Research Analyst. Contact him at Paul.Mondragon@navy.mil. Brian Borchers is Professor of Mathematics. His research interests are in interior point methods for linear and semi-definite programming, with applications to combinatorial optimization problems. Contact him at borchers@nmt.edu.

The software packages considered in this study are:

1. MATLAB codes by Hans Bruun Nielsen (2002).
2. GaussFit (Jeffreys, Fitzpatrick, McArthur, & McCartney, 1998).
3. Gnuplot (Crawford, 1998).
4. Microsoft Excel (Mathews & Seymour, (1994).
5. Minpack (More, Garbow, & Hillstrom, 1980).

Hiebert (1981) compared 12 Fortran codes on 36 separate nonlinear least squares problems. Twenty-eight of the problems used by Hiebert are given by Dennis, Gay, and Welch (1977) with the other eight problems given by More, Garbow, and Hillstrom, (1978). In their paper, More et al. (1978) used Fortran subroutines to test 35 problems. These 35 problems were a mixture of systems of nonlinear equations, nonlinear least – squares, and unconstrained minimization. We are not aware of any other published studies in which codes were tested on the NIST nonlinear regression problems.

Methodology

Following McCullough (1998), accuracy is determined using the log relative error (LRE) formula,

$$\lambda_q = -\log_{10}\left[\frac{|q-c|}{|c|}\right]. \quad (1)$$

where q is the value of the parameter estimated by the code being tested and c is the certified value. In the event that $q = c$ exactly then λ_q is not formally defined, but we set it equal to the number of digits in c. It is also possible for an LRE to exceed the number of digits in c; for example, it is possible to calculate an LRE of 11.4 even though c contains only 11 digits. This is because double precision floating point arithmetic uses binary, not decimal arithmetic. In such a case, λ_q is set equal to the number of digits in c. Finally, any λ_q less than one is set to zero.

Robustness is an important characteristic for a software package. In terms of accuracy, there is concern with each specific problem as individuals. Robustness, however, is a measure of how the software packages performed on the problems as a set. In other words, there must be a sense of how reliable the software package is so there may be some level of confidence that it will solve a particular nonlinear regression problem other than those listed in the NIST StRD.

In this sense, robustness may very well be more important to the user than accuracy. Certainly the user would want parameter estimates to be accurate to some level, but accuracy to 11 digits is often not particularly useful in practical application. However, the user would want to be confident that the software package they are using will generate parameter estimates accurate to perhaps 3 or 4 digits on most any problem they attempt to solve. If, on the other hand, a software package is extremely accurate on some problems, but returns a solution which is not close to actual values on other problems, the user would want to use this software package with extreme caution.

The codes were not compared on the basis of CPU time, for the reason that all of these codes solve (or fail to solve) all of the NIST test problems within a few seconds. CPU time comparisons would certainly be of interest in the context of problems with many variables, or in problems for which the model and derivative computations are extremely time consuming.

A closer look at the various software packages chosen for this comparative study follows. Some of the packages are parts of a larger package, such as Microsoft Excel. In this case, the parts of the larger package which were used in the completion of this study are considered. Others in the set of packages used are designed exclusively for solving nonlinear least – squares problems.

HBN MATLAB Code

The first software package used in this study is the MATLAB code written by Hans Bruun Nielson (2002). Nielson's code can work with a user supplied analytical Jacobian or it can compute the Jacobian by finite differences. The Jacobian was calculated analytically for the purpose of this study.

GaussFit

GaussFit (Jeffreys et al., 1998) was designed for astrometric data reduction with data from the NASA Hubble Space Telescope. It was designed to be a flexible least squares package so that astrometric models could quickly and easily be written, tested and modified. In this study, version 3.53 of GaussFit was used.

A unique feature of GaussFit is that although it is a special purpose system designed for estimation problems, it includes a full-featured programming language which has all the power of traditional languages such as C, Pascal, and Fortran. This language possesses a complete set of looping and conditional statements as well as a modern nested statement structure. Variables and arrays may be freely created and used by the programmer. There is therefore no theoretical limit to the complexity of model that can be expressed in the GaussFit programming language.

One of the onerous tasks that faces the implementer of a least squares problem is the calculation of the partial derivatives with respect to the parameters and observations that are required in order to form the equations of

condition and the constraint equations. GaussFit solves this problem automatically using a built-in algebraic manipulator to calculate all of the required partial derivatives. Every expression that the user's model computes will carry all of the required derivative information along with it. No numerical approximations are used.

Gnuplot

Gnuplot (Crawford, 1998) is a command-driven interactive function plotting program capable of a variety of tasks. Included among these tasks are plotting both two- or three-dimensional functions in a variety of formats, computations in integer, floating point, and complex arithmetic, and support for a variety of operating systems.

The 'fit' command can fit a user-defined function to a set of data points (x,y) or (x,y,z), using an implementation of the nonlinear least-squares Marquardt – Levenberg algorithm. Any user-defined variable occurring in the function body may serve as a fit parameter, but the return type of the function must be real.

For this study, gnuplot version 3.7 patchlevel 3 was used. Initially, gnuplot displayed only approximately 6 digits in its solutions to the estimation of the parameters. The source code was modified to display 20 digits in its solutions. For the purposes of this study, FIT_LIMIT was set to 1.0e-15, with the default values for the other program parameters.

Microsoft Excel

Microsoft Excel is a multi-purpose software package. As only a small part of its capabilities were used during the process of this study, discussion of Excel is limited to its 'Solver' capabilities. The Excel Solver function is a self-contained function in that all of the data must be located somewhere on the spreadsheet. The Solver allows the user to find a solution to a function that contains up to 200 variables and up to 100 constraints on those variables. A Quasi-Newton search direction was used with automatic scaling and a tolerance of 1.0e-15. (Mathews & Seymour, 1994).

MINPACK

Minpack (More et al., 1980) is a library of Fortran codes for solving systems of nonlinear equations and nonlinear least squares problems. Minpack is freely distributed via the Netlib web site and other sources. The algorithms proceed either from an analytic specification of the Jacobian matrix or directly from the problem functions. The paths include facilities for systems of equations with a banded Jacobian matrix, for least squares problems with a large amount of data, and for checking the consistency of the Jacobian matrix with the functions.

For the problems involved in this study a program and a subroutine had to be written. The main program calls the lmder1 routine. The lmder1 routine calls two user written subroutines which compute function values and partial derivatives.

Results

The problems given in the NIST StRD dataset are provided with two separate initial starting positions for the estimated parameters. The first position, Start 1, is considered to be the more difficult because the initial values for the parameters are farther from the certified values than are the initial values given by Start 2. For this reason, one might expect that the solutions generated from Start 2 to be more accurate, or perhaps for the algorithm to take fewer iterations. It is interesting to note that in several cases the results from Start 2 are not more accurate based upon the minimum LRE recorded.

The critical parameter used in the comparison of these software packages is the LRE as calculated in (1). The number of estimated parameters for these problems range from two to nine. It was decided that it would be beneficial for the results table to be as concise as possible, yet remain useful. As a result, after running a particular package from both starting values, the LRE for each estimated parameter was calculated. The minimum LRE for the estimated parameters from each starting position was then entered into the results table.

Table 1. Minimum Log Relative Error of Estimated Parameters.

Problem	Start	Excel	Gnuplot	GaussFit	HBN	Minpack
Misra1a	1	4.8	5.8	10.0	11.0	7.7
	2	6.1	5.8	10.0	10.3	7.7
Chwirut2	1	4.2	4.9	7.4	10.6	2.4
	2	4.6	4.9	8.6	9.1	2.4
Chwirut1	1	4.0	4.2	8.0	10.3	7.5
	2	4.9	4.3	8.5	10.1	7.5
Lanczos3	1	0.0	3.9	0.0	4.9	3.3
	2	0.0	3.9	7.9	5.1	3.3
Gauss1	1	4.7	5.1	8.7	6.9	8.0
	2	4.6	5.1	8.6	6.9	3.3
Gauss2	1	4.5	4.9	0.0	6.8	7.8
	2	4.4	4.9	0.0	6.8	7.2
DanWood	1	4.6	5.1	NS	10.2	6.6
	2	4.7	5.1	NS	8.7	6.6
Misra1b	1	4.4	5.8	0.0	10.9	2.7
	2	6.4	5.8	9.7	11.0	2.5
Kirby2	1	1.0	4.8	7.4	10.3	6.2
	2	1.9	4.9	7.9	10.4	6.2

Problem	Start	Excel	Gnuplot	GaussFit	HBN	Minpack
Hahn1	1	0.0	4.0	0.0	9.5	NS
	2	0.0	4.0	0.0	9.7	NS
Nelson	1	0.0	0.0	0.0	0.0	0.0
	2	0.0	0.0	1.4	0.0	0.0
MGH17	1	0.0	NS	NS	0.0	7.6
	2	1.4	3.7	NS	0.0	7.5
Lanczos1	1	0.0	10.0	0.0	4.9	4.3
	2	0.0	10.0	10.0	5.8	4.3
Lanczos2	1	0.0	5.4	0.0	5.7	3.5
	2	0.0	5.4	9.1	5.3	3.5
Gauss3	1	4.3	4.8	9.2	6.5	2.4
	2	4.1	5.0	9.1	6.5	2.4
Misra1c	1	0.0	5.9	0.0	10.8	7.6
	2	0.0	5.9	10.0	10.2	7.6
Misra1d	1	5.2	5.8	0.0	11.0	7.6
	2	4.4	5.9	8.9	11.0	7.6
Roszman1	1	3.5	4.1	8.7	4.0	0.0
	2	0.0	5.1	8.6	4.0	0.0
ENSO	1	0.0	1.6	3.7	6.5	0.0
	2	0.0	2.2	3.7	6.6	0.0

Problem	Start	Excel	Gnuplot	Gaussfit	HBN	Minpack
MGH09	1	0.0	3.6	0.0	5.0	6.3
	2	5.0	3.6	0.0	5.2	6.4
Thurber	1	1.7	3.2	0.0	7.8	0.0
	2	1.5	4.4	6.4	7.5	0.0
BoxBOD	1	0.0	4.5	NS	9.7	0.0
	2	5.6	3.8	NS	8.6	9.1
Rat42	1	5.3	4.2	8.0	10.3	7.1
	2	5.2	4.1	8.3	11.2	7.1
MGH10	1	0.0	NS	0.0	0.0	10.8
	2	0.0	4.4	0.0	0.0	11.0
Eckerle4	1	0.0	0.0	0.0	8.1	0.0
	2	5.1	4.8	8.3	7.2	1.2
Rat43	1	0.0	NS	NS	0.0	6.9
	2	3.2	2.6	NS	1.3	7.0
Bennett5	1	0.0	6.4	NS	3.7	0.0
	2	0.0	6.7	NS	3.7	1.5

Notes: NS – Software package was unable to generate any numerical solution. A score of 0.0 implies that the package returned a solution in which at least one parameter was accurate to less than one digit.

An entry of 0.0 in the results table is given if a software package generated estimates for the parameters but the minimum LRE was less than 1.0. For example if the minimum LRE was calculated to be 8.0e-1, rather than entering this, a 0.0 was entered. This practice was followed in an effort to be consistent with established practices (McCullough, 1998). If a software package did not generate a numerical estimate for the parameters, then an entry of 'NS' is entered into the results table.

Accuracy

As stated in the introduction, the accuracy of the solutions was evaluated in terms of the log relative error (LRE) using equation (1). Essentially the LRE gives the number of leading digits in the estimated parameter values that correspond to the leading digits of the certified values. Again, it should to be noted that the values given in the results table are the minimum LRE values for those problems. In other words, if a problem has five parameters to be estimated and four of the parameters are estimated accurately to seven digits, but the fifth is only accurate to one digit, it is reasonable to say that the problem was not accurately solved. On the other hand, if all five parameters were estimated to at least five digits, then one could feel confident that the package had indeed solved the problem.

Nielsen's MATLAB code had an average LRE score of 6.8 for the problems. For the problems this package was able to solve, the starting position did not seem to be of much importance. In fact, it is quite interesting that for several problems the LRE generated using the first set of initial values is larger than the LRE generated using the second set of initial values. This is interesting because the second set of initial values is closer to the certified values of the parameter estimates. Of the twenty-three problems that the parameters were estimated correctly to at least two digits, the average LRE was 7.96. This shows us that the accuracy of the estimated parameters was very high on those problems which this package effectively solved.

GaussFit had an average LRE score of 4.9. Unlike Nielsen's MATLAB code, GaussFit was very dependent upon the initial values given to the parameters. On eight of the problems GaussFit was unable to estimate all of the parameters to even one digit from the first starting position. From the second starting position GaussFit was able to estimate all of the parameters to over six digits correctly. This seemingly high dependence upon the starting values is a potential problem when using GaussFit for solving these nonlinear regression problems. There is no guarantee that one can find a starting value which is sufficiently close to the solution for GaussFit to effectively solve the problem.

Gnuplot has an average LRE score of 4.6. While this is actually lower than the average LRE score for GaussFit, gnuplot is not so heavily dependent upon the starting position in order to solve the problem. Rather, much like Nielsen's code, gnuplot seems quite capable of accurately estimating the parameter values to four digits whether the starting position is close or far from the certified values.

Microsoft Excel did not solve these problems well at all. The average LRE score for Excel is 2.32. Excel did perform reasonably well on the problems with a lower level of difficulty. For the eight problems with a lower level of difficulty the average LRE was 4.18. While these are probably reasonable results for these problems, we can see that for the problems with a moderate or high level of difficulty Excel did very poorly. Such results as this would cause one to have serious questions as to Excel being able to solve any particular least squares regression problem.

The Minpack library of Fortran codes also performed poorly on these particular problems. The average LRE for the twenty-six problems that Minpack did solve is 4.51. Minpack was significantly less accurate than the other packages on four of the problems, Misra1b, ENSO, Thurber, and Eckerle4. On the other hand, Minpack was considerably more accurate on the MGH10 problem. Minpack did not seem to be overly dependent upon starting position as in only two of the problems was there a significant difference in the minimum LRE for the different starting positions.

Table 2. Comparison of Robustness

Package	N	P(%)
Gnuplot	24	88.89%
Nielsen's MATLAB Code	23	85.19%
GaussFit	17	62.96%
Minpack	17	62.96%
Excel	15	55.56%

Robustness

Although the accuracy to which a particular software package is able to estimate the parameters is an important characteristic of the package, the ability of the package to solve a variety of nonlinear regression problems to an acceptable level of accuracy is perhaps more important to the user. Most users would like to have confidence that the particular software package in use is likely to estimate those parameters to an acceptable level of accuracy.

What is an acceptable level of accuracy? Such a question as this might elicit a variety of responses simply depending upon the nature of the study, the data, the relative size of the parameters, and many other variables which may need to be considered. For the purposes of this study we will consider an acceptable level of accuracy to be three digits. In Table 2, the various software packages are compared by the number (and percentage) of the problems which they were able to estimate the parameters accurately to at least three digits from either starting position.

Here, N is the number of problems which the package accurately estimated the parameters to at least three digits. P is the percentage of the problems which the package accurately estimated the parameters to at least three digits.

It can easily be seen here that as far as the robustness of the packages is concerned there are two distinct divisions. Nielsen's MATLAB code, and Gnuplot were both able to attain the 3 digit level of accuracy for over 80% of the problems. GaussFit, Excel, and Minpack, on the other hand were able to attain that level of accuracy on less than 65% of the problems.

Conclusion

The robustness of the codes tested in this study is surprisingly poor. In many cases, the results were quite accurate from one starting point, and completely incorrect from another starting point. In some cases the codes failed with an error message indicating that no correct solution had been obtained, while in other cases an incorrect solution was returned without warning.

Although some problems seemed to be easy for all of the codes from all of the starting points, there were other problems for which some codes easily solved the problem while other codes failed. In general, when reasonably accurate solutions were obtained, the solutions were typically accurate to five digits or better.

It is suggested that users of these and other packages for nonlinear regression would be well advised to carefully check the results that they obtain. Some obvious strategies for checking the solution include running a code from several different starting points and solving the problem with more than one package.

References

Crawford, D. (1998). Gnuplot Manual. Retrieved March 4, 2004, from www.ucc.ie/gnuplot/gnuplot.html.

Dennis, J. E., Gay, D. M., & Welch, R. E. (1997). An adaptive nonlinear least – squares algorithm, NBER working paper 196, M.I.T./C.C.R.E.M.S., Cambridge, Mass.

Hiebert, K. L. (1981). An Evaluation of Mathematical Software That Solves Nonlinear Least Squares Problems, *ACM Transactions on Mathematical Software*, *7*(1), 1-16.

Jefferys, W. H., Fitzpatrick, M. J., McArthur, B. E., & McCartney, J. E. (1998). *GaussFit: A system for least squares and robust estimation users manual*. Austin: University of Texas.

Kitchen, A. M., Drachenberg, R., & Symanzik, J. (2003). Assessing the reliability of web-based statistical software, *Computational Statistics*, *18*(1), 107-122.

Mathews, M., & Seymour S. (1994). *Excel for Windows: The Complete Reference*. (2nd ed.). NY:McGraw-Hill Inc.

McCullough, B. D. (1998). *Assessing reliability of statistical software: Part I: The American Statistician*, *52*(4), 358-366.

More, J. J., Garbow, B. S., & Hillstrom, K. E. (1978). Testing unconstrained optimization software. Rep. TM-324, Applied Math Division, Argonne National Lab., Argonne, IL.,

More, J. J., Garbow, B. S., & Hillstrom, K. E. (1980). User guide for MINPACK-1. Report ANL-80-74. Argonne, IL: Argonne National Lab.

National Institute of Standards and Technology. (2000). Statistical Reference Datasets (StRD). Retrieved March 4, 2004 http://www.itl.nist.gov/div898/strd/.

Nielsen, H. B. (2002). Nonlinear Least Squares Problems. Retrieved March 4, 2004 from www.imm.dtu.dk/~hbn/Software/#LSQ.

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 352

1538 – 9472/05/$95.00

Letter To the Editor

Abelson's Paradox And The Michelson-Morley Experiment

Sawilowsky, S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods, 2*(2), 467- 474.

Email correspondence was submitted to the Editorial Board pertaining to Sawilowsky's (2003) counter to the 'Einstein Gambit' in interpreting the 1887 Michelson-Morley experiment. To review, Carver (1978) claimed that hypothesis testing is detrimental to science, and educational research would be "better off even if [hypothesis tests are] properly used" (p. 398). Carver imagined (1993) that Albert Einstein would have been set back many years if he had relied on hypothesis tests. See Sawilowsky (2003) on why this gambit should be declined.

Carver (1993) obtained an effect size (eta squared) of .005 on some aspect of the Michelson-Morley data, although there was insufficient information given to replicate his results. Carver (1993) concluded "if Michelson and Morley had been forced … to do a test of statistical significance, they could have minimized its influence by reporting this effect size measure indicating that less that 1% of the variance in the speed of light was associated with its direction" (p. 289).

Sawilowsky (2003) noted that the experimental results were between 5 – 7.5 km/s. Although this did not support the static model of luminiferous *ether* that Michelson and Morley were searching for, which required 30 k/s, at more than 16,750 miles per hour it does represent a speed that exceeds the Earth's satellite orbital velocity. Thus, there is no legitimate reason to minimize this experimental result, which is clearly not zero, by dubbing it with the moniker of the most famous experiment in physics with a null result.

The author of the email correspondence noted that the magnitude of the speed is impressive, but perhaps Sawilowsky (2003) invoked a Huffian (Huff, 1954) maneuver in changing from the magnitude of variance explained to the speed in km/s. Although an invitation was declined to formalize the comment into a *Letter to the Editor*, the concern does merit a response.

Abelson (1985) sought to determine the contribution of past performance in explaining successful outcomes in the sport of professional baseball. There is no theory of success in baseball that denigrates the importance of the batting average. Yet, in Abelson's study, the amount of variance in successful outcomes that was due to batting average was a mere .00317.

Cohen (1988) emphasized "this is not a misprint – it is not .317, or even .0317. It is .00317, not quite one third of 1%" (p. 535)! Although a model that explains so little variance is probably misspecified, the response to the email query is to invoke Cohen's (1988) adage: "The next time you read that 'only X% of the variance is accounted for,' remember Abelson's Paradox" (p. 535).

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129-133.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.) Hillsdale, NY: Lawrence Erlbaum.

Huff, D. (1954). *How to lie with statistics*. NY: W.W. Norton & Company.

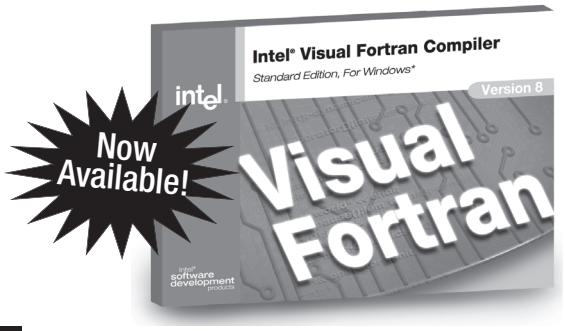
Sawilowsky, S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods*, *2*(2), 467- 474.

Shlomo S. Sawiowsky, Wayne State University.
Email: shlomo@wayne.edu

Two Years in the Making...

Intel® Visual Fortran 8.0

The next generation of Visual Fortran is here! Intel Visual Fortran 8.0 was developed jointly by Intel and the former DEC/Compaq Fortran engineering team.



Visual Fortran Timeline

1997 DEC releases Digital Visual Fortran 5.0

1998 Compaq acquires DEC and releases DVF 6.0

1999 Compaq ships CVF 6.1

2001 Compaq ships CVF 6.6

2001 Intel acquires CVF engineering team

2003 Intel releases Intel Visual Fortran 8.0

Intel Visual Fortran 8.0

- CVF front-end + Intel back-end
- Better performance
- OpenMP Support
- Real*16

Performance

Outstanding performance on Intel architecture including Intel® Pentium® 4, Intel® Xeon™ and Intel Itanium® 2 processors, as well as support for Hyper-Threading Technology.

Compatibility

- **Plugs into Microsoft Visual Studio* .NET**
- **Microsoft PowerStation4 language and library support**
- **Strong compatibility with Compaq* Visual Fortran**

Support

1 year of free product upgrades and Intel Premier Support

"The Intel Fortran Compiler 7.0 was first-rate, and Intel Visual Fortran 8.0 is even better. Intel has made a giant leap forward in combining the best features of Compaq Visual Fortran and Intel Fortran. This compiler... continues to be a 'must-have' tool for any Twenty-First Century Fortran migration or software development project."

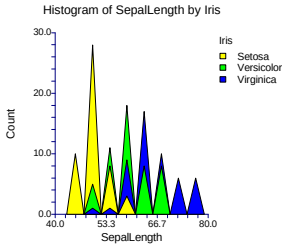
—Dr. Robert R. Trippi
Professor Computational Finance
University of California, San Diego

FREE trials available at:
programmersparadise.com/intel

Programmer's Paradise®

To order or request additional information call:
800-423-9990
Email: intel@programmers.com

NCSS
329 North 1000 East
Kaysville, Utah 84037



Announcing NCSS 2004
Seventeen New Procedures

NCSS 2004 is a new edition of our popular statistical **NCSS** package that adds seventeen new procedures.

New Procedures
Two Independent Proportions
Two Correlated Proportions
One-Sample Binary Diagnostic Tests
Two-Sample Binary Diagnostic Tests
Paired-Sample Binary Diagnostic Tests
Cluster Sample Binary Diagnostic Tests
Meta-Analysis of Proportions
Meta-Analysis of Correlated Proportions
Meta-Analysis of Means
Meta-Analysis of Hazard Ratios
Curve Fitting
Tolerance Intervals
Comparative Histograms
ROC Curves
Elapsed Time Calculator
T-Test from Means and SD's
Hybrid Appraisal (Feedback) Model

Documentation
The printed, 330-page manual, called ***NCSS User's Guide V***, is available for $29.95. An electronic (pdf) version of the manual is included on the distribution CD and in the Help system.

Two Proportions
Several new exact and asymptotic techniques were added for hypothesis testing (null, noninferiority, equivalence) and calculating confidence intervals for the difference, ratio, and odds ratio. Designs may be independent or paired. Methods include: Farrington & Manning, Gart & Nam, Conditional & Unconditional Exact, Wilson's Score, Miettinen & Nurminen, and Chen.

Meta-Analysis
Procedures for combining studies measuring paired proportions, means, independent proportions, and hazard ratios are available. Plots include the forest plot, radial plot, and L'Abbe plot. Both fixed and random effects models are available for combining the results.

Curve Fitting
This procedure combines several of our curve fitting programs into one module. It adds many new models such as Michaelis-Menten. It analyzes curves from several groups. It compares fitted models across groups using computer-intensive randomization tests. It computes bootstrap confidence intervals.

Tolerance Intervals
This procedure calculates one and two sided tolerance intervals using both distribution-free (nonparametric) methods and normal distribution (parametric) methods. Tolerance intervals are bounds between which a given percentage of a population falls.

Comparative Histogram
This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups.

Random Number Generator
Matsumoto's Mersenne Twister random number generator (cycle length > 10**6000) has been implemented.

Binary Diagnostic Tests
Four new procedures provide the specialized analysis necessary for diagnostic testing with binary outcome data. These provide appropriate specificity and sensitivity output. Four experimental designs can be analyzed including independent or paired groups, comparison with a gold standard, and cluster randomized.

ROC Curves
This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUC's and partial AUC's for paired and independent sample designs.

Hybrid (Feedback) Model
This new edition of our hybrid appraisal model fitting program includes several new optimization methods for calibrating parameters including a new genetic algorithm. Model specification is easier. Binary variables are automatically generated from class variables.

Statistical Innovations Products
Through a *special arrangement* with Statistical Innovations (S.I.), NCSS customers will receive $100 discounts on:

Latent GOLDÒ - latent class modeling
SI-CHAIDÒ - segmentation trees
GOLDMineRÒ - ordinal regression

For demos and other info visit:
www.statisticalinnovations.com

Please rush me the following products:

Qty

___ **NCSS 2004 CD upgrade from NCSS 2001**, $149.95 ..... $_____

___ **NCSS 2004 User's Guide V**, $29.95..... $_____

___ **NCSS 2004 CD, upgrade from earlier versions**, $249.95..... $_____

___ **NCSS 2004 Deluxe (CD and Printed Manuals),** $599.95..... $_____

___ **PASS 2002 Deluxe**, $499.95 ..... $_____

___ **Latent Gold® from S.I.,** $995 - $100 NCSS Discount = $895..... $_____

___ **GoldMineR® from S.I.**, $695 - $100 NCSS Discount = $595 ..... $_____

___ **CHAID® Plus from S.I.,** $695 - $100 NCSS Discount = $595.... $_____

Approximate shipping--depends on which manuals are ordered (U.S: $10 ground, $18 2-day, or $33 overnight) (Canada $24) (All other countries $10) (Add $5 U.S. or $40 International for any S.I. product) ..... $_____

Total..... $_____

TO PLACE YOUR ORDER

CALL: (800) 898-6109 **FAX:** (801) 546-3907

ONLINE: **www.ncss.com**

MAIL: NCSS, 329 North 1000 East, Kaysville, UT 84037

My Payment Option:

___ Check enclosed

___ Please charge my: __VISA __ MasterCard ___Amex

___ Purchase order attached_____

Card Number _____ Exp _____

Signature_____

Telephone:

( ) _____

Email:

_____

Ship to:

NAME _____

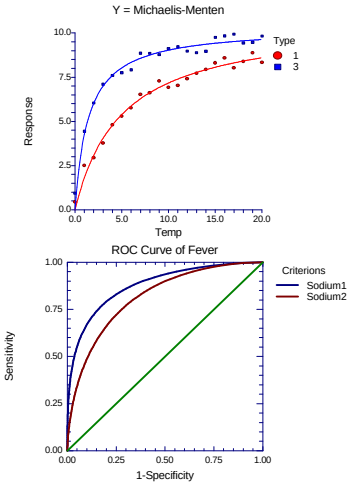
ADDRESS _____

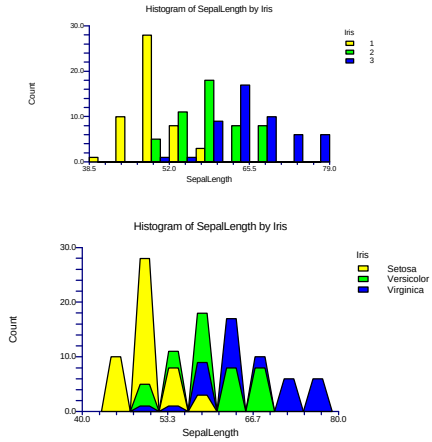
ADDRESS _____

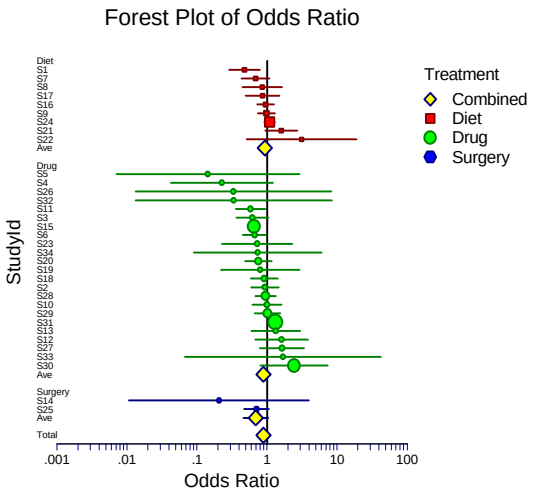
ADDRESS _____

CITY _____ STATE _____

ZIP/POSTAL CODE _____ COUNTRY _____







Statistical and Graphics Procedures Available in NCSS 2004

Analysis of Variance / T-Tests
Analysis of Covariance
Analysis of Variance
Barlett Variance Test
Crossover Design Analysis
Factorial Design Analysis
Friedman Test
Geiser-Greenhouse Correction
General Linear Models
Mann-Whitney Test
MANOVA
Multiple Comparison Tests
One-Way ANOVA
Paired T-Tests
Power Calculations
Repeated Measures ANOVA
T-Tests – One or Two Groups
T-Tests – From Means & SD's
Wilcoxon Test

Time Series Analysis
ARIMA / Box - Jenkins
Decomposition
Exponential Smoothing
Harmonic Analysis
Holt - Winters
Seasonal Analysis
Spectral Analysis
Trend Analysis

***New Edition in 2004**

Plots / Graphs
Bar Charts
Box Plots
Contour Plot
Dot Plots
Error Bar Charts
Histograms
Histograms: Combined*
Percentile Plots
Pie Charts
Probability Plots
ROC Curves*
Scatter Plots
Scatter Plot Matrix
Surface Plots
Violin Plots

Experimental Designs
Balanced Inc. Block
Box-Behnken
Central Composite
D-Optimal Designs
Fractional Factorial
Latin Squares
Placket-Burman
Response Surface
Screening
Taguchi

Regression / Correlation
All-Possible Search
Canonical Correlation
Correlation Matrices
Cox Regression
Kendall's Tau Correlation
Linear Regression
Logistic Regression
Multiple Regression
Nonlinear Regression
PC Regression
Poisson Regression
Response-Surface
Ridge Regression
Robust Regression
Stepwise Regression
Spearman Correlation
Variable Selection

Quality Control
Xbar-R Chart
C, P, NP, U Charts
Capability Analysis
Cusum, EWMA Chart
Individuals Chart
Moving Average Chart
Pareto Chart
R & R Studies

Survival / Reliability
Accelerated Life Tests
Cox Regression
Cumulative Incidence
Exponential Fitting
Extreme-Value Fitting
Hazard Rates
Kaplan-Meier Curves
Life-Table Analysis
Lognormal Fitting
Log-Rank Tests
Probit Analysis
Proportional-Hazards
Reliability Analysis
Survival Distributions
Time Calculator*
Weibull Analysis

Multivariate Analysis
Cluster Analysis
Correspondence Analysis
Discriminant Analysis
Factor Analysis
Hotelling's T-Squared
Item Analysis
Item Response Analysis
Loglinear Models
MANOVA
Multi-Way Tables
Multidimensional Scaling
Principal Components

Curve Fitting
Bootstrap C.I.'s*
Built-In Models
Group Fitting and Testing*
Model Searching
Nonlinear Regression
Randomization Tests*
Ratio of Polynomials
User-Specified Models

Miscellaneous
Area Under Curve
Bootstrapping
Chi-Square Test
Confidence Limits
Cross Tabulation
Data Screening
Fisher's Exact Test
Frequency Distributions
Mantel-Haenszel Test
Nonparametric Tests
Normality Tests
Probability Calculator
Proportion Tests
Randomization Tests
Tables of Means, Etc.
Trimmed Means
Univariate Statistics

Meta-Analysis*
Independent Proportions*
Correlated Proportions*
Hazard Ratios*
Means*

Binary Diagnostic Tests*
One Sample*
Two Samples*
Paired Samples*
Clustered Samples*

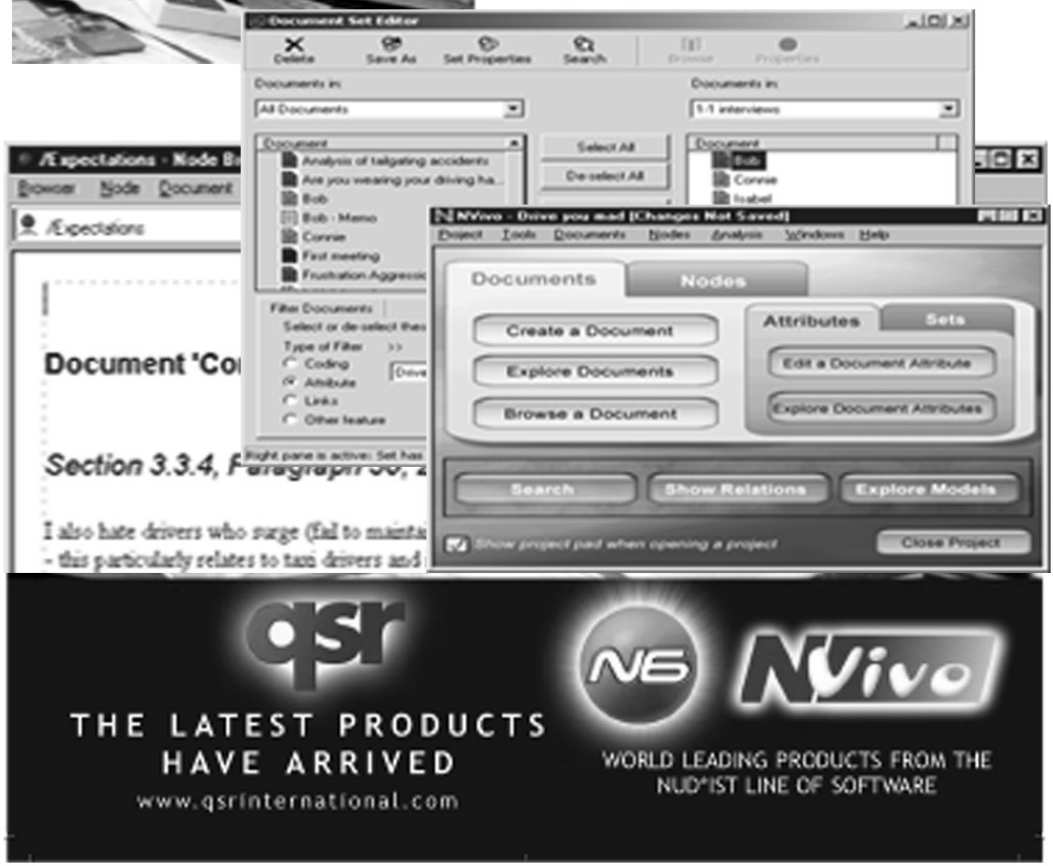
Proportions
Tolerance Intervals*
Two Independent*
Two Correlated*
Exact Tests*
Exact Confidence Intervals*
Farrington-Manning*
Fisher Exact Test
Gart-Nam* Method
McNemar Test
Miettinen-Nurminen*
Wilson's Score* Method
Equivalence Tests*
Noninferiority Tests*

Mass Appraisal
Comparables Reports
Hybrid (Feedback) Model*
Nonlinear Regression
Sales Ratios

Qualitative research has come a long way...

from this...

to this!



Read more about QSR software in this edition of JMASM.

PASS 2002

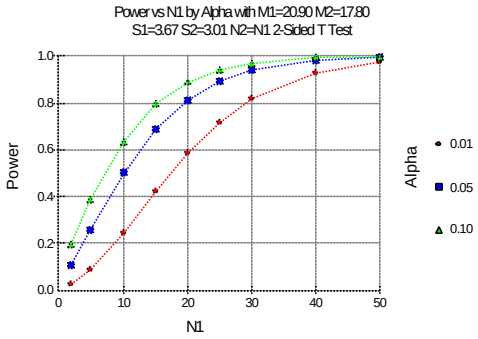
Power Analysis and Sample Size Software from NCSS

PASS performs power analysis and calculates sample sizes. Use it before you begin a study to calculate an appropriate sample size (it meets the requirements of government agencies that want technical justification of the sample size you have used). Use it after a study to determine if your sample size was large enough. ***PASS*** calculates the sample sizes necessary to perform all of the statistical tests listed below.

A power analysis usually involves several "what if" questions. ***PASS*** lets you solve for power, sample size, effect size, and alpha level. It automatically creates appropriate tables and charts of the results.

PASS is accurate. It has been extensively verified using books and reference articles. Proof of the accuracy of each procedure is included in the extensive documentation.

PASS is a standalone system. Although it is integrated with ***NCSS***, you do not have to own ***NCSS*** to run it. You can use it with any statistical software you want.



PASS comes with two manuals that contain tutorials, examples, annotated output, references, formulas, verification, and complete instructions on each procedure. And, if you cannot find an answer in the manual, our free technical support staff (which includes a PhD statistician) is available.

System Requirements

PASS runs on Windows 95/98/ME/NT/ 2000/XP with at least 32 megs of RAM and 30 megs of hard disk space.

PASS sells for as little as **$449.95**.

PASS Beats the Competition!

No other program calculates sample sizes and power for as many different statistical procedures as does *PASS*. Specifying your input is easy, especially with the online help and manual.

PASS automatically displays charts and graphs along with numeric tables and text summaries in a portable format that is cut and paste compatible with all word processors so you can easily include the results in your proposal.

Choose ***PASS***. It's more comprehensive, easier-to-use, accurate, and less expensive than any other sample size program on the market.

Trial Copy Available

You can try out ***PASS*** by downloading it from our website. This trial copy is good for 30 days. We are sure you will agree that it is the easiest and most comprehensive power analysis and sample size program available.

Analysis of Variance
Factorial AOV
Fixed Effects AOV
Geisser-Greenhouse
MANOVA*
Multiple Comparisons*
One-Way AOV
Planned Comparisons
Randomized Block AOV
New Repeated Measures AOV*

Regression / Correlation
Correlations (one or two)
Cox Regression*
Logistic Regression
Multiple Regression
Poisson Regression*
Intraclass Correlation
Linear Regression

Proportions
Chi-Square Test
Confidence Interval
Equivalence of McNemar*
Equivalence of Proportions
Fisher's Exact Test
Group Sequential Proportions
Matched Case-Control
McNemar Test
Odds Ratio Estimator
One-Stage Designs*
Proportions – 1 or 2
Two Stage Designs (Simon's)
Three-Stage Designs*

Miscellaneous Tests
Exponential Means – 1 or 2*
ROC Curves – 1 or 2*
Variances – 1 or 2

T Tests
Cluster Randomization
Confidence Intervals
Equivalence T Tests
Hotelling's T-Squared*
Group Sequential T Tests
Mann-Whitney Test
One-Sample T-Tests
Paired T-Tests
Standard Deviation Estimator
Two-Sample T-Tests
Wilcoxon Test

Survival Analysis
Cox Regression*
Logrank Survival -Simple
Logrank Survival - Advanced*
Group Sequential - Survival
Post-Marketing Surveillance
ROC Curves – 1 or 2*

Group Sequential Tests
Alpha Spending Functions
Lan-DeMets Approach
Means
Proportions
Survival Curves

Equivalence
Means
Proportions
Correlated Proportions*

Miscellaneous Features
Automatic Graphics
Finite Population Corrections
Solves for any parameter
Text Summary
Unequal N's

***New in PASS 2002**

NCSS Statistical Software • 329 North 1000 East • Kaysville, Utah 84037
Internet (download free demo version): http://www.ncss.com • Email: sales@ncss.com
Toll Free: (800) 898-6109 • Tel: (801) 546-0445 • Fax: (801) 546-3907

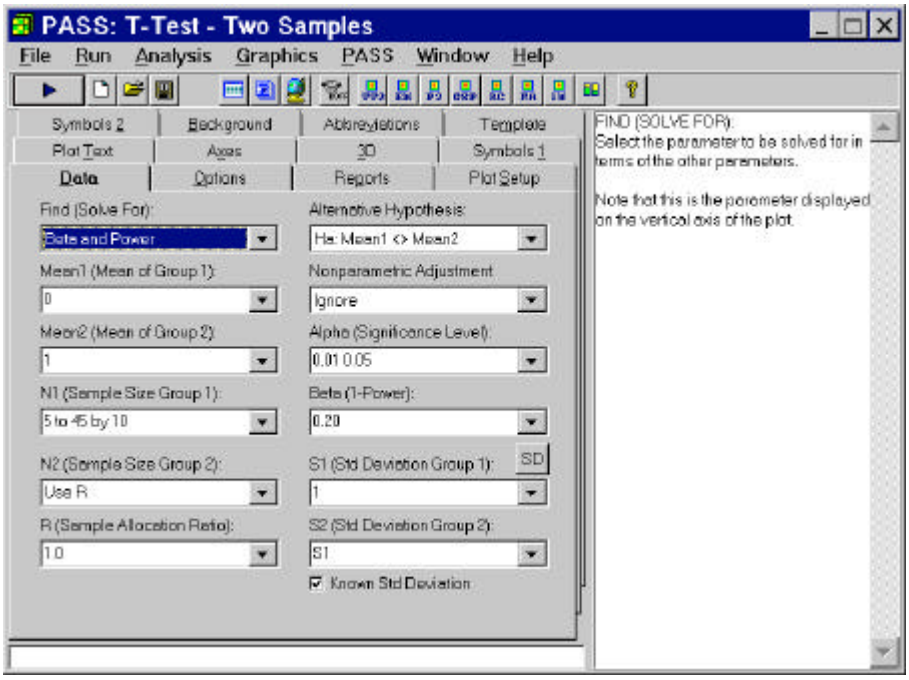
PASS 2002 adds power analysis and sample size to your statistical toolbox

WHAT'S NEW IN PASS 2002?
Thirteen new procedures have been added to ***PASS*** as well as a new home-base window and a new Guide Me facility.

MANY NEW PROCEDURES
The new procedures include a new multi-factor repeated measures program that includes multivariate tests, Cox proportional hazards regression, Poisson regression, MANOVA, equivalence testing when proportions are correlated, multiple comparisons, ROC curves, and Hotelling's T-squared.

TEXT STATEMENTS
The text output translates the numeric output into easy-to-understand sentences. These statements may be transferred directly into your grant proposals and reports.

GRAPHICS
The creation of charts and graphs is easy in ***PASS***. These charts are easily transferred into other programs such as MS PowerPoint and MS Word.



NEW USER'S GUIDE II
A new, 250-page manual describes each new procedure in detail. Each chapter contains explanations, formulas, examples, and accuracy verification.

The complete manual is stored in PDF format on the CD so that you can read and printout your own copy.

GUIDE ME
The new *Guide Me* facility makes it easy for first time users to enter parameter values. The program literally steps you through those options that are necessary for the sample size calculation.

NEW HOME BASE
A new home base window has been added just for PASS users. This window helps you select the appropriate program module.

COX REGRESSION
A new Cox regression procedure has been added to perform power analysis and sample size calculation for this important statistical technique.

REPEATED MEASURES
A new repeated-measures analysis module has been added that lets you analyze designs with up to three grouping factors and up to three repeated factors. The analysis includes both the univariate F test and three common multivariate tests including Wilks Lambda.

RECENT REVIEW
In a recent review, 17 of 19 reviewers selected ***PASS*** as the program they would recommend to their colleagues.

PASS calculates sample sizes for...

Please rush me my own personal license of ***PASS 2002.***

Qty

___ **PASS 2002 Deluxe** (CD and User's Guide)**:** $499.95.....$ _____

___ **PASS 2002 CD (electronic documentation):** $449.95.....$ _____

___ **PASS 2002 5-User Pack (CD & 5 licenses):** $1495.00.....$ _____

___ **PASS 2002 25-User Pack (CD & 25 licenses):** $3995.00....$ _____

___ **PASS 2002 User's Guide II (printed manual):** $30.00.....$ _____

___ **PASS 2002 Upgrade CD** for *PASS 2000* users**:** $149.95.....$ _____

Typical Shipping & Handling: USA: $9 regular, $22 2-day, $33 overnight. Canada: $19 Mail. Europe: $50 Fedex.....$ _____

Total: .....$ _____

FOR FASTEST DELIVERY, ORDER ONLINE AT

WWW.NCSS.COM

Email your order to sales@ncss.com
Fax your order to (801) 546-3907
NCSS, 329 North 1000 East, Kaysville, UT 84037
(800) 898-6109 or (801) 546-0445

My Payment Options:

___ Check enclosed
___ Please charge my: __VISA __MasterCard __Amex
___ Purchase order enclosed

Card Number
_____Expires_____

Signature_____
Please provide daytime phone:

( )_____

Ship my *PASS 2002* to:

NAME

COMPANY

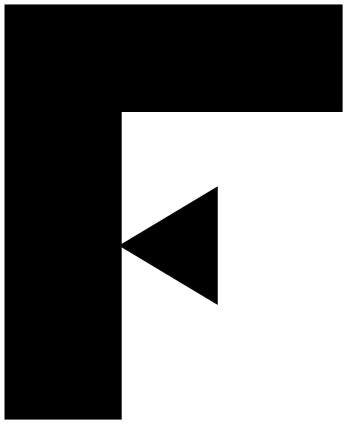
ADDRESS

CITY/STATE/ZIP

COUNTRY (IF OTHER THAN U.S.)

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."

- Antoine de Saint Exupery



F is a carefully crafted subset of the most recent version of Fortran, the world's most powerful numeric language.

Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- F is easier to use than other popular programming languages
- ***F compilers are free*** and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran—modules and data abstraction, for example—but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.

F encourages Module-Oriented Programming.

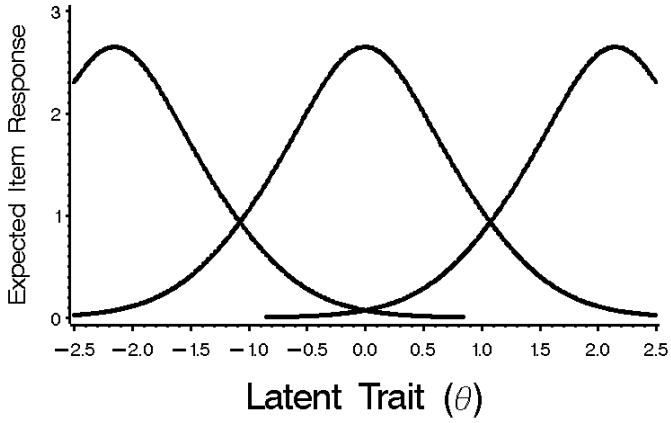
It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.

The Fortran Company
11155 E. Mountain Gate Place, Tucson, AZ 85749 USA
+1-520-256-1455 +1-520-760-1397 (fax)
http://www.fortran.com info@fortran.com

Introducing GGUM2004

Item Response Theory Models for Unfolding



The new GGUM2004 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2004 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected *a posteriori* (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. GGUM2004 is compatible with computers running updated versions of Windows 98 SE, Windows 2000, and Windows XP. The software is accompanied by a detailed technical reference manual and a new Windows user's guide. ***GGUM2004 is free*** and can be downloaded from:

http://www.education.umd.edu/EDMS/tutorials

GGUM2004 improves upon its predecessor (GGUM2000) in several important ways:

- It has a user-friendly graphical interface for running commands and displaying output.
- It offers real-time graphics that characterize the performance of a given model.
- It provides new item fit indices with desirable statistical characteristics.
- It allows for missing item responses assuming the data are missing at random.
- It allows the number of response categories to vary across items.
- It estimates model parameters more quickly.

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2004 today!

WWW.DIGARCH.COM

DIGITAL EXCAVATOR

Are you involved in Data Modeling or Data Mining?

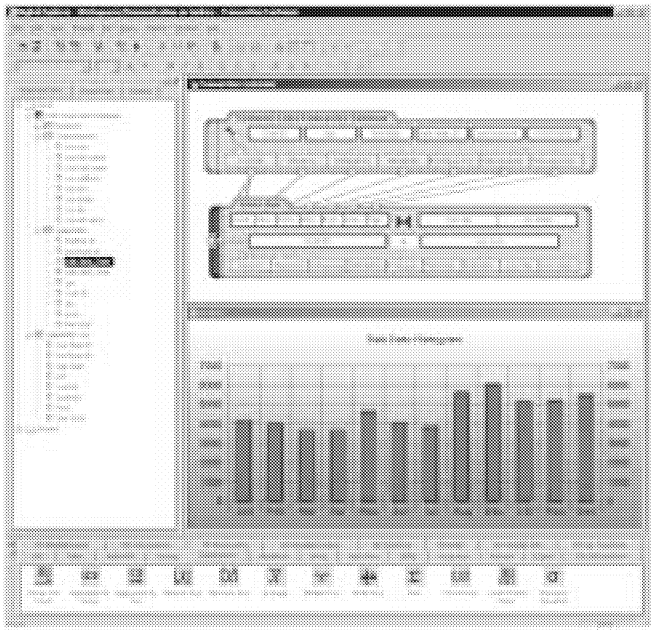
Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology (www.digarch.com). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent 60%-80% of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

Visit our website and see for yourself! >>>> www.digarch.com

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others



DIGITAL ARCHAEOLOGY
A DELANO COMPANY

15721 COLLEGE BOULEVARD
LENEXA, KS 66219
1-877-DIGARCH (344-2724)
WWW.DIGARCH.COM

DataMineIt℠

announces

PermuteIt™ v2.0

The fastest, most comprehensive and robust permutation test software on the market today.

Permutation tests increasingly are the statistical method of choice for addressing business questions and research hypotheses across a broad range of industries. Their distribution-free nature maintains test validity where many parametric tests (and even other nonparametric tests), encumbered by restrictive and often inappropriate data assumptions, fail miserably. The computational demands of permutation tests, however, have severely limited other vendors' attempts at providing useable permutation test software for anything but highly stylized situations or small datasets and few tests. PermuteIt™ addresses this unmet need by utilizing a combination of algorithms to perform non-parametric permutation tests very quickly – often more than an order of magnitude faster than widely available commercial alternatives when one sample is large and many tests and/or multiple comparisons are being performed (which is when runtimes matter most). PermuteIt™ can make the difference between making deadlines, or missing them, since data inputs often need to be revised, resent, or recleaned, and one hour of runtime quickly can become 10, 20, or 30 hours.

In addition to its speed even when one sample is large, some of the unique and powerful features of PermuteIt™ include:

- **the availability to the user of a wide range of test statistics for performing permutation tests on continuous, count, & binary data, including: pooled-variance t-test; separate-variance Behrens-Fisher t-test, scale test, and joint tests for scale and location coefficients using nonparametric combination methodology; Brownie et al. "modified" t-test; skew-adjusted "modified" t-test; Cochran-Armitage test; exact inference; Poisson normal-approximate test; Fisher's exact test; Freeman-Tukey Double Arcsine test**

- **extremely fast exact inference (no confidence intervals – just exact p-values) for most count data and high-frequency continuous data, often several orders of magnitude faster than the most widely available commercial alternative**

- **the availability to the user of a wide range of multiple testing procedures, including: Bonferroni, Sidak, Stepdown Bonferroni, Stepdown Sidak, Stepdown Bonferroni and Stepdown Sidak for discrete distributions, Hochberg Stepup, FDR, Dunnett's one-step (for MCC under ANOVA assumptions), Single-step Permutation, Stepdown Permutation, Single-step and Stepdown Permutation for discrete distributions, Permutation-style adjustment of permutation p-values**

- **fast, efficient, and automatic generation of all pairwise comparisons**

- **efficient variance-reduction under conventional Monte Carlo via self-adjusting permutation sampling when confidence intervals contain the user-specified critical value of the test**

- **maximum power, and the shortest confidence intervals, under conventional Monte Carlo via a new sampling optimization technique (see Opdyke, JMASM, Vol. 2, No. 1, May, 2003)**

- **fast permutation-style p-value adjustments for multiple comparisons (the code is designed to provide an additional speed premium for many of these resampling-based multiple testing procedures)**

- **simultaneous permutation testing and permutation-style p-value adjustment, although for relatively few tests at a time (this capability is not even provided as a preprogrammed option with any other software currently on the market)**

For Telecommunications, Pharmaceuticals, *f*MRI data, Financial Services, Clinical Trials, Insurance, Bioinformatics, and just about any data rich industry where large numbers of distributional null hypotheses need to be tested on samples that are not extremely small and parametric assumptions are either uncertain or inappropriate, PermuteIt™ is the optimal, and only, solution.

To learn more about how PermuteIt™ can be used for your enterprise, and to obtain a demo version, please contact its author, J.D. Opdyke, President, DataMineIt℠, at JDOpdyke@DataMineIt.com or www.DataMineIt.com.

DataMineIt℠ is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. PermuteIt™ is its flagship product.

JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to *Psychological Methods* or *Psychological Assessment* (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional $18)
- *The Score* – the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: $38 (**APA membership not required**); student membership is only $8

For further information, please contact the Division's Membership Chair, Yossef Ben-Porath (ybenpora@kent.edu) or check out the Division's website:

http://www.apa.org/divisions/div5/

ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the **Special Interest Group - Educational Statisticians** of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.
We also publish a twice-yearly electronic newsletter.

Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at http://orme.uark.edu/edstatsig.htm

To join SIG-ES you must be a member of AERA. Dues are $5.00 per year.

For more information, contact Joan Garfield, President of the SIG-ES, at jbg@umn.edu.

Lahey
Computer Systems, Inc.
SOFTWARE SOLUTIONS
for Science & Engineering

Lahey/Fujitsu Fortran

The standard for Fortran programming from the leader in Fortran language systems

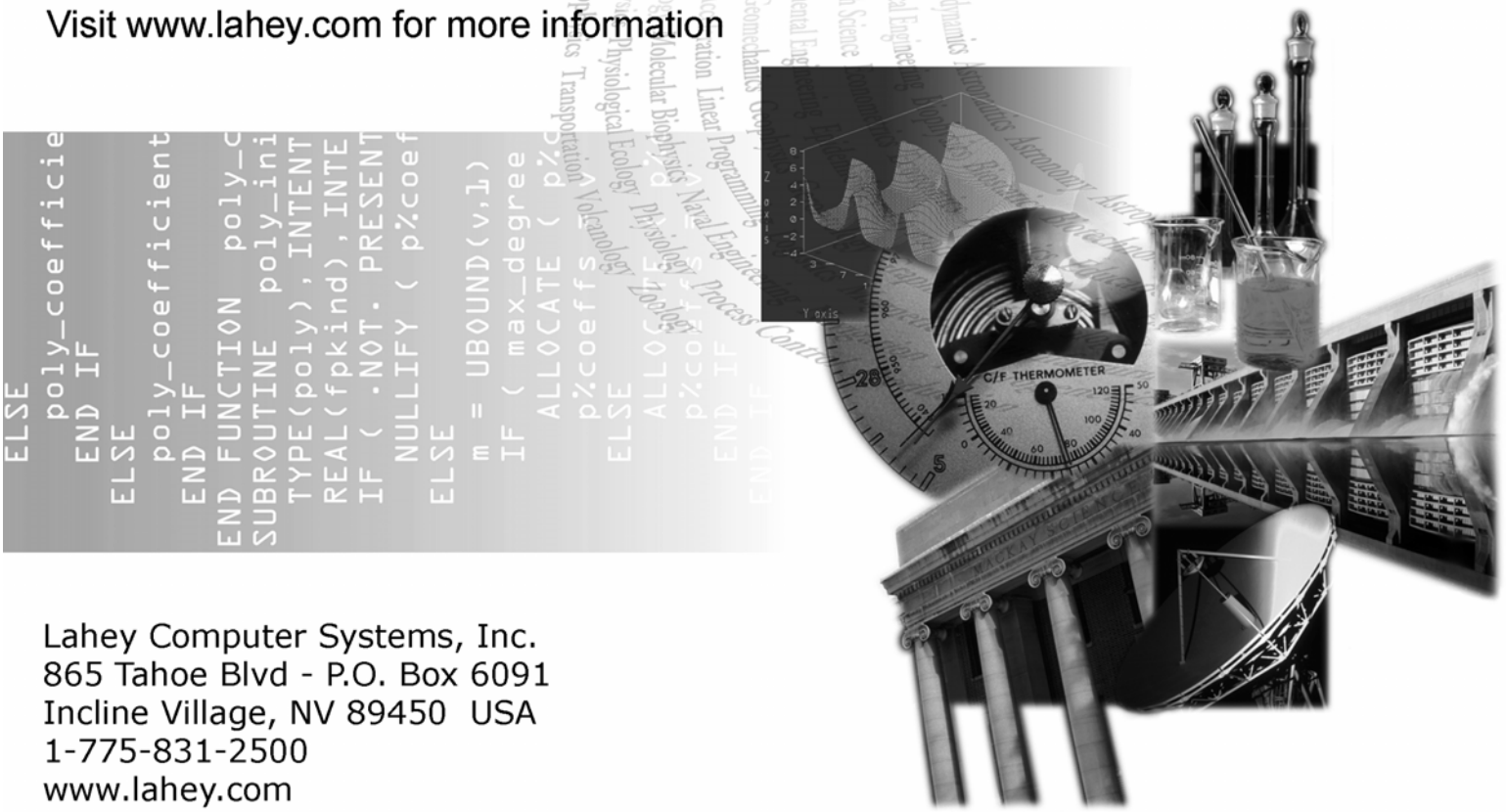
LF95 Fortran for Linux and Windows

- Full Fortran 95/90/77 support
- Unsurpassed diagnostics
- Intel and AMD optimizations
- IMSL compatible
- Fujitsu SSL2 math library
- Wisk graphics package

LF Fortran for the Microsoft® .NET Framework - Coming Soon !

- Visual Studio integration
- Windows / Web Forms designer
- Project and code templates
- On-line integrated help
- XML Web services
- ADO.NET support

Visit www.lahey.com for more information



Lahey Computer Systems, Inc.
865 Tahoe Blvd - P.O. Box 6091
Incline Village, NV 89450 USA
1-775-831-2500
www.lahey.com

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-352

1538 – 9472/04/$95.00

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@edstat.coe.wayne.edu. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are not acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified, indent optional.
8. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
9. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "&" instead of "and" in multiple author listings.
10. *Suggestions for style*: Instead of "I drew a sample of 40" write "A sample of 40 was selected". Use "although" instead of "while", unless the meaning is "at the same time". Use "because" instead of "since", unless the meaning is "after". Instead of "Smith (1990) notes" write "Smith (1990) noted". Do not strike spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professionals are US $95 per year; for graduate students are US $47.50 per year; and for libraries, universities, and corporations are US $195 per year. Subscribers outside of the US and Canada pay a US $10 surcharge for additional postage. Online access is currently free at http://tbf.coe.wayne.edu/jmasm. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@edstat.coe.wayne.edu.

Notice To Advertisers

Send requests for advertising information to jmasm@edstat.coe.wayne.edu.

STATISTICIANS

HAVE YOU VISITED THE

Mathematics Genealogy Project?

The Mathematics Genealogy Project is an ongoing research project tracing the intellectual history of all the mathematical arts and sciences through an individual's Ph.D. advisor and Ph.D. students. Currently we have over 80,000 records in our database. We welcome and encourage all statisticians to join us in this endeavor.

Please visit our web site

http://genealogy.math.ndsu.nodak.edu

The information which we collect is the following:
The full name of the individual, the school where he/she earned a Ph.D., the year of the degree, the title of the dissertation, and, MOST IMPORTANTLY, the full name of the advisor(s). E.g., Fuller, Wayne Arthur; Iowa State University; 1959; ***A Non-Static Model of the Beef and Pork Economy*;** Shepherd, Geoffrey Seddon

For additions or corrections for one or two people a link is available on the site. For contributions of large sets of names, e.g., all graduates of a given university, it is better to send the data in a text file or an MS Word file or an MS Excel file, etc. Send such information to:

harry.coonce@ndsu.nodak.edu

The genealogy project is a not-for-profit endeavor supported by donations from individuals and sales of posters and t-shirts. If you would like to help this cause please send your tax-deductible contribution to:
Mathematics Genealogy Project, 300 Minard Hall, P. O. Box 5075, Fargo, North Dakota 58105-5075E