“AN ADVANCE STUDY OF COVARIANCE STRUCTURE”

A Thesis submitted to the
Saurashtra University
for the Degree of Doctor of Philosophy
in
Statistics

By
VIREN B. MANEK

Under the Guidance of
Dr. G.C.BHIMANI
Professor of Statistics
Department of Statistics
Saurashtra University
Rajkot – 360005
Gujarat (India)

May - 2011

ACKNOWLEDGEMENT

I would like to thank my guide Dr. G. C. Bhimani for his thoughtful guidance and support. Besides being an excellent guide he has been a great mentor who has helped me develop my self – confidence and always encouraged me to push my boundaries and take on new challenges in life.

I appreciatively acknowledge the rendered by Dr. D. K. Ghosh, Professor & Head, Department of Satistics and other faculty of the deparment for being a constant source of inspiration.

Ths thesis is dedicated to my parents, son and family members. My parents & family members integrity, humility, love and compassion for all has left an indelible impression in my life. I am eternally grateful for their constant encouragement and for setting.

I express thanks to all colleagues for their co – operation.

Microsoft Office 2007 was used to prepare this thesis and all calculation were done with the statistical software. This thesis contains much of thesis effort not in term of paragraphs or tables, rather their understanding and support all the way.

Viren B. Manek

CERTIFICATE

This is to certify that the Thesis entitled **"An Advance Study Of Covariance Structure"** submitted by **Mr. Viren B. Manek** for the award of the **Degree of Doctor of Philosophy in Statistics** is a bonafide research work done independently by my guidance.

(Dr. G. C. Bhimani)
Professor of Statistics
Department of Statistics
Saurashtra University
Rajkot – 360005

DECLARATION

I hereby declared that the research work on **"An Advance Study Of Covariance Structure"** is carried out by me and results of this work have not submitted at any other university for the award of the Degree of Doctor of Philosophy in Statistics.

(Viren B. Manek)
Ph.D. Student
Department of Statistics
Saurashtra University
Rajkot – 360005

Contents

Chapter – 1
Introduction

Chapter – 1
Introduction

This article examines the adjustment of normal theory methods for the analysis of covariance structures to make them applicable under the class of elliptical distributions. It is shown that if the model satisfies a mild scale invariance condition and the data have an elliptical distribution, the asymptotic covariance matrix of sample covariances has a structure that results in the retention of many of the asymptotic properties of normal theory methods. If a scale adjustment is applied, the likelihood ratio tests of fit have the usual asymptotic chi-squared distributions. Difference tests retain their property of asymptotic independence, and maximum likelihood estimators retain their relative asymptotic efficiency within the class of estimators based on the sample covariance matrix. An adjustment to the asymptotic covariance matrix of normal theory maximum likelihood estimators for elliptical distributions is provided. This adjustment is particularly simple in models for patterned covariance or correlation matrices. These results apply not only to normal theory maximum likelihood methods but also to a class of minimum discrepancy methods. Similar results also apply when certain robust estimators of the covariance matrix are employed.

A considerable part of classical multivariate analysis is devoted to hypotheses concerning the population covariance matrix, Z. The associated statistical inference is well developed under the assumption that the sample is drawn from a normally distributed population (e.g., Muirhead 1982). These normal theory methods can, however, be sensitive to deviations from normality and, in

particular, to the kurtosis of data distributions. Multivariate distributions that are convenient for investigating the sensitivity of normal theory methods to kurtosis are the elliptical distributions (Chmielewski 1981; Devlin, Gnanadesikan, and Kettenring 1976; Muirhead 1982, sec. 1.5). The elliptical class of distributions incorporates a single additional kurtosis parameter, K, and contains the multivariate normal distribution as a special case with $\boldsymbol{K} = 0$.

Elliptical distributions have been employed in two general approaches yielding somewhat different results. In one, an N x p data matrix is regarded as being distributed according to an Np-dimensional elliptical distribution. Elements in different rows of the data matrix are regarded as uncorrelated but not independent if K # 0. Under these conditions certain normal theory likelihood ratio tests remain valid without correction (Anderson, Fang, and Hsu 1986; Chmielewski 1980).

The present article will adopt the other approach where rows of the data matrix are regarded as being independently and identically distributed according to a p-variate elliptical distribution. Under these assumptions a number of situations were found (Muirhead 1982; Muirhead and Waternaux 1980) where normal theory likelihood ratio tests retain their asymptotic chi-squared distribution if divided by a correction factor dependent on kurtosis. Tyler (1983) gave a class of null hypotheses defined by equality constraints on elements of Z where these scale corrections for the likelihood ratio test are applicable. A class of structural models Z = Z(8), where scale corrections for likelihood ratio goodness-of-fit tests are applicable and normal theory maximum likelihood parameter estimators retain

their relative asymptotic efficiency within a certain class of estimators, was given in Browne (1982, 1984). Tyler (1982, 1983) also showed that correction factors can be found when the usual sample covariance matrix is replaced by an alternative estimator from the class of Mestimators (Maronna 1976) or by an estimator of Z that is a maximum likelihood estimator under the assumption of some specific elliptical distribution.

The present article unifies and extends the findings of Tyler (1982, 1983) and Browne (1982, 1984). This is done by showing that their two superficially different sets of conditions on Z both imply a property of the model that justifies scale corrections to the test statistic. We consider a class of minimum discrepancy test statistics, which includes the previously considered normal theory likelihood ratio statistic, and show that similar scale corrections also apply to difference tests with constrained alternative hypotheses. A new test statistic that does not require a scale correction for the kurtosis of an elliptical distribution is also obtained. The result concerning robustness of the asymptotic efficiency of maximum likelihood estimators given by Browne (1982, 1984) is extended to other discrepancy functions based on the wider class of covariance matrix estimates considered by Tyler (1982, 1983). In addition, we provide a new correction factor of rank 1to the asymptotic covariance matrix of estimators that is more direct and simpler to apply than the correction factor given in Browne (1982, 1984).

A general approach to the analysis of covariance structures is considered, in which the variances and covariances or correlations of the observed variables are directly expressed in terms of the parameters of interest. The statistical

problems of identification, estimation and testing of such covariance or correlation structures are discussed.

Several different types of covariance structures are considered as special cases of the general model. These include models for sets of congeneric tests, models for confirmatory and exploratory factor analysis, models for estimation of variance and covariance components, regression models with measurement errors, path analysis models, simplex and circumplex models. Many of the different types of covariance structures are illustrated by means of real data.

The search for structure in correlated psychological variables has been one of the main objectives in psychometrics for several decades. Traditionally this search was done by using factor analysis to detect and assess latent sources of variation and covariation in observed measurements. Seldom do these measurements represent pure psychological traits or functions. Rather, as Thurstone [1947] assumed in his multiple factor model, each measure depends on a limited number of traits or functions and one tries to identify, and ultimately estimate, the components of the observed measurements associated with different traits or functions.

In factor analysis the correlation matrix is subjected to a suitable method for estimation of the factor space, the solution rotated to obtain projections of the test vectors on certain reference vectors, called factors, and, by examining the contents of the tests which have large projections on a particular reference vector, a trait or function is inferred to be common to these psychological tests. The trait or function, treated as an explanatory variable is then named and considered to be

a source of one of the components of covariation or correlation in the tests analyzed. Individual differences in this component can then be estimated as so called factor scores.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Chapter – 2

Some remarks on estimating a covariance structure from a sample correlation matrix

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Chapter – 2
Some remarks on estimating a covariance structure from a sample correlation matrix

2.1 Introduction

In covariance structure analysis, one wishes to model the variances and covariances of the observed variables. That is, one assumes that the population covariance matrix $\sum$ of the observed variables depends on a parameter vector θ, say $\sum(\theta)$, whereas no structure is imposed on the population mean vector μ. The objective is then to estimate the parameter vector θ from a sample covariance matrix. In contrast, in correlation structure analysis, one wishes to model the correlations among the observed variables. Thus, in this case it is the population correlation matrix $\boldsymbol{P}$ which is assumed to depend on a parameter vector θ, say $\boldsymbol{P}(\theta)$, whereas, as before, no structure is imposed on the population mean vector μ. Correlation structure analysis is often chosen when the observed variables have different and arbitrary scales. In this case, researchers may feel that it is more meaningful to transform the observed variables to standard deviation scales. In contrast, when all observed variables are on the same scale, researchers may feel that it is more appropriate to fit a covariance structure.

It is not the aim of this paper to elaborate on when to perform covariance vs. correlation structure analysis. Rather, this paper aims at discussing the case in which a researcher wishes to estimate a covariance structure but s/he is unable to do so from a sample covariance matrix because only a sample correlation matrix is available for analysis. Estimating a covariance structure from a sample correlation

matrix is not a trivial matter. Cudeck thoroughly reviewed this topic pointing out that doing this may result in (a) fitting a different model that the one intended, (b) incorrect !2 and other goodness-of-fit measures, and (c) incorrect standard errors. Given these problems, one should estimate a covariance structure from a sample covariance matrix if at all possible. However, in some cases is not possible. For instance

a) When all observed variables are categorical. Although covariance structure analysis was originally developed as a technique for continuous variables, over the last fifteen years the most popular software packages for structural equation modeling (LISREL: Jöreskog & Sörbom, MPLUS: Muthén & Muthén, EQS: Bentler,) have incorporated routines for performing covariance structure analysis for categorical dependent variables as well by assuming that these arise by discretizing a multivariate normal distribution according to a set of thresholds. Nevertheless, when all the observed variables are categorical, then the parameters of the underlying covariance structure can not be estimated from a sample covariance matrix, as only the correlation matrix of the underlying normal variates (a matrix of tetrachoric/polychoric correlations) may be estimated.

b) When all observed variables are continuous but only a correlation matrix of is available (e.g., when one is interested in estimating a covariance structure from a published correlation matrix). Since in this case only the correlation matrix is available, estimation must proceed under multivariate normal assumptions.

Clearly, the first instance will be encountered more frequently than the second, and correspondingly, it will be the main focus of the present research. The standard procedure to fit a covariance structure to categorical observed variables when no restrictions are imposed on the thresholds consists in estimating each sample threshold and polychoric correlation separately from the first and second order marginals of the observed contingency table. Then, the parameters of the underlying covariance structure are estimated from the sample tetrachoric/polychoric correlations alone using a weighted least squares discrepancy function. By using this approach, one can estimate the covariance model parameters, obtain asymptotically correct goodness-of-fit measures and standard errors for the parameter estimates. But, as we shall show, if and only if the covariance structure being fitted is scale invariant. If this procedure is used to estimate a covariance structure that is not scale invariant, then one ends up fitting a different (and more restricted) covariance structure than the one intended. We shall also show that to fit covariance structure that is not scale invariant to categorical observed variables one must use in the final stage of the estimation procedure a weighted least squares discrepancy function using both the sample thresholds and tetrachoric/polychoric correlations. To illustrate our discussion, we shall provide a numerical example in which we fit scale invariant and non-scale invariant factor models to the well known LSAT 6 dataset (Bock & Lieberman).

Next, we shall discuss how to fit a covariance structure model to a sample correlation matrix of continuous variables. Covariance structure models can be estimated from a sample correlation matrix by minimizing a normal theory

generalized least squares function of the sample correlations under normality assumption (Jennrich, Browne & Shapiro). However, this is actually not needed. One can estimate any covariance structure from a sample correlation matrix by minimizing a normal theory discrepancy function for sample covariances. This is convenient, because to our knowledge discrepancy functions for sample correlations have not been implemented in standard software packages such as LISREL, EQS or MPLUS. Unfortunately, no standard software package can currently estimate a non-scale invariant covariance structure from a sample correlation matrix. To illustrate our discussion of the continuous case we shall use some data originally published by Jöreskog and also considered by Cudeck.

Because determining whether a model is scale invariant is critical in applications in which a covariance structure is estimated from a sample correlation matrix, we provide in an appendix computer algebra code in Mathematica (Wolfram) that may be employed to determine whether a covariance structure model is scale invariant using results from Bekker, Merckens and Wansbeek. Also, because when estimating a covariance structure from a sample correlation matrix not all covariance structure parameters may be identified we provide in another appendix Mathematica computer algebra code to be used to investigate the identification of the model parameters.

2.2 Covariance structure analysis for categorical dependent variables

Let $\mathbf{y}^* \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ and suppose that each variable y_i^* has been categorized using

$$y_i = h \quad \text{if} \quad \alpha_{i_h} < y_i^* < \alpha_{i_{h+1}} \qquad h = 0, \cdots, k-1;\ i = 1, \ldots, n \qquad (1)$$

where $\alpha_{i_0} = -\infty, \alpha_{i_k} = \infty$. That is, for notational ease, we assume that all variables y_i have the same number of categories, k. Our objective is to estimate the q-dimensional parameter vector θ from the observed categorical variables $\mathbf{y}$.

According to this model, the probability of observing any categorical pattern is $\mathbf{y_c}$

$$\Pr(\mathbf{y}_c) = \int_{\mathbf{R}} \cdots \int \phi_n\left(\mathbf{y}^* : \boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right) d\mathbf{y}^* \qquad c = 1, \ldots, k^n \qquad (2)$$

where $\phi_n(\bullet)$ denotes a n-variate normal density and the intervals of the area of integration

$$\mathbf{R} \text{ are } R_i = \left(\alpha_{i_h}, \alpha_{i_{h+1}}\right) \text{ if } y_i = h\ .$$

Because the underlying variables **y*** are normal, the pattern probabilities (2) are unchanged when we standardize each $\mathbf{y_i}$* by subtracting its mean and dividing it by its standard deviation using

$$\mathbf{z}^* = \mathbf{D}_{\boldsymbol{\theta}}\left(\mathbf{y}^* - \boldsymbol{\mu}\right) \qquad \mathbf{D}_{\boldsymbol{\theta}} = \mathrm{Diag}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)^{-\frac{1}{2}} \qquad (3)$$

where Diag($\bullet$) denotes a square matrix whose non-diagonal elements have been set to 0. Denoting by $\sigma_{ii}(\theta)$ a diagonal element of $\sum(\theta)$, the diagonal elements of $\mathbf{D}_\theta$ are of the type

$$\delta_i = \frac{1}{\sqrt{\sigma_{ii}(\boldsymbol{\theta})}} \tag{4}$$

As a result of (3), **z*** has mean zero and correlation structure

$$\mathbf{P}(\theta) = \mathbf{D}_\theta \Sigma(\theta) \mathbf{D}_\theta \tag{5}$$

i.e., $\boldsymbol{P}(\theta)$ has ones along its diagonal. Furthermore, defining

$$\tau_{i_h} := \delta_i \left(\alpha_{i_h} - \mu_i \right) \tag{6}$$

when we change the variable of integration in (2) using (3) we find that at $y_i^* = \alpha_{i_h}$ $z_i^* = \tau_{i_h}$

,

. . Thus, (2) can be equivalently written as

$$\Pr(\mathbf{y}_c) = \int_{\tilde{\mathbf{R}}} \cdots \int \phi_n \left(\mathbf{z}^* : \mathbf{0}, \mathbf{P}(\boldsymbol{\theta}) \right) d\mathbf{z}^* \tag{7}$$

with intervals of integration $\tilde{R}_i = \left(\tau_{i_h}, \tau_{i_{h+1}} \right)$ if $y_i = h$, where $\tau_{i_0} = -\infty, \tau_{i_k} = \infty$.

Now, because (2) and (7) are equivalent, we see that only the correlation structure (5) can be identified (estimated) from categorical data. There is an additional identification problem in (6), namely, that the μ's can not be separately estimated from the α's. The easiest way to solve this identification problem is to assume in applications that $\mu = \mathbf{0}$. We shall do so in the remainder of this paper. Note, however, that if we were to generate data according to this model with $\mu \neq \mathbf{0}$, we would be estimating $\alpha_{i_h}^\circ = \alpha_{i_h} - \mu_i$ rather than α_{i_h}

We shall now introduce some notation. Let $\alpha_h = \left(\alpha_{1_h}, \cdots, \alpha_{n_k}\right)'$, $\alpha' = \left(\alpha_1', \cdots, \alpha_{k-1}'\right)$, $\vartheta' = \left(\alpha', \theta'\right)$ and $\tau_h(\vartheta) = \left(\tau_{1_h}(\vartheta), \cdots, \tau_{n_k}(\vartheta)\right)'$, where from (6) and the identification restriction $\mu = \mathbf{0}$,

$$\tau_h(\vartheta) = \mathbf{D}_\theta \alpha_h \qquad (8)$$

Furthermore, let $\tau'(\vartheta) = \left(\tau_1'(\vartheta), \cdots, \tau_{k-1}'(\vartheta)\right)$, and $\kappa'(\vartheta) = \left(\tau'(\vartheta), \rho'(\vartheta)\right)$ where $\boldsymbol{p}(\vartheta)$ is obtained by stacking the lower diagonal elements of $\boldsymbol{P}(\vartheta)$ excluding the diagonal onto a column vector. Note that in fact ρ depends only on the covariance structure parameters θ,

As pointed out in the introduction, standard software programs such as EQS, LISREL and MPLUS estimate θ using several stages (see Jöreskog, 1994; Lee, Poon & Bentler, 1995; Muthén, 1978, 1984, 1993; Muthén, du Toit & Spisic, in press; Muthén & Satorra, 1995). First, the sample thresholds τ are estimated from the first order marginals of the contingency table. Then, the polychoric correlations ρ are estimated from second order marginals of the contingency table given the estimated sample thresholds.

Consider now the estimation of ϑ from the parameters estimated in the first two stages, $\hat{\kappa}' = \left(\hat{\tau}', \hat{\rho}'\right)$. Before estimating the model parameters in the last stage using (9),however, we must investigate its identification. Most often, when estimating ϑ from κ, θ will not be identified even if the covariance structure model $\Sigma(\theta)$ is identified. Denoting by θ^* the subset of identified parameters in

θ, a general approach to estimate the identified model parameters from κ is by minimizing $\vartheta^{*\prime} = \left(\alpha', \theta^{*\prime}\right)$

$$F_1(\vartheta) = \left(\hat{\kappa} - \kappa(\vartheta)\right)' \hat{\mathbf{W}} \left(\hat{\kappa} - \kappa(\vartheta)\right) \tag{9}$$

where .**W** is a matrix converging in probability to **W**, a non-negative definite matrix, and from (5) and (8)

$$\tau_h\left(\vartheta^*\right) = \mathbf{D}_{\theta^*}\alpha_h \qquad \mathbf{P}\left(\vartheta^*\right) = \mathbf{D}_{\theta^*}\Sigma\left(\theta^*\right)\mathbf{D}_{\theta^*} \tag{10}$$

To use this general approach we need to be able to model simultaneously the thresholds and tetrachoric/polychoric correlations. In addition, we need to be able to enforce the complex non-linear constraints (4). MPLUS (Muthén & Muthén, 1998) can be used to do the former, but not the latter. LISREL (Jöreskog & Sörbom, 1993) and EQS (Bentler, 1995) only have capabilities for modeling tetrachoric/polychoric correlations.

Letting Ξ be a consistent estimate of the asymptotic covariance matrix of κ, then, obvious choices of .**W** in (9) are **W** Ξ^{-1} (WLS: Muthén, 1978), **W** = *diag* $(\Xi)^{-1}$ (DWLS: Muthén, du Toit & Spisic, in press), and **W** = **I** (ULS: Muthén, 1993). WLS estimation has asymptotically optimal properties (i.e., minimum variance) among the class of estimators (9). However, it has been found repeatedly in simulation studies (e.g., Muthén & Kaplan, 1992; Muthén, 1993; Reboussin & Liang, 1998) that unless the model is very small and the sample size very large WLS has an unacceptable small sample behavior. Furthermore, ULS and DWLS behave well in small samples (Muthén, 1993; Muthén et al., in press), the difference between the two being negligible (Maydeu-Olivares, in press).

Suppose now that the covariance structure $\sum(\theta)$ is scale invariant. A covariance structure is scale invariant (e.g., Browne & Shapiro, 1991) if for any parameter vector θ belonging to the parameter space Θ and a diagonal matrix $\mathbf{D}_\delta$ with non-zero and distinct elements δ_i, one can find a parameter vector θ belonging to Θ such that

$$\boldsymbol{\Sigma}\left(\tilde{\theta}\right) = \mathbf{D}_\delta \boldsymbol{\Sigma}(\theta)\mathbf{D}_\delta \qquad (11)$$

Since (8) is a special case of (11), when a covariance structure $\sum(\theta)$ is scale invariant (a) one can always find a parameter vector θ satisfying $\boldsymbol{P}(\theta) = \sum(\theta)$, and (b) exactly n elements of θ will not be identified because θ must satisfy the constraint (Cudeck, 1989: p. 319)

$$\mathrm{Diag}\left(\boldsymbol{\Sigma}\left(\tilde{\theta}\right)\right) = \mathbf{I} \qquad (12)$$

Thus, when a covariance structure $\sum(\theta)$ is scale invariant one can always find a subset of identified parameters in θ, say θ^*, such that (12) is satisfied. Then, letting $\tilde{\alpha}_h := \mathbf{D}_{\theta^*}\alpha_h$ and $\tilde{\vartheta}^{*\prime} = \left(\tilde{\alpha}', \tilde{\theta}^{*\prime}\right)$,

$$\tau_h\left(\tilde{\vartheta}^*\right) = \tilde{\alpha}_h \qquad \mathbf{P}\left(\tilde{\vartheta}^*\right) = \boldsymbol{\Sigma}\left(\tilde{\theta}^*\right) \qquad (13)$$

is equivalent to (10), where $\sum(\theta^*)$ has ones along its diagonal and has the same functional form as the original covariance structure $\sum(\theta^*)$.

Thus, when $\sum(\theta)$ is scale invariant it is always possible to reparameterize ϑ^* as $\tilde{\vartheta}^*$, where the latter is greatly preferable from a computationally point of view. On the one hand, when the thresholds and polychoric correlations are parameterized as a function of $\tilde{\vartheta}^*$ one takes rid of the non-linear constraints (4). On the other hand, as there is a one to one relationship between the parameter

vector $\tilde{\alpha}$ and $\boldsymbol{T}$, and as $\boldsymbol{p}$ depends only on $\tilde{\theta}^*$, one may estimate the (reparameterized) covariance structure parameters $\tilde{\theta}^*$ from the estimated tetrachoric/polychoric correlations $\boldsymbol{p}$ only by minimizing

$$F_2\left(\tilde{\theta}^*\right) = \left(\hat{\rho} - \rho\left(\tilde{\theta}^*\right)\right)' \hat{\tilde{\mathbf{W}}}\left(\hat{\rho} - \rho\left(\tilde{\theta}^*\right)\right) \qquad (14)$$

In Appendix 1 we show, following Muthén (1978, p. 554), that when is estimated using (14) with ,*diag*, and a consistent estimate of the asymptotic covariance matrix of , one obtains the same parameter estimates for than when (9) is minimized with respect to with, d*iag* respectively. Furthermore, However, the estimates for will be the same if ULS or DWLS is employed, but not when WLS is employed. LISREL (Jöreskog & Sörbom, 1993), MPLUS (Muthén & Muthén, 1998) and EQS (Bentler, 1995) all have capabilities for estimating a covariance structure from categorical data using a sequential procedure with (14) in the last stage.

In sum, scale invariance of is a sufficient condition to estimate the parameter vector only from the sample polychoric correlations. It is a necessary condition as well.

Often times, we can turn into a correlation structure by enforcing Diag. In so doing we are fitting to the data the model on the left hand side of (11). When is not scale invariance, then the models on the left and right hand side of (11) are not equivalent and the model on the left hand side of (11) is a restrictive version of the model on the right hand side of (11). Hence, when is not scale invariant and we turn it into a correlation structure by enforcing Diag, we are actually fitting to the data at hand a different and more restrictive model than the one intended.

Because when is scale invariant considerable computational gains are obtained in performing covariance structure analysis when all the observed variables are categorical, it becomes critical in applications to be able to assess whether is scale invariant. In Appendix 2 we provide computer algebra code in Mathematica (Wolfram, 1999) that will enable researchers to determine whether is locally scale invariant.

We shall now apply this general theory to a particular class of covariance structures.

3. An application of the general theory: The common factor model

Consider the class of covariance structures implied by the common factor model,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \qquad (15)$$

Where is a diagonal matrix. We shall assume that enough restrictions have been imposed on the model so that the covariance structure is identified and that for identification purposes. The threshold and correlation structure implied by this model are by (8) and (5)

$$\tau_h(\vartheta) = \mathbf{D}_\theta \alpha_h \qquad \mathbf{P}(\vartheta) = \mathbf{D}_\theta \left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}\right)\mathbf{D}_\theta \qquad (16)$$

where diag. We shall now consider how to estimate the parameter vector from the estimated thresholds and polychoric correlations. One way of estimating any member of this class is to introduce enough restrictions in so that (16) is

identified. The identified parameters,, are then estimated from . simultaneously with using (9).

Consider now the subset of models of (15) that are scale invariant. For these models

$$\tau_h\left(\tilde{\vartheta}\right) = \tilde{\alpha}_h \qquad \boldsymbol{\Sigma}\left(\tilde{\vartheta}\right) = \tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\tilde{\boldsymbol{\Lambda}}' + \tilde{\boldsymbol{\Psi}} \tag{17}$$

is equivalent to (16) where

$$\tilde{\alpha}_h = \mathbf{D}_\theta \alpha_h \qquad \tilde{\boldsymbol{\Lambda}} = \mathbf{D}_\theta \boldsymbol{\Lambda} \qquad \tilde{\boldsymbol{\Psi}} = \mathbf{D}_\theta \boldsymbol{\Psi} \mathbf{D}_\theta \tag{18}$$

To identify (17) and fulfill (12) we may simply let

$$\tilde{\boldsymbol{\Psi}} = \mathbf{I} - \mathrm{Diag}\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\tilde{\boldsymbol{\Lambda}}'\right) \tag{19}$$

Substituting (19) in (17), we obtain

$$\tau_h\left(\tilde{\alpha}_h\right) = \tilde{\alpha}_i \qquad \mathbf{P}\left(\tilde{\theta}^*\right) = \tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\tilde{\boldsymbol{\Lambda}}' + \tilde{\boldsymbol{\Psi}} \tag{20}$$

Thus, in this case, one can estimate in the last stage of the estimation procedure simply using (14). When the number of categories *k* and the number of items *n* are large this is greatly preferable from a computational viewpoint than to estimate all the identified parameters in using (9) with (16).

We shall now consider the results of estimating a covariance structure that is not scale invariant by introducing the constraints (19) to identify the model. When we use (20) with (19) we are effectively postulating that y^* , rather than y^* , has the parametric structure . When the model is scale invariant this has no effect as y^* and y^* have the same structure. However, when the model is not scale invariant y^* and y^* have different structures and thus fitting the model to y^* rather than to y^* results in fitting a more restrictive model. Another way to put it

is to say that applying (19) with (20) implies 13 fitting to the standardized variables **z***, rather than to the unstandardized variables **y***.

Again, when is scale invariant, it is irrelevant whether one imposes this structure on **z*** or on **y***. But when it is not scale invariant, however, the covariance structures of **z*** and **y*** have different parametric forms. Thus, when is not scale invariant, imposing this structure on **z*** will always results in poorer fit that imposing the same structure on **y***. To illustrate the present discussion, consider a *n*-variate normal distribution **y*** with mean zero that have been dichotomized via a threshold relationship (1). Note that since we are considering dichotomous variables, there is only one set of thresholds. The following four covariance structures for **y*** will be considered

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi} \tag{21}$$

$$\boldsymbol{\Sigma}\left(\tilde{\boldsymbol{\theta}}\right) = \tilde{\lambda}^2\mathbf{1}\mathbf{1}' + \tilde{\boldsymbol{\Psi}} \tag{22}$$

$$\boldsymbol{\Sigma}\left(\breve{\boldsymbol{\theta}}\right) = \mathbf{1}\mathbf{1}' + \breve{\boldsymbol{\Psi}} \tag{23}$$

$$\boldsymbol{\Sigma}\left(\dot{\boldsymbol{\theta}}\right) = \dot{\lambda}^2\mathbf{1}\mathbf{1}' + \mathbf{I} \tag{24}$$

where all matrices are diagonal with elements . The covariance structures (21) and (22) correspond to the well-known one factor and tau-equivalent models, respectively. Using the computer algebra code provided in Appendix 2, one may easily verify that (21) is scale invariant, whereas (22), (23), and (24) are not scale invariant.

By (16), the threshold and correlation structures corresponding to models (21) to (24) have elements

$$\tau_i(\vartheta) = \frac{\alpha_i}{\sqrt{\lambda_i^2 + \psi_i}} \qquad \rho_{ii'}(\vartheta) = \frac{\lambda_i \lambda_{i'}}{\sqrt{\lambda_i^2 + \psi_i}\sqrt{\lambda_{i'}^2 + \psi_{i'}}} \tag{25}$$

$$\tau_i(\hat{\vartheta}) = \frac{\hat{\alpha}_i}{\sqrt{\hat{\lambda}^2 + \hat{\psi}_i}} \qquad \rho_{ii'}(\tilde{\vartheta}) = \frac{\tilde{\lambda}^2}{\sqrt{\tilde{\lambda}^2 + \tilde{\psi}_i}\sqrt{\tilde{\lambda}^2 + \tilde{\psi}_{i'}}} \tag{26}$$

$$\tau_i(\breve{\vartheta}) = \frac{\breve{\alpha}_i}{\sqrt{1 + \breve{\psi}_i}} \qquad \rho_{ii'}(\breve{\vartheta}) = \frac{1}{\sqrt{1 + \breve{\psi}_i}\sqrt{1 + \breve{\psi}_{i'}}} \tag{27}$$

$$\tau_i(\dot{\vartheta}) = \frac{\dot{\alpha}_i}{\sqrt{\dot{\lambda}^2 + 1}} \qquad \rho_{ii'}(\dot{\vartheta}) = \frac{\dot{\lambda}^2}{\dot{\lambda}^2 + 1} \tag{28}$$

After introducing suitable (if any) identification constraints, any of these threshold and correlation structures can be estimated employing (9). In Appendix 3 we provide computer algebra code in Mathematica (Wolfram, 1999) that will enable users to determine whether these threshold and correlation structures are locally identified using results of Bekker, Merckens and Wansbeek (1994).

We shall first consider the one factor model (25). Using the code in Appendix 3 we find that *n* constraints need to be introduced in this model for this structure to be identified. The constraint identifies the model. One set of identified parameters is therefore

$\vartheta^* = (\alpha_1, \cdots, \alpha_n, \lambda_1, \cdots, \lambda_n)'$ and we may rewrite (25) as

$$\tau_i(\vartheta^*) = \frac{\alpha_i}{\sqrt{\lambda_i^2 + 1}} \qquad \rho_{ii'}(\vartheta^*) = \frac{\lambda_i \lambda_{i'}}{\sqrt{\lambda_i^2 + 1}\sqrt{\lambda_{i'}^2 + 1}} \tag{29}$$

Now, because the one factor model is scale invariant using (19) and (20) we can reparameterize it as

$$\tau_i\left(\tilde{\vartheta}^*\right) = \tilde{\alpha}_i \qquad\qquad \rho_{ii'}\left(\tilde{\vartheta}^*\right) = \tilde{\lambda}_i\tilde{\lambda}_{i'} \qquad\qquad (30)$$

with *diag* The parameterization (30) is considerably more convenient than (29) because the parameters of the covariance structure can be estimated in the third stage as a correlation structure problem using (14) rather than as a threshold and correlation structure problem using (9). Furthermore, the non-linear restrictions in (30) are considerably simpler than in (29). The relationship between the parameterizations (29) and (30) is given by

$$\tilde{\alpha}_i = \frac{\alpha_i}{\sqrt{\lambda_i^2 + 1}} \qquad\qquad \tilde{\lambda}_i = \frac{\lambda_i}{\sqrt{\lambda_i^2 + 1}} \qquad\qquad (31)$$

Consider now the tau-equivalent model (26). Using computer algebra we find that just one constraint needs to be introduced in this model to identify it. The constraint identifies the model. Alternatively, the constraint also identifies the model. If we use

to identify the model, (26) becomes

$$\tau_i\left(\tilde{\vartheta}^*\right) = \frac{\tilde{\alpha}_i}{\sqrt{1+\tilde{\psi}_i}} \qquad\qquad \rho_{ii'}\left(\tilde{\vartheta}^*\right) = \frac{1}{\sqrt{1+\tilde{\psi}_i}\sqrt{1+\tilde{\psi}_{i'}}} \qquad\qquad (32)$$

which is identical to (27). But if we substitute

$$\tilde{\alpha}_i = \frac{\alpha_i}{\sqrt{1+\tilde{\psi}_i}} \qquad\qquad \tilde{\lambda}_i = \frac{1}{\sqrt{1+\tilde{\psi}_i}} \qquad\qquad (33)$$

into (30), we also see that (27) is equivalent to (30). Thus, the covariance structures (21), (22) and (23) are equivalent when only categorical data is observed. This is a remarkable result. We shall now consider the results of applying (20) with (19) to a covariance structure that is not scale invariant, such as

the tau-equivalent covariance structure (22), in order to estimate it as a correlation structure only via (14). In this case, letting **.** we would estimate a threshold and correlation structure with elements

$$\tau_i\left(\tilde{\dot{\vartheta}}^*\right) = \tilde{\dot{\alpha}}_i \qquad\qquad \rho_{ii'}\left(\tilde{\dot{\vartheta}}^*\right) = \tilde{\dot{\lambda}}^2 \qquad\qquad (34)$$

Clearly, (30) and (34) are not equivalent models. Thus, applying (20) with (19) to estimate a covariance structure that is not scale invariant from a sample correlation matrix results in estimating a different, more restrictive, model than the one intended. To what covariance structure for **y*** corresponds the threshold and correlation structure (34)? Consider the covariance structure (24). It can be readily verified by substituting

$$\tilde{\dot{\alpha}}_i = \frac{\dot{\alpha}_i}{\sqrt{\dot{\lambda}^2 + 1}} \qquad\qquad \tilde{\dot{\lambda}}^2 = \frac{\dot{\lambda}^2}{\dot{\lambda}^2 + 1} \qquad\qquad (35)$$

into (34) that (34) and (28) are equivalent, and therefore that by fitting (34) we are actually estimating the covariance structure (24).

We shall now provide a numerical example to illustrate our discussion. The covariance structures (21) to (24) will be fitted to a small binary dataset. We chose the well studied LSAT 6 dataset (Bock & Lieberman, 1970) for this example. This dataset consists of 1000 observations on 5 binary variables.

The following table summarizes the covariance structures fitted, the parameterization employed in their threshold and correlation structure, and how they were estimated in the last stage of the sequential procedure employed.

Model	Covariance structure	Parameterized as	Estimated in the third stage using
A	(21)	(29)	(9)
B	(21)	(30)	(14)
C	(22)	(32)	(9)
D	(24)	(35)	(9)
E	(24)	(34)	(14)

To estimate these models, the elements of and their asymptotic covariance matrix **,** were estimated as in Muthén (1978). Parameter estimates, their asymptotic standard errors and goodness of fit tests for the structural restrictions were obtained employing DWLS in the third stage as in Muthén, du Toit and Spisic (in press).

The parameter estimates and standard errors for these models are shown in Table 1.

Insert Table 1 about here

The so called models {A, B, C} are equivalent (they are just reparameterizatios of each other) and so are models {D, E}. The Satorra-Bentler's scaled statistic for assessing the structural restrictions imposed on the threshold and correlation structures by models {A, B, C} is $Ts = 4.741$, 5 d.f., $p = 0.448$, and for models {D, E} is $Ts = 5.269$, 9 d.f., $p = 0.810$. A nested test (Satorra and Bentler, 1999) reveals that the less restricted models {A, B, C} do not fit significantly better these data than the more restricted ones: $Tdif = 0.856$, 4 d.f., $p = 0.931$. Furthermore, one can verify in Table 1 the equivalencies among the models: Parameter estimates for

models A and B are related by (31), for models B and C by (33), and for models D and E are by (35).

4. Estimating a covariance structure model from a sample correlation matrix of continuous variables

When a covariance structure is to be estimated from a sample correlation matrix one must obtain the population correlation structure associated with the covariance structure. We saw in previous sections that there are two ways to do this: By using scaling constraints

$$\mathbf{P}(\theta) = \mathbf{D}_\theta \Sigma(\theta) \mathbf{D}_\theta \qquad \mathbf{D}_\theta = \mathrm{Diag}\left(\Sigma(\theta)\right)^{-\frac{1}{2}} \qquad (36)$$

$$\mathbf{P}(\theta) = \Sigma\left(\tilde{\theta}\right) \qquad (37)$$

were one can employ (37) if and only if is scale invariant, whereas (36) can be used to estimate a covariance structure from a sample correlation regardless of whether is scale invariant or not. The application of (37) to estimate a covariance structure that is not scale invariant results in estimating a different and more restrictive covariance structure than intended. In any case, not all the parameters in θ can be estimated, and the same number of identification constraints must be imposed if one uses (36) or (37) to estimate a scale invariant covariance structure. To identify (37) one simply needs to enforce Diag, whereas identifying (36) is more complex and we provide computer algebra code in Appendix 3 to do so.

In sum, because in general estimating a covariance structure from a sample correlation matrix (a) requires enforcing complex constraints among the covariance structure parameters, and (b) not all the parameters of the covariance

structure can be estimated, one should not estimate a covariance structure from a correlation matrix unless one is forced to do so because only the sample correlation matrix is available. When only the sample correlation matrix among the observed continuous variables is available, then estimation must proceed under multivariate normal assumptions.

One can estimate the identified subset of by minimizing a normal theory (NT) generalized least squares (GLS) discrepancy function for sample correlations (Jennrich, 1970; Browne & Shapiro, 1990). To our knowledge this discrepancy function has not been implemented in any standard software package for covariance structure analysis. Fortunately, it is not needed to employ a NT discrepancy function for sample correlations to correctly estimate a covariance structure from sample correlations. One may simply employ a NT discrepancy function for sample covariances provided (a) the degrees of freedom are correctly computed as $n\ n\ q'$ where q^* is the number of identified parameters, and (b) one imposes the constraints among the identified parameters is employed, or Diag if (37) is employed.

Both LISREL and MPLUS can be used to fit scale invariant covariance structures to a sample correlation matrix by using (37) enforcing Diag and a NT discrepancy function for sample covariances. To our knowledge, the current version of EQS can not enforce constraints Diag**I** and hence, it can not be used to correctly estimate a covariance structure from a sample correlation matrix of continuous variables. Neither LISREL, MPLUS, nor EQS can enforce the complex non-linear constraints implied by (36), and hence, these programs can not be used

to estimate a non-scale invariant covariance structure from a sample correlation matrix. To illustrate our present discussion numerically, we shall use a sample covariance matrix considered by Cudeck (1989) and originally published in Jöreskog (1978). The sample covariance matrix and its corresponding correlation matrix are given in Table 2.

Insert Table 2 about here

Consider a factor analysis covariance structure Σ with the following constraints:

Model A

$$\Lambda' = \begin{pmatrix} \lambda_1 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & \lambda_4 \end{pmatrix} \tag{38}$$

Model B

$$\Lambda' = \begin{pmatrix} \lambda_1 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & \lambda_2 \end{pmatrix} \tag{39}$$

Model B is not. Both models are identified if estimated from sample covariances. The following table summarizes the various submodels to be fitted.

Submodel	Cov. structure	Associated corr. structure obtained by	Estimated from sample
A	A	--	covariances
A_1	A	scaling constraints	correlations
A_2	A	reparameterization	correlations
B	B	---	covariances
B_1	B	scaling constraints	correlations

Estimation in all cases was performed by minimizing a maximum likelihood discrepancy function for sample covariances. The resulting parameter estimates, standard errors and goodness of fit tests are shown in Table 3.

Insert Table 3 about here

Consider first Model A. To fit this model from sample correlations, we obtain its associated correlation structure using (36). Because the covariance structure is scale invariant, when estimating it from a sample correlation matrix exactly n elements in can not be estimated. Using the methods given in Appendix 3, we find that the constraint **I** identifies the model. This is submodel A1. Because Model A is scale invariant when estimating it from sample covariances or correlations we obtain the same (a) goodness of fit, (b) parameter estimates and standard errors for scale free parameters (in this case for <) –see

Cudeck (1989). However, we obtain different parameter estimates for the elements of - in A and A1 because these parameters are not scale free. In this example, because we have both the sample covariance and correlation matrices we can obtain the same parameter estimates or using correlations than covariances if instead of fixing **I** when estimating the model from correlations, we fix these values at the values estimated using covariances. This is submodel A1

Note that the standard errors for non-scale free parameters estimated from correlations are larger. Finally, because Model A is scale invariant, we can alternatively use the reparameterization approach (37) to fit it from sample

correlations and estimate the reparameterized matrices of factor loadings (18) and uniquenesses (19).

Consider now Model B. Because it is not scale invariant, in can only be estimated from correlations using scaling constraints. Using the methods given in Appendix 3, we find that two constraints need to be introduced in the parameter vector to estimate it from sample correlations. The constraints $\theta_1 = 1, \theta_2 = 1$ identify the model. This is model B1. The goodness of fit of models B and B1 are different because model B1 is a constrained version of model B. In fact, B1 is equivalent to models A1 and A2. That is, although Models A and B are distinct covariance structures, they have equivalent associated correlation structures.

5. Conclusions

When fitting a covariance structure from a sample correlation matrix one must consider the population correlation structure associated with it under the null hypothesis This is obtained by pre and post-multiplying the covariance structure specified by the null hypothesis by a model-based diagonal matrix. That is, this diagonal matrix consists of the inverse of the square root of the diagonal of the covariance structure under consideration. As a result, in general, estimating a covariance structure from a sample correlation matrix requires estimating complicated non-linear functions of the covariance structure parameters.

However, it is well known (see for instance Cudeck, 1989) that if the covariance structure is scale invariant then one can find a reparameterization of this correlation structure that has the same functional form as the covariance

structure specified by the null hypothesis. This reparameterization approach to estimate covariance structures is greatly preferable from a computational point of view, but it is only possible with scale invariant models.

Furthermore, the goodness of fit indices obtained when estimating a covariance structure from sample correlations and from a sample covariances will be the same only if the covariance structure is scale invariant because not all the parameters of the covariance structure can be estimated from sample correlations. Hence the substantive conclusions a researcher may reach if s/he estimates a covariance structure that is NOT scale invariant from sample covariances or correlations may be different. Hence, assessing whether a covariance structure is scale invariant is critical in estimating it from a sample correlation matrix.

When all the observed variables are categorical these problems can not be avoided, as in this case one can only estimate a matrix of sample tetrachoric/polychoric correlations. Furthermore, we have shown that in this case the common practice of estimating the covariance structure parameters from a matrix of sample tetrachoric/polychoric correlations when no restrictions are imposed on the thresholds is admissible only if the covariance structure specified by the null hypothesis is scale invariant. Otherwise, one estimates a covariance structure that is more restrictive than that specified by the null hypothesis. We have also shown that to correctly estimate a covariance structure that is not scale invariant from categorical observed variables, one has to do so jointly from the sample thresholds and tetrachoric/polychoric correlations.

Proof of the equivalence of (9) and (14) for scale invariant models

When $\Sigma(\theta)$ is scaled invariant, $\kappa\left(\tilde{\vartheta}^*\right)' = \left(\tau(\tilde{\alpha})', \rho\left(\tilde{\theta}^*\right)'\right)$. Now, letting $\mathbf{e}_1 := \hat{\tau} - \tau(\tilde{\alpha})$, $\mathbf{e}_2 := \hat{\rho} - \rho\left(\tilde{\theta}^*\right)$, and partitioning $\mathbf{W}$ according to the partitioning of κ, (9) may be rewritten as

$$F_1 = \begin{pmatrix} \mathbf{e}_1' & \mathbf{e}_2' \end{pmatrix} \begin{pmatrix} \hat{\mathbf{W}}_{11} & \hat{\mathbf{W}}_{21}' \\ \hat{\mathbf{W}}_{21} & \hat{\mathbf{W}}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = \mathbf{e}_1'\hat{\mathbf{W}}_{11}\mathbf{e}_1 + 2\mathbf{e}_1'\hat{\mathbf{W}}_{21}'\mathbf{e}_2 + \mathbf{e}_2'\hat{\mathbf{W}}_{22}\mathbf{e}_2 \qquad (40)$$

Now, since there is a one-to-one relationship between and , from the first order condition for minimizing (14)

$$\frac{\partial F_1}{\partial \tilde{\alpha}'} = \frac{\partial F_1}{\partial \tau'} = 2\hat{\mathbf{W}}_{11}\mathbf{e}_1 + 2\hat{\mathbf{W}}_{21}'\mathbf{e}_2 = 0 \quad \Rightarrow \quad \mathbf{e}_1 = -\hat{\mathbf{W}}_{11}^{-1}\hat{\mathbf{W}}_{21}'\mathbf{e}_2 \qquad (41)$$

and substituting this into (40), we obtain

$$F_1 = \mathbf{e}_2'\left(\hat{\mathbf{W}}_{22} - \hat{\mathbf{W}}_{21}\hat{\mathbf{W}}_{11}\hat{\mathbf{W}}_{21}'\right)\mathbf{e}_2 = \mathbf{e}_2'\left[\hat{\mathbf{W}}^{-1}\right]_{22}\mathbf{e}_2 := F_2 \qquad (42)$$

where the last equality follows from a well-known result for the inverse of a partitioned matrix (e.g., Mardia, Kent & Bibby, p. 459).

Hence, since 1 2 , when the covariance structure parameters are estimated by minimizing *F*2 the resulting parameter estimates and their standard errors will equal those obtained had these been estimated by minimizing *F*1. If one is interested in estimating the threshold parameters after minimizing *F*2, from (41) one may use

Assessing local scale invariance using computer algebra

Assessing whether is scale invariant amounts to verifying if we can find an alternative parameter vector such that (11) is satisfied, under the additional conditions that (a) and belong to the same parameter space and (b) the elements of the diagonal matrix are non-zero and distinct elements. Most often I a non-linear function of . In that case, it is very difficult to solve the system of non-linear equations (11) unless the model is small, even with the aid of software systems capable of performing symbolic computations, such as Mathematica (Wolfram, 1999).

Note, however, that is nested within. Thus assessing scale invariance amounts to assessing whether two nested models are equivalent. To do so, we may apply a result due to Bekker et al. (1994: Section 2.8) by which, under appropriate regularity conditions, and are locally equivalent (and hence will be scale invariant) if and only if

$$\operatorname{rank}\left(\frac{\partial \operatorname{vecs}\left(\mathbf{D}_{\delta}\Sigma(\theta)\mathbf{D}_{\delta}\right)}{\partial\left(\theta', \delta'\right)}\right) = \operatorname{rank}\left(\frac{\partial \operatorname{vecs}\left(\Sigma(\theta)\right)}{\partial\theta'}\right) + n \qquad (44)$$

where *n* is the number of observed variables, and vecs(*) denotes a column vector obtained by stacking the lower triangular elements of a matrix, including the diagonal, into a column vector. Condition (44) can be very easily verified using a software package with symbolic computational capabilities, often for large models. Consider the covariance structure models A and B described in Section 4. We shall now provide some very simple Mathematica code to assess whether these models are scale invariant using (44). The code consists of four parts.

We first need the following function definitions

```
T[matrix_List] := Transpose[matrix]

L[matrix_List] := Length[matrix]

Diag[matrix_List] := Table[If[i == j, matrix[[i, j]], 0], {i, L[matrix]}, {j, L[matrix]}] (45)

VecLow[matrix_List] := Flatten[MapIndexed[Take[#1, First[#2] - 1] &, matrix]]

VecLowDiag[matrix_List] := Flatten[MapIndexed[Take[#1, First[#2]] &, matrix]]

VecDiag[matrix_List] := Table[matrix[[i, i]], {i, Length[matrix]}]
```

where VecDiag(*), VecLow(*), and VecLowDiag(*) vectorize the diagonal, below the diagonal, and below and diagonal elements of a matrix, respectively. Diag(*) simply sets the off-diagonal elements of a matrix equal to zero. The second block of the program simply constructs ***!***(*"*). For model A, this would simply be

```
n = 4;

la = {{l1,0},{l2,0},{0,l3},{0,l4}};

phi = {{1,r},{r,1}}; (46)

psi = DiagonalMatrix[Table[ToExpression["ps" <> ToString[i]], {i, n}]]

sigma = la . phi . T[la] + psi:
```

The third block of the program constructs vecs . The latter is accomplished by vectorizing and , putting them together and dropping constants and repeated parameters.

```
omega=VecLowDiag[sigma];

Print["This is the parameter vector theta"] (47)
```

theta=Cases[Union[Flatten[la],VecLowDiag[phi],VecDiag[psi]],_Symbol] Finally, the fourth block constructs and informs the user of whether is (locally) scale invariant or not by verifying where F"E denotes a basis for the nullspace of the Jacobian matrix E.

```
j =Outer[D,omega,theta];
lnu1=L[NullSpace[j]];
d=DiagonalMatrix[Table[ToExpression["d"<>ToString[i]],{i,n}]];
j2 =Outer[D,VecLowDiag[d . sigma . d],Join[theta,VecDiag[d]]]; (48)
lnu2=L[NullSpace[j2]];
```

If[lnu1 + n == lnu2,Print["The covariance structure is scale invariant"],Print["The covariance structure is NOT scale invariant Using (45), (46), (47) and (48) one may readily verify model A is scale invariant but model B is not.

Assessing local model identification using computer algebra

Following Bekker et al. (1994) a necessary and sufficient condition (under appropriate regularity conditions) for the local identification of in the parametric structure **.** is that the Jacobian matrix be of full column rank. This condition may be verified by nstructing a basis for the nullspace of E, say F , such that FE0 % 0 , and checking that F is an empty set. Whenever the model is not identified, the number of constraints we need to introduce in the parameter vector will be given by the rank of F . Furthermore, a zero column in F indicates an identified parameter.

We shall now provide some very simple Mathematica code to assess whether a threshold and correlation structure is locally identified using these results. We shall apply it to investigate the identification of the tau-equivalent covariance structure (22) for binary data. The code consists of four blocks.

The first block is simply (45). The second block constructs the threshold and correlation structure of the model of interest. In this case, it would be

In this example, the program reports that none of the parameters is identified, and that one constraint must be introduced in the model to identified. At this point, one can check whether the estimated F is actually a basis of the nullspace of E verifying that Simplify yields a zero matrix, or print F using MatrixForm[nu], which in this example yields, Finally, one can fix one of the non-identified parameters, say G = 1, and re-run the program to verify that the model is identified for any number of observed variables *n*. A word of caution. Because of the non-linear constraints (4), finding a basis for the nullspace in these models requires considerable computer resources unless the model is small.

Chapter - 3

The Model-Size Effect on Traditional and ModifiedTests of Covariance Structures

Chapter - 3
The Model-Size Effect on Traditional and ModifiedTests of Covariance Structures

3.1 Introduction.

In the practice of structural equation modeling (SEM) one can observe that an increasing number of large models are estimated; that is, models with lots of indicators and latent variables, and consequently in most cases many degrees of freedom. This may raise a number of problems. First, it is not always possible and it is often too expensive to get large sample sizes needed to estimate such big models. Second, the distribution of the large number of observed variables involved can rarely be approximated by a multivariate normal density. Third, the combination of large models, relatively small sample sizes, and non normal data appears to be accountable for the inflated Type I error rates of the traditional maximum likelihood ratio test statistic, TML, for global model fit (see, e.g., Hoogland, 1999). The apparent consequence—which can be verified from the literature—is that in applied SEM, researchers increasingly rely on alternative fit measures rather than TML. Decisions and conclusions regarding model fit are frequently based on more popular statistics and fit indexes, applying partly subjective cutoff criteria. A brief outline of the goals of our study follows.

It is argued that the effect of model size, measured by the number of degrees of freedom d (cf. Kenny & McCoach, 2003), and its interaction with sample size requires more attention in applied research, because (a) the model-size effect makes investigators more reluctant to report p values of model fit statistics

in their studies—even if of no single use—and (b) other popular statistics (e.g., the Tucker–Lewis index [TLI], and the root mean square error of approximation [RMSEA]) are affected by the inflated values of TML as well. Because relatively little is known about the effects of model size on familiar model test statistics, the first aim of our study is to quantify the impact of large model size on the finite sampling distribution of TML in SEM. In general, for the evaluation of model-size effects on model test statistics Type I error rates are of specific, although not of single importance.

Although not very obvious at first glance, a family of chi-square corrections introduced by Satorra and Bentler (1988, 1994) might be one promising approach to handle the model-size effect. Two of them are the scaled (mean-corrected) statistic, TSC, and the adjusted (mean- and variance-corrected) statistic, TAD (Satorra & Bentler, 1994, p. 407f), based on theoretical work by Bartlett (1937) and Satterthwaite (1941), respectively, and a classical paper by Box (1954). It is well known that these corrections have first and foremost been developed to make TML robust against effects of nonnormality. It should be noted, however, that Satorra and Bentler (2001) suggested (in their abstract) that their corrections might also work for small samples and large models, relative to distribution-free estimation methods, that is. In addition, the studies by Fouladi (2000) and Nevitt and Hancock (2004) provided empirical evidence that, relative to TML, these corrections might also improve small-sample performance even when the normality assumption is not violated at all. As large models need large sample sizes for the asymptotic properties of test statistics to hold (Muthén, 1993,

p. 228), it is reasonable to assume that these statistics will also perform well in large models. Unfortunately, little is known about the finite-sample behavior of TSC and TAD in large models and about the interaction of sample-size and model-size effects. Therefore, our second aim is to check whether it is beneficial (focusing on Type I error rates as well as on complete distribution functions) to favorTSC or TAD over TML for the test of large models even under conditions of multivariate normality. In this study we do not consider analyses of nonnormal data because, as a baseline, a detailed investigation of the effect of increasing d under the normality assumption is needed first. Once more, we included the Satorra–Bentler statistics in our research design, not because of their wellknown performance for the non normal case (e.g., Hu, Bentler, & Kano, 1992), but because they seem to be promising for correcting model-size effects under normality conditions as well.

Another straightforward approach to attack the problem of model size is to compute the corresponding Bartlett corrections of the three model fit statistics, TMLb, TSCb, and TADb, as proposed by Fouladi (2000) and more recently by Nevitt and Hancock (2004). Although Bartlett (1950) developed his type of corrections for exploratory factor modeling, these researchers found an acceptable performance under conditions of small sample size for general SEM as well. Because of the dependency of sample-size requirements on model size, as mentioned earlier, it is expected that these corrections might also work in large models. Because their behavior in large models is not precisely known, it is investigated whether these statistics turn out to be adequate corrections of model-

size effects. Hence, our third aim is to investigate the Type I error rates produced by TMLb, TSCb, and TADb, and to compare them to those of TML, TSC , and TAD, respectively, in large models under conditions of multivariate normality.

A less well-known correction of TML has been developed by Swain (1975). According to Browne (1982), this approach "seem[s] to result in an improvement of the approximation of the chi-squared distribution" (p. 98). With the exception of the Monte Carlo study by Fouladi (2000), to our knowledge the finite-sample behavior of this statistic is undocumented. Fouladi found a good performance of the statistic, especially for small sample sizes. For similar reasons as for the Bartlett corrections, it could be claimed that the corresponding Swain corrections TMLs , TSCs , and TADs might yield better Type I error rates compared to those of TML, TSC, and TAD. Therefore, the fourth aim of this study is to investigate the performance of the Swain corrections in large models under multivariate normality.

In summary, the purpose of our study is (a) to investigate the bias in Type I error rates produced by TML; (b) to compare the results of TML with those of TSC and TAD; (c) to evaluate the performance of TMLb, TSCb, and TADb; and (d) to check whether the behavior of TMLs , TSCs , and TADs is appropriate for testing covariance structure models with many degrees of freedom when multivariate normality assumptions hold.

Before we turn to the next section, it is emphasized that a careful investigation of TML, TSC, and TAD in large models was demanded by several researchers (e.g., Hoogland, 1999; Kenny & McCoach, 2003; Muthén, 1993, p.

228; Muthén & Satorra, 1995). To our present knowledge, no systematic Monte Carlo study of the behavior of chi-square statistics in very large models exists, although the investigation of such models "will probably result in findings that are more disappointing regarding the chi-square statistic" (Hoogland, 1999, p. 51). As indicated before, an exception is a study on some fit measures (RMSEA, TLI, and the comparative fit index [CFI]) by Kenny and McCoach (2003). Two remarks on this first investigation of the behavior of fit statistics in large models can be made. First, the study aimed at two measures (CFI and TLI) with rather subjective cutoff criteria for model fit evaluation, not at the regular chi-square statistic for overall model fit. Second, in applied research, model decision criteria for the RMSEA are mainly based on practical experience (Browne & Cudeck, 1992, p. 239), which is not undisputable: Jöreskog (2005) favored a p value for the test of close fit associated with the RMSEA of at least 0.50.

The article is structured as follows. First, the test statistics under study are defined and the corresponding asymptotic theory is presented briefly. Second, research hypotheses are developed based on findings of previous simulation studies; that is, expectations regarding the behavior of the test statistics under study are formulated. Third, based on results from a Monte Carlo research design, the expectations are tested and consequences for applied research are deduced. The practical implications of our findings are further exemplified by correcting the fit of a large structural equation model that was published recently. Finally, some limitations of this study and directions of future research are briefly mentioned.

3.2 TEST STATISTICS AND THEIR ASYMPTOTIC DISTRIBUTION

In this section, all test statistics under study are defined and the asymptotic theory underlying their distribution is summarized.

Likelihood **Ratio Statistic**

Consider ***p*** random variables **z** (***p*** x 1) with an empirical sample covariance matrix **S**(***p*** x ***p***) based on ***N*** = n + 1 independent observations, and a population model of underlying relations among these variables with covariance structure $\sum(\theta)$ (***p*** x ***p***), where $\sum$ (***t*** x 1) is the vector of independent model parameters to be estimated. If the observed variables **z** follow a multivariate normal distribution, the sample covariance matrix **S** based on independently and identically distributed observations has a Wishart distribution (Anderson, 1958). The maximization of the corresponding log-likelihood function, conditional on the sample covariance matrix S, is equivalent to minimizing the function

$$F_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] - \log|\mathbf{S}| - p, \qquad (1)$$

which is a discrepancy function as defined by Browne (1984, p. 64); log denotes the natural logarithm here. The parameter vector $\hat{\theta}$, defining the minimum of F_{ML} [**S**;. $\sum(\theta)$], contains the so-called maximum likelihood estimates of θ. Asymptotically, as N goes to infinity, the maximum likelihood estimates are normally distributed with expectation vector E. ($\hat{\theta}$)= θ, and asymptotic covariance matrix acov($\hat{\theta}$,$\hat{\theta}$)=$\boldsymbol{\Gamma}^{1}(\theta)$, the inverted Fisher information matrix of order (***t*** x ***t***), which can be estimated (cf. Bollen, 1989, p. 109), yielding estimates

of the standard errors of the t parameter estimates as well as estimated covariances between those parameter estimates.

Let $\sum(\boldsymbol{p} \times \boldsymbol{p})$ denote the population covariance matrix of the $\boldsymbol{p}$ observed variables $\mathbf{z}$, $\sum(\theta \mathrm{j})$ the population covariance matrix implied by a postulated model $\boldsymbol{M}_{\mathrm{j}}$, and let c be an "irrelevant constant" (Bollen, 1989, p. 263). One can then test the null hypothesis $\mathrm{H}_0 : \sum = \sum(\theta \mathrm{o})$; that is, that the postulated model holds, with the corresponding log-likelihood function, evaluated at $\theta \mathrm{o} = \hat{\theta}\,\mathrm{o}$,

$$\log L_0 = \log L[\boldsymbol{\Sigma}(\hat{\theta}_0); \mathbf{S}] = -\frac{n}{2}\{\log|\boldsymbol{\Sigma}(\hat{\theta}_0)| + \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\hat{\theta}_0)]\} + \log c, \quad (2)$$

$$\log L_1 = \log L(\boldsymbol{\Omega}; \mathbf{S}) = -\frac{n}{2}[\log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{S}^{-1})] + \log c \quad (3)$$

$$= -\frac{n}{2}(\log|\mathbf{S}| + p) + \log c$$

It can then be shown that under H0, the distribution of the likelihood ratio statistic, defined as

$$T_{ML} \equiv -2\log\frac{L_0}{L_1} = -2\log\frac{L[\boldsymbol{\Sigma}(\hat{\theta}_0); \mathbf{S}]}{L(\boldsymbol{\Omega}; \mathbf{S})} = nF_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\theta}_0)], \quad (4)$$

converges with increasing sample size 1 to a chi-square distribution with 2-t degrees of freedom (Wilks, 1938); the likelihood criterionoe $L_0=L_1$ in Equation 4 was introduced by Neyman and Pearson (1928).From Equations 1 and 4 it follows that the likelihood ratio test statistic, T_{ML}, is by definition n times the minimum of the maximum likelihood discrepancy function evaluated . Hence, the likelihood ratio test statistic can beused to test whether the proposed model is implausible at a given level of significance. In practice, the behavior of this statistic depends, of course, on its robustness against violations of underlying assumptions (independent observations, multivariate normality with covariance structure , and a large sample size, mainly).

Satorra–Bentler Statistics

Because non normal data are very common in practice, Satorra and Bentler (1988, 1994) introduced two corrections to a family of model test statistics, aimed to yield distributional behavior that more closely follows the chi-square reference distribution that is used in structural equation model testing. Relative to distribution-free methods, these statistics can be useful when the sample size is small or the estimated model is large (Satorra & Bentler, 2001, p. 507). The corrections can, in principle, be applied to a family of test statistics, including the normal theory weighted least square model test statistic, TWLSN, as it is used in the LISREL program (see Jöreskog, Sörbom, Du Toit, & Du Toit, 2001, Appendix A). In this study, we only apply it to TML.

The mean-corrected, scaled statistic (Satorra & Bentler, 1988, 1994, p. 407) is defined as

$$T_{SC} \equiv \frac{d}{\mathrm{tr}(\mathbf{A})} T_{ML}, \qquad (5)$$

where matrix A is a slightly complicated function of a matrix of first-order derivatives of the ML-discrepancy function to the parameters to be estimated and an estimate of the asymptotic covariance matrix of sample covariances (cf. Muthén, 2004, Equation 105). If the distribution of z is elliptical, the scaling factor d=tr.A in Equation 5 provides an estimate of the common relative kurtosis of z (Satorra & Bentler, 1994, p. 407), which implies a correction for non normality.

As usual, the test statistic TSC is evaluated as having (approximately) a chisquare distribution with C 1/=2 t degrees of freedom. For certain distributions

of the observed variables, for example, elliptical ones, the asymptotic distribution of TSC is exactly chi-square with d degrees of freedom. In principle, however, the correction of TML involves a scaling to the correct mean, so that for general distributions asymptotically the first moment of the distribution of TSC is matched to the number of degrees of freedom d. Under conditions of multivariate normality, TSC has asymptotically an exact chi-square distribution with d degrees of freedom, because a multivariate normal density is also elliptical.

Furthermore, Satorra and Bentler (1988, 1994, p. 408) used a procedure developed by Satterthwaite (1941, 1946) to correct not only for the mean but for the variance of TML as well. This is possible by an adjustment of the number of degrees of freedom to d0, which is the integer closest to a function of the matrix A (cf. Muthén, 2004, Equation 110): by definition

$$d' = \mathtt{int}\left\{\frac{[\mathrm{tr}(\mathbf{A})]^2}{\mathrm{tr}(\mathbf{A}^2)}\right\}. \qquad (6)$$

$$T_{AD} \equiv \frac{d'}{\mathrm{tr}(\mathbf{A})}\, T_{ML}, \qquad (7)$$

It should be noted that the value of d0 may vary from sample to sample. Substituting d0 for d in Equation 5, we get (cf. Muthén, 2004, Equation 108): which is the adjusted chi-square test statistic; adjusted for mean and variance that is. Again, for general distributions of observed variables, TAD has asymptotically not an exact chi-square distribution with d0 degrees of freedom, but it matches the first- and second-order moment of that distribution (Satorra & Bentler, 1994, p. 408). For multivariate normal observations, TAD has asymptotically an exact chi-square distribution with d0 degrees of freedom.

It should be stressed that if distributional assumptions or conditions for asymptotic robustness hold, both corrections of TML discussed in this section are "automatically inactive (asymptotically)" (Satorra & Bentler, 1994, p. 414). Notice, however, the adverb in parentheses: asymptotically. It has to be reemphasized, that TML also follows a chi-square distribution only asymptotically

.Bartlett-Corrected Statistics

For exploratory factor analysis models (more specifically, for principal components models) Bartlett (1950, 1954) developed a correction of the chi-square test statistic for small sample sizes. In general, Bartlett's correction consists of multiplying , where oe is the likelihood ratio criterion of Neyman and Pearson (1928), by a scale factor that results in a statistic having the same moments as ¦2, ignoring quantities of order n2 (cf. Lawley, 1956). As pointed out by Lawley (1956), this scaling device was first employed by Bartlett (1937).

From Equation 9, it can be seen that Bartlett's correction for unrestricted factor models is a function of the number of latent variables k, the number of observed variables p, and the sample size N D n C 1. Fouladi (2000) and Nevitt and Hancock (2004) studied the Bartlett correction for the analysis of general structural equation models, and applied it to the three model test statistics discussed so far, TML, TSC, and TAD. The corresponding Bartlett corrections for these statistics are defined as respectively,

$$T_{MLb} \equiv b\,T_{ML}\ ,\ T_{SCb} \equiv b\,T_{SC}\ ,\ \text{and}\ T_{ADb} \equiv b\,T_{AD}, \qquad (8)$$

where

$$b = 1 - \frac{4k + 2p + 5}{6n}. \qquad (9)$$

It follows from Equations 8 and 9 that asymptotically the distribution of the Bartlett-corrected statistics matches the asymptotic distributions of TML, TSC, and TAD, respectively. The specific form of Equation 9 was derived by Bartlett (1950, Equation 3) from expansion of a moment generating function. Independently, Box (1949) derived approximations of chi-square statistics for tests on correlation matrices identical to those of Bartlett.

Swain-Corrected Statistics

As we have emphasized, the Bartlett correction in Equation 9 is the appropriate small-sample correction for exploratory or unrestricted factor models only. For general covariance structure models, Bartlett's correction is strictly speaking not appropriate. In fact, for each class of models a specific multiplier or correction factor would be needed. Because this is quite troublesome for applied researchers, Swain (1975) developed four small-sample corrections of TML for general covariance structure models. We only study the one that seemed most promising among those four; see also Browne (1982, p. 98), who claimed that Swain used "heuristic arguments" in proposing these correction factors. It should be noted in advance that Swain (1975) is very cautious about the applicability of the corrections he proposed: "For any particular model the worth of the forms suggested [correction factors of the form 1 k1=n C O.n 2/, where k1 is a

function of p and d] would, of course, have to be carefully evaluated before routine application" (p. 78).

From their basic derivations it is clear that both Bartlett and Swain corrections should be considered as multiplying or scale factors of nFMLOES;†. O™0/ , not as multipliers of just the discrepancy function FMLOES;†. O™0/ . Hence, it would be improper to suggest that these corrections can or should be interpreted as a modification of just the sample size.

For the special case of maximum likelihood estimation of structural equation models that are invariant under a constant scaling factor (cf. Browne, 1982, p. 77), the most promising small-sample correction of TML introduced by Swain (1975) is defined as

$$s = 1 - \frac{p(2p^2 + 3p - 1) - q(2q^2 + 3q - 1)}{12dn}, \qquad (10)$$

where

$$q = \frac{\sqrt{1 + 4p(p+1) - 8d} - 1}{2}, \qquad (11)$$

p is the number of observed variables, d is the number of degrees of freedom, is the sample size, as before. Equations 10 and 11 correspond to Swain's (1975) Equations 4.14 and 4.10. The Swain corrections for the three test statistics TML, TSC, and TAD are now, respectively, defined as

$$T_{MLs} \equiv s\,T_{ML}\ , T_{SCs} \equiv s\,T_{SC}\ ,\ \text{and}\ T_{ADs} \equiv s\,T_{AD}. \qquad (12)$$

From Equation 10 it can be seen that Swain's correction is a function of p, d, and N. Because, Equations 10 and 11 can also be written as a function of t instead of d, along with p and N, of course (cf. Browne, 1982, p. 98).

It follows from Equations 10 and 12 that asymptotically the distributions of the Swain-corrected statistics match those of TML, TSC, and TAD, respectively

3.3 EXPECTATIONS OF FINITE SAMPLE BEHAVIOR

In this section we discuss the expected finite sample performance of the nine statistics for global model fit in large models, TML, TSC, TAD, TMLb, TSCb, TADb, TMLs , TSCs , and TADs , as defined previously. Statistical theory does not yield clear guidelines as to the choice among these statistics, nor does it help unequivocally to come up with proper, theory-based expectations about the issue under investigation (cf. Bentler & Yuan, 1999). In our case, the design of the study has two main factors, model size and sample size: The number of latent variables in the factor models ranges from 4 to 16, with three indicators for each latent variable, and the sample sizes are 200, 400, and 800 (details of the design are reported in the next section). In general it can be expected that the behavior of the model test statistics will improve with increasing sample size (consistent estimators, the functioning of asymptotic theory) for any given model size.

Generally, it is also expected that the statistics will show improved behavior with decreasing model size for a given sample size. There exists empirical evidence and arguments for this claim. First, the results of a meta-analysis by Hoogland (1999, section 3.3) show that the performance of the chi-

square model statistics improves with a decreasing number of degrees of freedom d. Second, there are several rules of thumb in the literature indicating that one might need a specific minimal number of observations for each observed variable or for each model parameter to be estimated. Such recommendations suggest that if the number of observed or latent variables increases, more observations are needed to obtain proper estimates. As to the comparison of the test statistics under study, statistical theory is not providing solid predictions for their finite sample behavior, but in most cases it is possible to contrive expectations about the results of our investigations from the findings of previous simulation studies.

Likelihood Ratio Statistic

Under conditions of multivariate normality, for test statistic TML Hoogland (1999) found a trend to an overrejection of true models for $N < 400$, and this tendency increased as models got larger. This finding is supported by other simulation studies with various designs (Curran, Bollen, Paxton, Kirby, & Chen, 2002; Hau & Marsh, 2004; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998). We therefore expect that the empirical rejection rates will be inflated more or less seriously for very large models.

Scaled Satorra–Bentler Statistic

The studies by Hu, Bentler, and Kano (1992), Curran, West, and Finch (1996), Bentler and Yuan (1999), Hoogland (1999), Nevitt and Hancock (2001), and Hau and Marsh (2004) revealed that the test statistic TSC produces even

higher rejection rates than TML when multivariate normal variables are analyzed, and this liberal tendency increased with model size as well. Therefore, we expect that TSC will perform worse than TML in large models under conditions of normality. The explanation for this expected tendency could very well be that TSC requires the estimation of the asymptotic covariance matrix of sample covariances, which involves estimation of fourth-order moments and the computation of the inverse of often huge matrices.

Adjusted Satorra–Bentler Statistic

There is not a great deal of information about the finite sample behavior of TAD in the literature. In a recent Monte Carlo investigation, Asparouhov (2005) found the adjusted chi-square statistic to have excellent Type I error rates compared to TML and TSC. Fouladi (2000) conducted an extensive simulation study with 12 different test statistics and found TAD to outperform all other statistics with respect to Type I error rate "under more general non normal distributional conditions" (p. 400; cf. p. 371, Table 1). She concluded that TAD "shows the most rapid convergence to the nominal level and as such can be used with smaller samples than the other procedures" (p. 401). We therefore expect that TAD will outperform TML and TSC in large models.

Bartlett-Corrected Statistics

Fouladi (1999, 2000) and Nevitt and Hancock (2004) examined the performance of Bartlett corrections in the context of SEM. The results of Nevitt

and Hancock, in particular, indicate that TMLb, TSCb, and TADb tend to underestimate the nominal levels when N decreases and when d increases. Based on this finding, it is reasonable to expect that the Bartlett corrections will clearly underestimate the nominal error levels, when the model to be analyzed is larger than the models studied by Nevitt and Hancock (2004), which ranged between d D 85 and d D 196.

Swain-Corrected Statistics

To our knowledge, the only study on the Swain correction is the Monte Carlo investigation by Fouladi (2000). For the analysis of covariance structures, she found that “the normal theory procedures with the best small sample Type I error control under conditions of extremely mild distributional non normalitywere the 0-factor Bartlett rescaling or Swain rescaling of the standard ML covariance structure analysis test statistic" (p. 400). Unfortunately, she only investigated very small models with no more than 12 variables. However, as discussed earlier in the introductory section, it seems legitimate to expect an improved performance of the Swain statistics compared to TML in large models because of its favorable small-sample properties.

Summary

In summary, it is expected that TAD will perform better than TML, and that TML will be more accurate than TSC for large models under conditions of multivariate normality. We do not have much information about the Bartlett and

the Swain statistics, but it seems reasonable to expect an improved performance compared to TML when the number of degrees of freedom increases.

Although we formulated expectations based on empirical findings from the literature mainly, our study has a partly explorative character. Where appropriate, published results are revalidated by our investigations, but we seek to elaborate and to generalize them to large structural equation models.

3.4 MONTE CARLO DESIGN

Sample Size Conditions

Sample sizes of 200, 400, and 800 are used. It can be problematic to investigate sample sizes of $N < 200$ because it is well known that estimates of parameters and standard errors may be biased seriously. Also, non convergence problems and Heywood cases are more likely to occur for such small sample sizes (Boomsma, 1982, pp. 171, 1985; Boomsma & Hoogland, 2001). In practice, getting more observations than 800 is not always possible or too expensive.

Population Models and Model Size

Most Monte Carlo studies reported in the literature examined very small population models; see, for example, Asparouhov (2005) and Fouladi (2000). As for the factor models in Hoogland's (1999) meta-analysis, d ranged from 2 to 98. For our study, it was decided to restrict the population models to confirmatory factor analysis (CFA) models, because in practice these measurement models are most widely applied.

In general, a factor model without an intercept term is defined as DŸ•, where is a vector of observed variables, *f* . matrix of factor loadings on k common factors Ÿ1; Ÿ2; : : : ; Ÿk, and • .p_1/ a vector with unique scores (measurement error), where uncorrelated with Ÿ. Under the usual assumptions, the population covariance matrix of z has the form where is a diagonal matrix with unique score or error variances.

To study a variety of model sizes, the number of factors k was set at 4, 6, 8, 10, 12, 14, and 16. Each factor has three indicators, so the number of observed variables p ranges from 12 to 48. To achieve identifiable models, the variance of each latent construct was fixed to the value of one. Furthermore, the population factor loadings were set to 0.70 and the error variance to 0.51 for each indicator. The correlation between each pair of factors was set to 0.30. Table 1 gives an overview of characteristics of the seven factor models.

Number of Replications

A total number of NR D 1,200 replications was used. Although 300 replications would have been a "reasonable trade off between precision, and the amount of information to be handled" (Hoogland, 1999, p. 59), it was decided to use four times as many replications to lower the standard error of percentages presented in Tables 2, 3, and 4 (see next section). For example, under the null hypothesis that the nominal value of a 5% significance level holds, the standard error of the percentages reported in the cells of these tables equals 0.629%, where it would have been twice as large if only 300 replications had been used.

TABLE 1
Overview of Factor Models of the Monte Carlo Design and Seed Values for Data Generation

					Seed		
k	p	p^*	t	d	$N = 200$	$N = 400$	$N = 800$
4	12	78	30	48	77703570	49330350	71578326
6	18	171	51	120	83444508	39023988	68738111
8	24	300	76	224	16159776	44724671	97116941
10	30	465	105	360	71034416	06466931	85864123
12	36	666	138	528	56460497	36267030	98682926
14	42	903	175	728	64459199	07380304	07013316
16	48	1176	216	960	48795874	79583898	23965379

Note. k is the number of factors; $p = 3k$ the number of observed variables; $p^* = p(p+1)/2$ the number of independent elements of S; t the number of parameters to be estimated; $d = p^* - t$ the number of degrees of freedom.

Data Generation and Model Estimation

Multinormal variables were generated to isolate the effect of model size (and sample size) on the test statistics, and to set a normal baseline for comparison with non normal data in future research. The population covariance matrix of these normal variables is defined by the population factor structure of the models under study: †.™j /, j D 1; 2; : : : ; 7. Both the generation of the sample data and the estimation of the models was performed using the Mplus software program (Version 3.11; Muthén & Muthén, 2004). The seed values for the pseudo-random draws of samples from the multivariate normal population distributions for each cell in the design are listed in Table 1. The starting values for the model parameter estimates were fixed at their population values. The factor models were estimated using the primary estimation setting of maximum likelihood (ML) in Mplus. For

the mean-adjusted and mean- and variance-adjusted estimation of the chi-square statistic, the estimation option in Mplus was MLM and MLMV, respectively, which are both maximum likelihood procedures. For the statistical analyses of the generated model estimates, R software (Version 2.1.1) was used (see, e.g., Venables & Smith, 2005).

Statistics

The sampling distributions of the nine test statistics based on the 1,200 replications were observed. First, the empirical rejection rates on the 5% Type I error level were inspected. A tolerable rejection rate is defined here as one that falls in the two-sided 99% adjusted Wald confidence interval estimate, calculated as [3.5, 6.8]; see Agresti and Coull (1998). If the observed rejection rate falls outside this interval, it is concluded that the population rejection rate differs from 0.05; that is, rejecting the null hypothesis that the population rejection rate equals 0.05, using a 1% significance level. A 99% interval estimate was chosen because of the large number of replications, hence slightly reducing the power of the test compared to a 95% interval estimate.

Second, by means of a one-sample Kolmogorov–Smirnov test (e.g., Birnbaum, 1952) it was tested at a 1% significance level whether the empirical sampling distributions of the fit statistics follow the proper theoretical chi-square distribution. Because the value of the number of degrees of freedom for ADbased test statistics varies over sample covariance matrices, the rounded mean value over 1,200 replications was used as the number of degrees of freedom of the theoretical

chi-square distribution. In Tables 2 through 7, this rounded mean value is shown in brackets in column 12; in all cases it was equal to the median value of d0. In addition, selected PP and QQ plots (percentile-percentile and quantile-quantile plots), were used to illustrate the findings, so as to provide a visual reply to the question: How do the deviations from the theoretical chisquare distributions look?

Information about the discrepancies between empirical and theoretical distributions of test statistics, by means of both Kolmogorov–Smirnov tests and PP and QQ plots, is reported here for two reasons. First, 5% Type I error rates are quite arbitrary; sometimes 1% or 10% significance levels might be preferred. Second, in applied research p values of estimated model fit statistics are reported quite often, especially if in favor of the postulated model. If we had confined ourselves to rejection rate behavior at a 5% significance level, not only would it be difficult to generalize results to other significance levels, but also, and more important, no information about the empirical distribution function of the statistics as compared to the theoretical chi-square distribution would have
been obtained.

In the statistical analyses, all 1,200 replications were used for all cells in the design, because no convergence problems and no improper solutions occurred in model estimation.

3.5 FINDINGS AND RECOMMENDATIONS

In this section, we first focus on the empirical rejection rates of the nine test statistics for model fit and compare them with the rejection rates predicted by

asymptotic theory. Second, the sampling distributions of the test statistics are compared to the theoretical chi-square distributions by means of a one-sample Kolmogorov–Smirnov test. Third, the findings are further visualized by means of PP and QQ plots of the empirical sampling distributions of the test statistics. Finally, based on the results of these analyses, recommendations are formulated for the use of appropriate model test statistics in applied research when large models are at stake. In addition, the implications of our findings are briefly illustrated by correcting the fit of a recently published applied model.

Type I Error Rates

The empirical rejection rates were computed across the 1,200 replications. The differences of these rejection rates to the nominal 5% value are summarized in Table 2 (N D 200), Table 3 (N D 400), and Table 4 (N D 800). Values larger than zero indicate that the population model is rejected too frequently, whereas values smaller than zero indicate that the corresponding statistic is too conservative. The boldfaced numbers in these tables indicate acceptable rejection rates, for nominal ’ D 0:05 defined as O’ 2 OE0:035; 0:068 , implying that acceptable difference rates in the tables are within the range 1:5%;C1:8%.

Likelihood ratio statistic. The quantile bias of this statistic reduces with increasing sample size and decreasing model size. It can be seen that TML performs extremely badly. In fact, the rejection rate is not acceptable for all model sizes for a sample size of N D 200 and N D 400. This latter finding is in line with research findings of Boomsma (1983, Table 4.4.16, Model 4CM), who analyzed a

very similar model. The amount of this bias is considerable: For the largest model with d D 960 and N D 200 the progressive bias is 70.7%. Furthermore, the performance is not even acceptable for N D 800 when models with six or more factors are analyzed.

As a consequence of these findings, it is not recommendable to employ TML for the test of large models. Although the effect of increasing degrees of freedom has been reported frequently, the amount of the bias detected here is quite alarming. The effect of increasing degrees of freedom seems to be comparable to the effect of testing models with non normal variables. Curran et al. (1996), for example, reported empirical rejection rates of 48% for the nominal 5% Type I error rate when severely non normal variables (univariate kurtoses of 21.0 and skewnesses of 3.0) were analyzed (Curran et al., 1996, p. 22, Table 1). The rejection rate bias in our study is similar to the bias reported by these authors.

Therefore, one could argue that, in both theoretical and applied research, theissue of model size should deserve similar attention as the robustness against non normality.

Scaled Satorra–Bentler statistic. Like for TML, the finite sample bias of the test statistic TSC reduces with increasing sample size and decreasing model size. As expected, and therefore consistent with the results of simulation studies mentioned earlier, the performance of TSC is slightly worse compared to that of TML. For nearly all investigated sample sizes, the rejection rates are not acceptable. For N D 200 and 16 factors, the bias in the empirical rejection rates is

76.4%. It follows that the use of TSC is no option for the evaluation of large models.

TABLE 2
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 200$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)[a]	$N{:}t$
4	3.2	3.8	**1.1**	**.3**	**1.4**	**−1.0**	**1.4**	2.0	**−.2**	48 (36)	6.7
6	4.9	6.3	**−.6**	**−.8**	**−.5**	−3.5	**.4**	**1.2**	−3.1	120 (69)	3.9
8	9.7	13.2	**−.5**	−1.7	**−.7**	−4.6	**.8**	2.7	−3.5	224 (98)	2.6
10	20.3	24.9	**−.5**	−2.9	−1.7	−4.7	**.8**	3.2	−4.4	360 (120)	1.9
12	33.3	38.9	**.8**	−3.3	−2.4	−4.9	2.5	4.6	−4.7	528 (136)	1.4
14	50.9	57.1	**1.2**	−3.8	−3.4	−5.0	2.8	4.3	−5.0	728 (149)	1.1
16	70.7	76.4	4.2	−4.3	−4.0	−5.0	3.2	6.9	−5.0	960 (158)	.9

[a] $\overline{d}'$ denotes the rounded mean of d' for T_{AD}, T_{ADb}, and T_{ADs} over 1,200 replications.

Note. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

TABLE 3
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 400$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	2.6	3.1	**1.6**	**1.2**	**1.7**		**1.5**	2.0	**.7**	48 (41)	13.3
6	3.1	3.8	**.7**	**.5**	**1.1**	−1.6	**1.3**	1.9	**−1.1**	120 (88)	7.8
8	3.6	4.5	**−1.5**	−1.8	**−1.0**	−3.6	**−1.3**	**.3**	−3.2	224 (136)	5.3
10	6.5	8.3	**−.9**	**−1.1**	**−.7**	−4.0		**1.3**	−3.3	360 (179)	3.8
12	11.4	14.3	**−1.0**	−2.0	**−1.1**	−4.8	**.2**	**1.3**	−4.6	528 (215)	2.9
14	21.0	22.0	−1.9	−2.8	−2.2	−5.0	**1.4**	2.9	−4.7	728 (245)	2.3
16	26.0	29.7	−1.7	−3.4	−2.8	−5.0	**.8**	2.1	−4.6	960 (268)	1.9

Note. Blank cell indicates that the empirical error rate equals the nominal rate of 5%. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

TABLE 4
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 800$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	**1.4**	**1.7**	**1.1**	**1.0**	**1.3**	**.7**	**1.1**	**1.3**	**.7**	48 (44)	26.7
6	**2.2**	**2.7**	**.7**	**.6**	**1.0**	**−.6**	**1.2**	**1.6**	**−.4**	120 (101)	15.7
8	**3.1**	**3.0**	**.8**	**.8**	**1.3**	**−1.5**	**1.7**	2.1	**−.5**	224 (169)	10.5
10	1.9	2.6	**−1.1**	**−1.0**	**−.7**	−3.3	**−.2**	**−.1**	−2.6	360 (238)	7.6
12	5.6	6.1	**−1.1**	**−1.0**	**−.8**	−3.6	**.7**	**1.8**	−2.9	528 (305)	5.8
14	5.7	6.6	−1.7	−1.8	−1.6	−4.4	**−.1**	**.3**	−3.8	728 (365)	4.6
16	8.8	10.9	−2.1	−2.3	−1.7	−4.7	**−.1**	**.8**	−4.2	960 (418)	3.7

Note. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

Adjusted Satorra–Bentler statistic. For TAD with N D 200, there is a slight tendency of a reduced finite sample bias when model size decreases, but this tendency is much weaker compared to that of TML and TSC. For N D 400 and N D 800, TAD slightly underestimates nominal Type I error levels when the model size increases. Overall, however, the results indicate that TAD clearly outperforms TML and TSC for all models under study. The rejection rates on the 5% error level are nearly perfect for N D 200 and models with up to 14 factors. Therefore, our study revalidates the finding of Fouladi (2000) that test statistic TAD has excellent Type I error control. The reason for the good performance of TAD seems to be Satterthwaite's (1941, 1946) variance correction, which adjust the tail of the distribution of TML adequately.

In general, our expectations with respect to the behavior of the mean- and variance-adjusted test statistic TAD are not refuted. Recall that Fouladi (2000)found that TAD outperforms 12 other statistics with respect to Type I error control under various distributional conditions and for different models. Therefore, TAD seems to be relatively robust against model size, small sample size, and nonnormality. Nevitt and Hancock (2004) seem to be disinclined to recommend this statistic, because it slightly underestimates the nominal Type I error rates when non normal variables are analyzed. Their conclusions challenge those of Fouladi (2000); more research on this issue is therefore necessary. Nevertheless, after inspection of the empirical rejection rates, it seems legitimate to use TAD with approximately normal data, but a more final judgment will be postponed after inspection of the Kolmogorov–Smirnov test results.

Bartlett-corrected statistics. All Bartlett statistics underestimate the nominal rejection rates with increasing model size. Where most statistics are progressive (i.e., the null hypothesis is rejected too often, or the rejection rates are too high) for N D 200, the Bartlett corrections show a conservative trend (i.e., the null hypothesis is "conserved" too often, the rejection rates are too low). This is consistent with our expectation based on the results of Nevitt and Hancock (2004). Compared to TAD, the statistics TMLb, TSCb, and TADb are slightly more influenced by model size. Interestingly, TSCb performs better than TMLb. It seems that the progressive tendency of TSC dominates for smaller model sizes, whereas a general conservative effect of the Bartlett corrections dominates when the models get larger. Based on the empirical rejection rate performance only, we are slightly hesitant to recommend the use of Bartlett statistics, because these statistics are too conservative and do not reveal an adequate Type I error control, at least not for large models and small sample sizes.

Swain-corrected statistics. The results indicate that TMLs is less affected by model size compared to TMLb. The statistic TMLs has appropriate rejection rates for N D 200 up to 10 factors. Compared to all other statistics, TMLs is less influenced by the model-size effect, especially when the sample size is 400 or 800. TSCs performs equally well compared to TSCb. TADs is clearly too conservative. Thus, it seems legitimate to use TMLs in applied research, but again, a more final judgment will be formulated after looking at the results of the Kolmogorov–Smirnov test.

Intermediate conclusion. To summarize the results presented so far, we conclude that (a) TMLs , (b) TAD, and (c) TSCs or TSCb in that order yield the best 5% Type I error control in large models.

Kolmogorov–Smirnov Tests

To check whether the empirical sampling distributions of the test statistics, FNR.x/, deviate significantly from their reference chi-square distribution,F with d degrees of freedom, the one-sample Kolmogorov–Smirnov test statistic DNR supxOE was computed. The DNR values are presented in Table 5 (N D 200), Table 6 (N = 400), and Table 7 (N = 800). In the evaluation of test results we applied a two-sided 1% significance level. In our case, with NR D 1,200 replications, the critical value of the DNR statistic at that 1% level equals 1:63=p1,200 = 0:047 (Massey, 1951). Nonsignificant DNR values, indicating closeness of fit, are boldfaced in the tables.

For the smallest sample size N D 200, TMLs clearly outperforms all other statistics for large models. Although significant deviations for the larger models are reported, the relatively good performance of TMLs compared to the other statistics under study is obvious. The statistic TAD does not perform well, although it produced Type I error rates close to those of TMLs . When the sample size increases to N = 400, TSCb is the second best statistic. For N = 800, TMLs and TSCs are the best performing statistics regarding their expected distributional match.

TABLE 5
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 200$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.087	.110	.139	**.022**	**.025**	.070	**.029**	**.044**	.085	48 (36)	6.7
6	.138	.167	.203	.060	**.037**	.078	**.013**	**.043**	.111	120 (69)	3.9
8	.253	.295	.292	.068	**.027**	.100	.054	.097	.151	224 (98)	2.6
10	.368	.414	.367	.133	.076	.151	.057	.116	.178	360 (120)	1.9
12	.482	.528	.443	.195	.141	.186	.060	.124	.213	528 (136)	1.4
14	.626	.668	.516	.284	.205	.275	.099	.148	.230	728 (149)	1.1
16	.761	.800	.598	.362	.283	.301	.104	.189	.264	960 (158)	.9

Note. The critical value of $D_{1,200}$ at a two-sided 1% significance level equals 0.047. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.

TABLE 6
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 400$ ($NR = 1{,}200$)

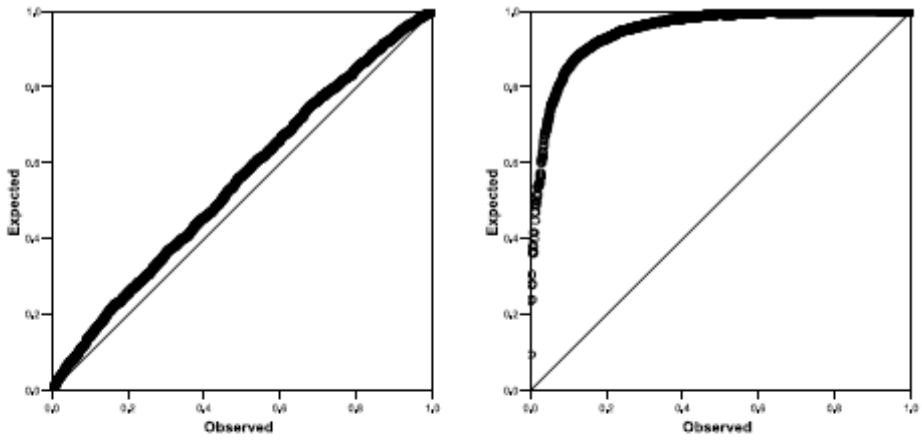
k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.084	.089	.102	**.033**	**.044**	.050	**.044**	.054	.060	48 (41)	13.3
6	.086	.102	.092	**.038**	**.030**	**.033**	**.021**	**.037**	**.042**	120 (88)	7.8
8	.145	.169	.176	**.031**	**.016**	.076	**.038**	.063	.093	224 (136)	5.3
10	.186	.211	.212	.070	**.044**	.105	**.036**	.059	.109	360 (179)	3.8
12	.260	.292	.291	.103	.070	.121	**.034**	.065	.151	528 (215)	2.9
14	.351	.385	.332	.118	.085	.164	.055	.092	.157	728 (245)	2.3
16	.428	.463	.399	.184	.138	.199	**.047**	.093	.190	960 (268)	1.9

Note. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.

TABLE 7
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 800$ ($NR = 1{,}200$)

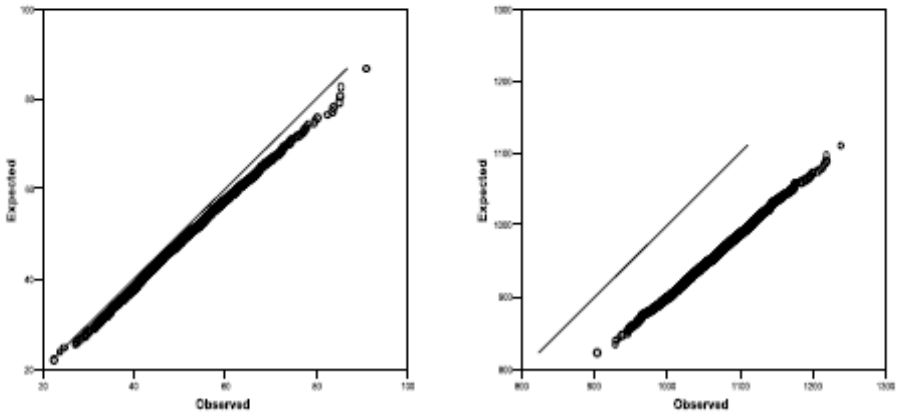
k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.048	.055	.074	**.030**	**.025**	**.043**	**.025**	**.029**	.048	48 (44)	26.7
6	**.044**	**.047**	.064	**.026**	**.023**	**.031**	**.020**	**.023**	**.039**	120 (101)	15.7
8	.096	.104	.109	**.018**	**.023**	**.047**	**.037**	**.046**	.061	224 (169)	10.5
10	.087	.096	.126	.062	.053	.072	**.023**	**.022**	.074	360 (238)	7.6
12	.135	.157	.159	.063	.054	.073	**.024**	**.040**	.086	528 (305)	5.8
14	.175	.192	.208	.072	.055	.109	**.027**	**.044**	.108	728 (365)	4.6
16	.235	.257	.268	.090	.065	.130	**.037**	.056	.143	960 (418)	3.7

Note. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.



(a) $d = 48$ (b) $d = 960$

FIGURE 1 PP plots for T_{ML} ($N = 200$; $NR = 1{,}200$).



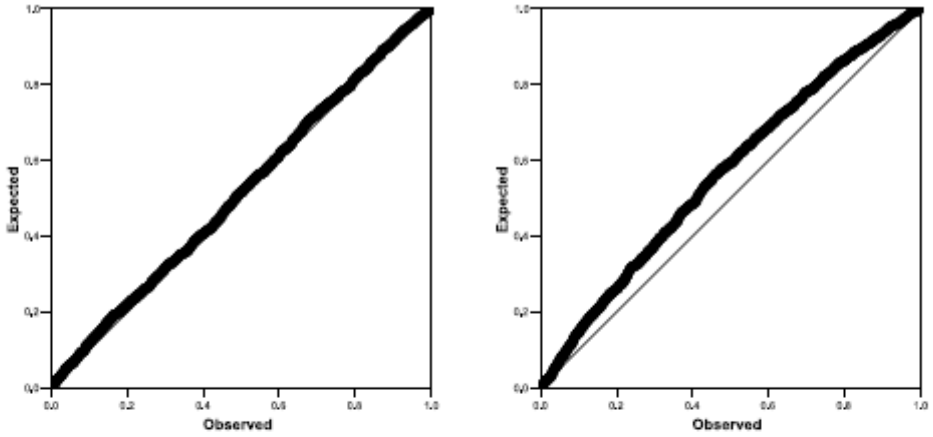
(a) $d = 48$ (b) $d = 960$

FIGURE 2 QQ plots for T_{ML} ($N = 200$; $NR = 1{,}200$).

PP Plots and QQ Plots

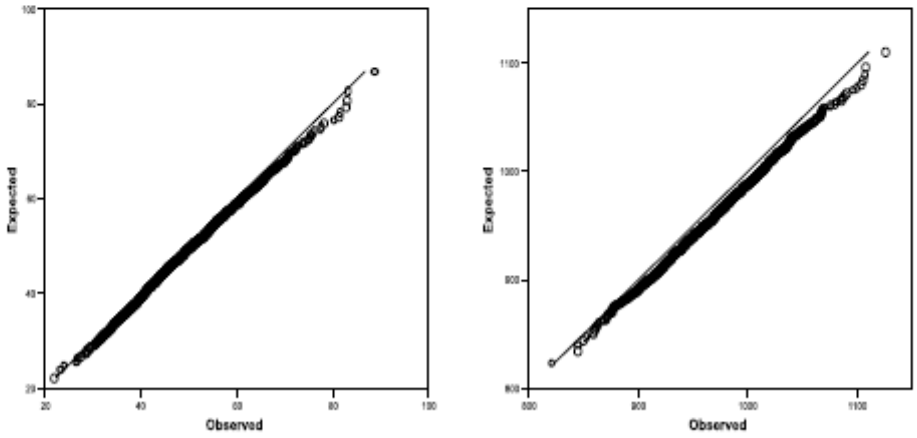
Graphical comparisons of the sampling distributions of the statistics to their reference chi-square distributions are provided to visualize information from Tables 2 through 7. Both PP plots and QQ plots are shown because PP plots are more sensitive to deviations in the middle of a distribution, whereas QQ plots are more sensitive to deviations in its tails (Gnanadesikan, 1977). The plots for TML (Figures 1 and 2) are included because TML serves here as the reference statistic

to illustrate the potential benefits of using TMLs (Figures 3 and 4). In addition, Figures 5 and 6 demonstrate the extremely bad distributional



(a) $d = 48$ (b) $d = 960$

FIGURE 3 PP plots for T_{MLs} ($N = 200$; $NR = 1{,}200$).

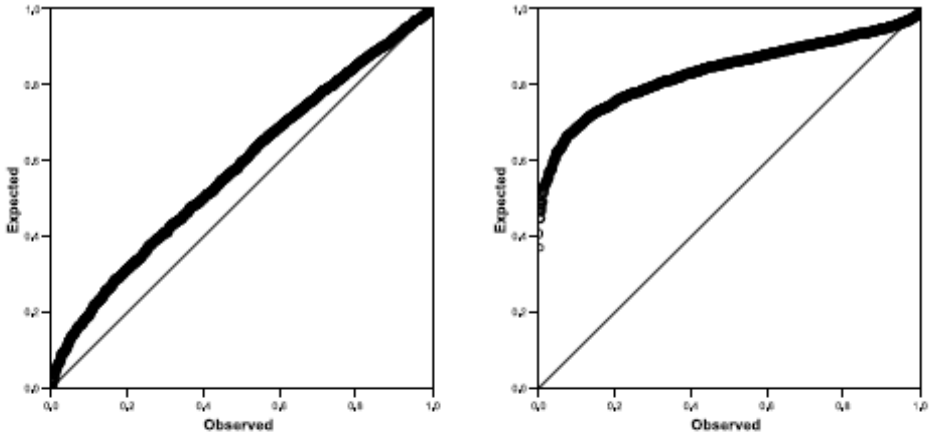


(a) $d = 48$ (b) $d = 960$

FIGURE 4 QQ plots for T_{MLs} ($N = 200$; $NR = 1{,}200$).

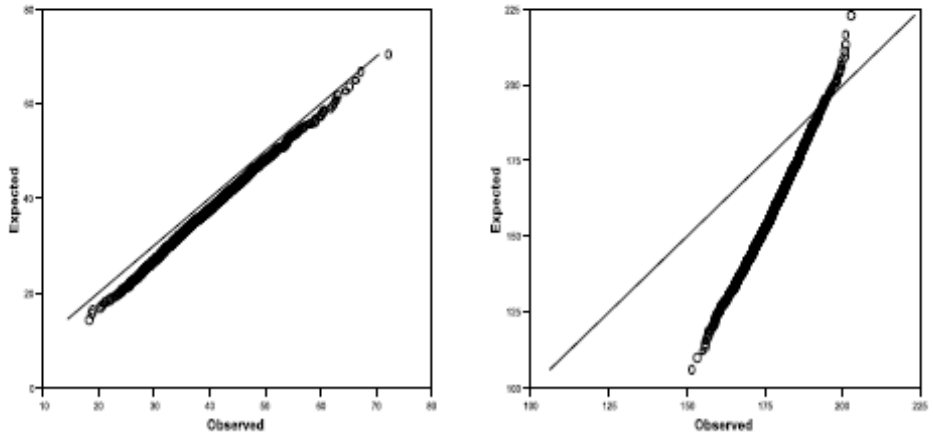
performance of TAD: The 5% Type I error rate is approximately correct but the overall behavior is clearly deviant. The plots for the smallest model (d = 48) and the largest model (d = 960) are shown for the worst case scenario where N = 200.

When comparing Figures 1 and 2 to Figures 3 and 4, the disastrous results for TML clearly emerge. Overall, TMLs has a very close approximation to the reference chi-square distribution. Therefore, we reconfirm our recommendation to use this correction of TML in applied research when large structural equation models are analyzed.



(a) $\bar{d} = 36$ (b) $\bar{d} = 158$

FIGURE 5 PP plots for T_{AD} ($N = 200$; $NR = 1{,}200$).



(a) $\bar{d} = 36$ (b) $\bar{d} = 158$

FIGURE 6 QQ plots for T_{AD} ($N = 200$; $NR = 1{,}200$).

Final Conclusion

In summary, the best performing statistic with respect to Type I error control and the approximation of the reference chi-square distribution is TMLs . Therefore, we recommend using this statistic when many (approximately) multi normal distributed variables are under study in SEM. From Equations 10 through 12 it can be seen that the correction will have only a very small effect on the chisquare value for smaller models or larger sample sizes. From that perspective it would make sense to apply the correction quite generally

Software

A Retrospective View on Applied Research In the following we briefly discuss the consequences of our results for past applied research using large covariance structure models. Even if the estimated models in those applications were specified correctly, with variables having nearly normal distributions, we suspect that the fit of most models was underestimated. Two strategies might have been used when small p values of the chi-square model fit statistics occurred. First, the chi-square statistic for global model fit might be neglected completely and refuge might be taken to other fit statistics (e.g., the RMSEA) or fit indexes (e.g., the TLI, the CFI, and the standardized root mean square residual, SRMR). Apart from the RMSEA, which is asymptotically based on a non central chi-square distribution, research on the distribution of the latter statistics is still at its beginning (e.g., Hu & Bentler, 1999; Ogasawara, 2001). The sampling The calculation of TMLs is quite easy once the value of TML is available, because

Swain's correction factor is a simple function of known values of p, N, and d or t . The p values for the test statistic TMLs are also easily computed with computer software, for example with the function pchisq(x,d), where x D TMLs , and d is the number of degrees of freedom, from freely available R software (cf. Venables & Smith, 2005, section 8.1). Although this is a small effort in practice (the R-function swain for the calculation of TMLs and its corresponding p value can be downloaded from http://www.gmw.rug.nl/_boomsma), we would recommend implementing the Swain correction in standard SEM software.

Example

To illustrate the effects of using TMLs, the value of TML was corrected in a recently published article. Ramaswami and Singh (2003) estimated a confirmatory factor model with N = 154, k = 13, p = 51, d = 1,147, and t = 179. They reported TML = 1,307 with a p value of 0.0007, which would lead to a rejection of the model if a formal test was applied at significance levels of 5% or 10%, say. When the Swain correction is applied, the value of TMLs equals 1,146 with a relatively large increase of the p value to 0.5034. Hence, the model is certainly not rejected when this Swain-corrected test of exact fit is performed. Of course, chi-square dependent statistics like the RMSEA are also affected by the model-size effect: The RMSEA test statistic for close fit would drop from 0.0302 (Ramaswami and Singh reported 0.0320) to 0.0000 when using TMLs.

3.6 DISCUSSION

A Retrospective View on Applied Research

In the following we briefly discuss the consequences of our results for past applied research using large covariance structure models. Even if the estimated models in those applications were specified correctly, with variables having nearly normal distributions, we suspect that the fit of most models was underestimated. Two strategies might have been used when small p values of the chi-square model fit statistics occurred.

First, the chi-square statistic for global model fit might be neglected completely and refuge might be taken to other fit statistics (e.g., the RMSEA) or fit indexes (e.g., the TLI, the CFI, and the standardized root mean square residual, SRMR). Apart from the RMSEA, which is asymptotically based on a noncentral chi-square distribution, research on the distribution of the latter statistics is still at its beginning (e.g., Hu & Bentler, 1999; Ogasawara, 2001). The sampling distribution of most fit indexes is just unknown. Researchers therefore rely on certain cut-off values for such indexes, that have been recommended in the literature (e.g., Hu & Bentler, 1999). These cut-off values are partly arbitrary, and moreover, the blindfolded use of such "golden rules" has proven to be inaccurate under circumstances (Kaplan, 1988; Marsh, Hau, & Wen, 2004; Saris, Den Ronden, & Satorra, 1987). More important, however, is the fact that most fit statistics and indexes are also affected by the inflated TML, because they are a function of this statistic when maximum likelihood estimation is applied. Given the results of our study, it would make sense to substitute TMLs for TML when

calculating these fit statistics and fit indexes. For incremental fit indexes it is not clear whether the fit statistic for the independence model needs to be adjusted similarly; these are issues in need of further research (for first results see Herzog & Boomsma, 2006).

Second, in applied (exploratory) SEM, modification indexes (Sörbom, 1989) are often used extensively, as a last resort in the search for models that cannot be rejected. In many cases, restrictions on covariances among measurement errors are removed without interpreting their meaning, or explaining why such covariances make sense from a theoretical point of view in the first place. This seems to become a common practice, although Jöreskog (1993, p. 297) and many others explicitly criticized this kind of pseudo-theory testing. Given our research findings, the reliability of such model explorations, with TML as its basis, must be questioned even further when at least 12 observed variables are analyzed with sample sizes of up to N = 800.

The results of our study also suggest that it is not unlikely that there may have been many studies in the past where correctly specified large models were not published, because the models were rejected due to the inflated TML. Such phenomena, also labeled "file drawer" problems (e.g., Scargle, 2000), clearly attenuate scientific progress.

The N:t Ratio Criterion

The robustness of model test statistics against model size is not unimportant, as our study shows. An obvious overall remedy to avoid the problem

of inflated values of test statistics is to increase sample size N relative to the number of degrees of freedom d, or to increase N relative to the number of parameters to be estimated t , because t can in principle be interpreted as a measure of model size as well. Certain rules of thumb regarding an adequate sample size relative to the number of parameters t , the N: t ratio, can be found in the literature. Bentler (1995), for example, recommended a ratio of at least 5:1 when TML is used and the assumption of multivariate normality holds. Although such rules of thumb are not without criticism (e.g., Jackson, 2003), we could evaluate our results also in terms of the N: t ratio, that is, the relative sample adequacy. The last column of Tables 2 through 7 shows the value of this ratio. We can now compare our results with earlier N: t recommendations and try to formulate general guidelines in terms of relative sample adequacy for proper behavior of model test statistics. One should realize, however, that the N: t ratio is a simplifying rule of thumb regarding only two of the many factors that matter in a research design.

Our results clearly show that Bentler's 5:1 rule of thumb is not sufficient for the sampling distribution of TML to be approximately chi-square. Even for our smallest model and our largest sample size (d = 48, t = 30, N D = 800), with a N: t ratio of 26.7:1, the Kolmogorov–Smirnov test for TML indicates a significant deviation from the chi-square reference distribution (see Table 7). For our second smallest model (d = 120, t = 51, and N D = 800), a N: t ratio of 15.7:1 is not large enough for proper Type I error behavior of TML at the 5% significance level (see Table 4). Also, in contrast to Fouladi (2000, p. 401), we would not conclude that TAD can be applied under conditions of small N: t ratios. The results in Table 7

show that a ratio of 26.7:1 is insufficient for proper behavior of TAD in moderately large models when inspecting its sampling distribution as a whole, not just its 5% Type I error rates.

Earlier we discussed evidence that the Bartlett statistics suffer from an increasingly conservative trend when model size increases. This effect may be due to the fact that these corrections were originally developed for exploratory factor analyses and not for general covariance structure analyses. For TSCb, this effect is masked by the slightly more liberal tendency of TSC compared to TML. Thus, for the models under study here, we do not observe and cannot conclude, unlike Nevitt and Hancock (2004), that the Bartlett corrections "frequently delivered acceptable Type I error rates at N: t _ 2:1"

The most salient conclusion of our study is that overall the Swain-corrected statistic TMLs performs best. The results in Tables 2 through 7 validate the (strong) conclusion that for the models under study, apart from single smallsample fluctuations, TMLs is robust against large model size if N: t 2:1 under conditions of normality. As will be indicated in the next section, more research is needed to investigate the interaction of nonnormality and model size.

However, although it seems convenient for applied researchers to have rules of thumb like N: t (or N: p ratios for that matter) it would be unwise to follow these guidelines blindly; compare the sincere warnings of Marsh et al. (1998) and Boomsma and Hoogland (2001, p. 142f). First, the mild requirement that for the use of TMLs the N: t ratio should be at least 2:1 should certainly not be interpreted as an encouragement to always stay away from large models, or to use a small

number of indicators per factor, which, as a start, would increase the occurrence of non convergent and improper solutions. Second, easy formulated rules of thumb regarding the N: t ratio also should not overshadow sample size requirements related to the stability of parameter estimates or the size of estimated standard errors of parameter estimates, and considerations as to the power of model test statistics, either locally or globally.

3.7 Limitations and Future Work

- It is well known that non normality has an inflating effect on chi-square model fit statistics (cf. Boomsma, 1983). It should be investigated how well the test statistics, and in particular the Swain-corrected scaled Satorra–Bentler statistic, behave in large models under conditions of non normality.
- This study was confined to factor models. It seems necessary to expand the scope of structural equation models under investigation to a broader range. For these other types of models a main question is also whether and to which extent Bartlett adjustments are effective in comparison with Swain's correction
- Another issue concerns the specific value of 0.70 of the factor loadings that was used in our study. According to the research by Hoogland (1999), the rejection rates are more accurate for smaller factor loadings. Maybe the same pattern will be observed for the test statistics from our study as well.
- The test statistic TMLs deserves additional attention from a statistical power perspective. After assessing the Type I error rates, future studies

should also focus on the power of this corrected test statistic in comparison with a few other promising ones. Emphasis would then turn more to Type II error rates (cf. Nevitt & Hancock, 2004).

- As mentioned earlier, the effect of the proposed corrections of TML on other fit statistics and indexes, like the RMSEA, the TLI, and the CFI, requires further attention. It needs to be investigated to which extent other fit measures are affected by corrected global test statistics (for first results see Herzog & Boomsma, 2006). The SRMR, in our view a fit measure that needs to be inspected in all circumstances, certainly is not.
- This simulation study emphasized the importance of investigating the finite sample behavior of statistics in large models. The disastrous results for TML and TSC may raise questions regarding the generalizations made in many previous simulation studies. One direction of further investigation could be to revisit those studies, and to check whether reported findings generalize to larger models.
- Wakaki, Eguchi, and Fujikoshi (1999) derived a (relatively complex) Bartlett adjustment factor for the test of general covariance structures. In a first simulation study, this correction significantly improved the performance of TML (Kensuke, Takahiro, & Kazuo, 2005). Therefore, it would be of interest to compare its performance with that of the statistics presented here.
- Within the framework of Bayesian estimation of structural equation models, Lee and Song (2004) made a comparison with the classical,

frequentist use of TML, and found that the Bayesian posterior predictive p values are less biased compared to the maximum likelihood p values under conditions of small sample sizes (cf. Scheines, Hoijtink, & Boomsma, 1999). They also found that the posterior predictive p values are not accurate when non normal variables are analyzed. A comparison of the performance of the Bayesian approach to that of TMLs for large models would be intriguing

3.8 CONCLUSION

Some years ago, Kaplan (1988) came to the conclusion that the chi-square model statistic “should be taken seriously as a means of formally testing model specification" (p. 85). For large models, it has been shown here that researchers should seriously consider corrected model test statistics if such a formal approach of model testing is being taken. Otherwise, biased inference might be an undesirable consequence. If this problem is acknowledged, and proper corrections are indeed applied, there are enough obstacles to clean inference left (cf. Jöreskog, 1993).

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Chapter - 4

Modeling covariance structure in the analysis of repeated measures data

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Chapter - 4
Modeling covariance structure in the analysis of repeated measures data

1. INTRODUCTION

Statistical linear mixed models state that observed data consist of two parts, fixed effects and random effects. Fixed effects define the expected values of the observations, and random effects define the variance and covariances of the observations. In typical comparative experiments with repeated measures, subjects are randomly assigned to treatment groups, and observations are made at multiple time points on each subject. Basically, there are two fixed effect factors, treatment and time. Random effects result from variation between subjects and from variation within subjects Measures on the same subject at different times almost always are correlated, with measures taken close together in time being more highly correlated than measures taken far apart in time. Observations on different subjects are often assumed independent, although the validity of this assumption depends on the study design. Mixed linear models are used with repeated measures data to accommodate the fixed effects of treatment and time and the covariation between observations on the same subject at different times. Cnaan et al. [1] extensively discussed the use of the general linear mixed model for analysis of repeated measures and longitudinal data. They presented two example analyses, one using BMDP 5V [2] and the other using PROC MIXED of the SAS System [3]. Although Cnaan et al. discussed statistical analyses in the context of

unbalanced data sets, their description of modelling covariance structure also applies to balanced data sets.

The objectives of repeated measures studies usually are to make inferences about the expected values of the observations, that is, about the means of the populations from which subjects are sampled. This is done in terms of treatment and time effects in the model. For example, it might be of interest to test or estimate difference between treatment means at particular times, or difference between means at deferent times for the same treatment. These are inferences about the fixed effects in the model.

Implementation of mixed models ordinarily occurs in stages. Deferent data analysts may use deferent sequences of stages. Ideally, deferent data sets would be used to choose model form and to estimate parameters, but this is usually not possible in practice. Here we present the more realistic situation of choosing model form using data to be analysed. We prefer a four stage approach, which is similar to recommendations of others, such as Diggle [4] and Bollinger [5].

The first stage is to model the mean structure in sufficient generality to ensure unbiasedness of the _xed eject estimates. This usually entails a saturated parameter specification for fixed effects, often in the form of effects for treatment, time, treatment-by-time interaction, and other relevant covariables. The second stage is to specify a model for the covariance structure of the data. This involves modelling variation between subjects, and also covariation between measures at deferent times on the same subject. In the third stage, generalized least squares methods are used to _t the mean portion of the model. In the fourth stage the fixed

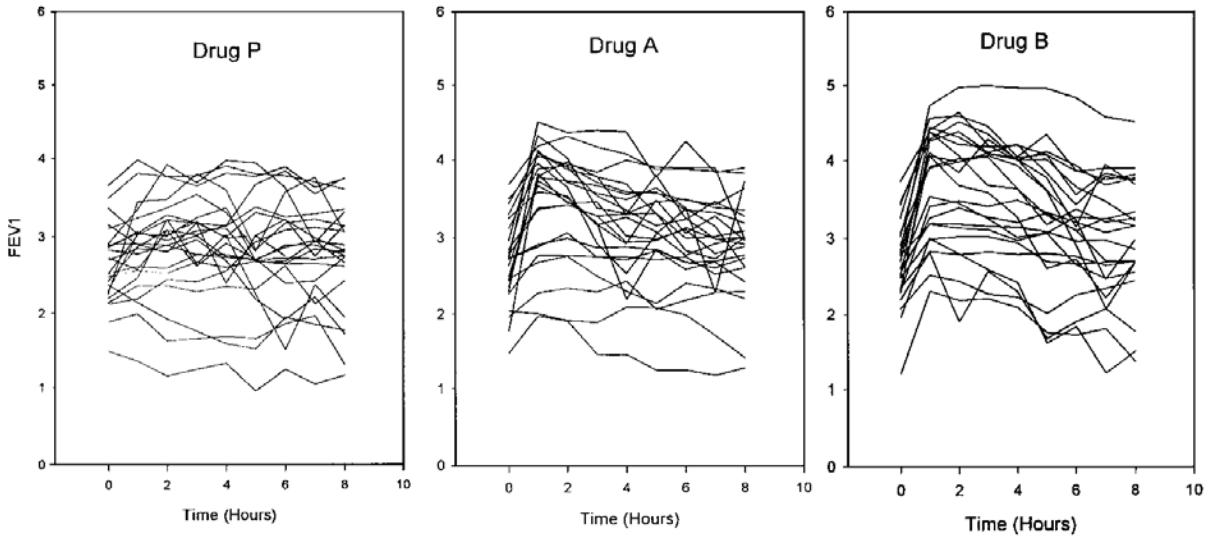
effects portion may be made more parsimonious, such as by fitting polynomial curves over time. Then, statistical inferences are drawn based on fitting this final model.

In the present paper, we illustrate the four-stage process, but the major focus is on the second stage, modelling the covariance structure. If the true underlying covariance structure were known, the generalized least squares mixed effects estimates would be the best linear unbiased estimates (BLUE). When it is unknown, our goal is to estimate it as closely as possible, thus providing more efficient estimates of the fixed effects parameters. The MIXED procedure in the SASJ system [3] provides a rich selection of covariance structures from which to choose. In addition to selecting a covariance structure, we examine the effects of choice of covariance structure on tests of fixed effects, estimates of differences between treatment means, and on standard errors of the differences between means.

2. EXAMPLE DATA SET

A pharmaceutical example experiment will be used to illustrate the methodology. Objectives of the study were to compare effects of two drugs (A and B) and a placebo (P) on a

MODELLING COVARIANCE STRUCTURE FOR REPEATED MEASURES DATA



measure of respiratory ability, called FEV1. Twenty-four patients were assigned to each of the three treatment groups, and FEV1 was measured at baseline (immediately prior to administration of the drugs), and at hourly intervals thereafter for eight hours. Data were analysed using PROC MIXED of the SAS System, using baseline FEV1 as a co-variable. An SAS data set, named FEV1UN1, contained data with variables DRUG, PATIENT, HR (hour), BASEFEV1 and FEV1. Data for individual patients are plotted versus HR in Figure 1 for the three treatment groups. The drug curves appear to follow a classic pharmacokinetic pattern and thus might be analysed using a non-linear mean model. However, we will restrict our attention to models of the mean function which are linear in the parameters. Estimates of between-patient variances within drug group at each hour are printed

Figure 1. FEV1 repeated measures for each patient.

Table I. REML covariance and correlation estimates for FEV1 repeated measures data.

			Unstructured				
Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8
0.226	0.216	0.211	0.204	0.175	0.163	0.128	0.168
0.893	0.259	0.233	0.243	0.220	0.181	0.156	0.195
0.880	0.908	0.254	0.252	0.219	0.191	0.168	0.204
0.784	0.892	0.915	0.299	0.240	0.204	0.190	0.226
0.688	0.807	0.813	0.822	0.286	0.232	0.204	0.247
0.675	0.698	0.745	0.735	0.855	0.258	0.214	0.245
0.516	0.590	0.643	0.670	0.733	0.812	0.270	0.233
0.642	0.701	0.742	0.755	0.845	0.882	0.820	0.299

Variances on diagonal, covariances above diagonal, correlations below diagonal.

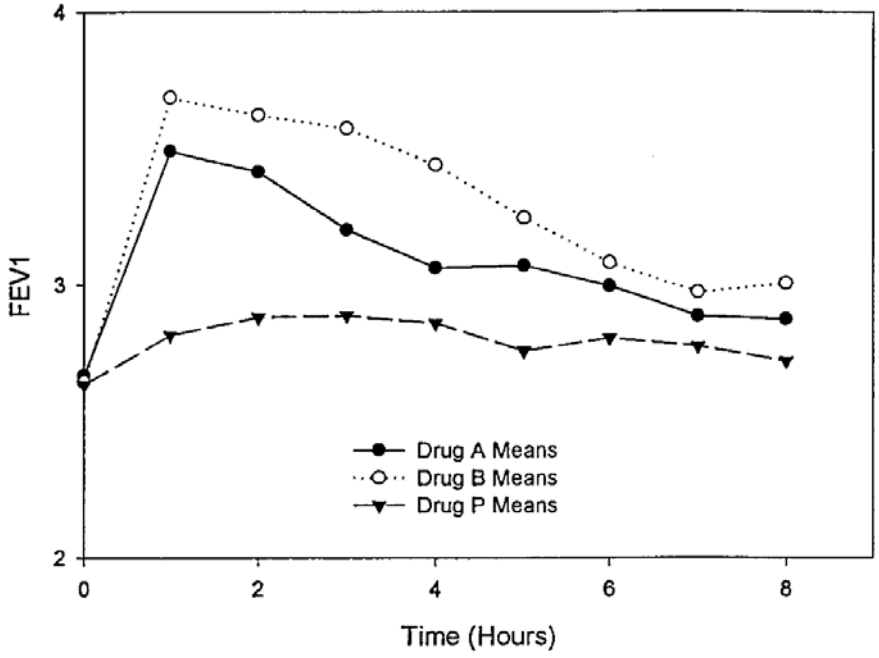


Figure 2. FEV1 repeated measures means for each drug. in the diagonal of the matrix of Table I. It appears from these plots and variance estimates that variances between patients within drug groups are approximately equal across times. Therefore, an assumption of equal variances seems reasonable.

Treatment means are plotted versus HR in Figure 2. The graph shows that means for the three treatment groups are essentially the same at HR =0 (baseline). At HR =1 the mean for drug B is larger than the mean for drug A, and both of the drug means are much larger than the placebo mean. Means for drugs A and B continue to be larger than the placebo means for subsequent hours, but the magnitudes of the differences decrease sharply with time. It is of interest to estimate differences between the treatment group means at various times, and to estimate differences between means for the same treatment at different times.

Co-variances and correlations are printed above and below the diagonal, respectively, of the matrix in Table I. The correlations between FEV1 at HR =1 and later times are in the first column of the matrix. Correlations generally decrease from 0.893 between FEV1 at HR =1 and HR =2 down to 0.642 between FEV1 at HR =1 and HR =8. Similar decreases are found between FEV1 at HR =2 and later times, between FEV1 at HR =3 and later times etc. In short, correlations between pairs of FEV1 measurements decrease with the number of hours between the times at which the measurements were obtained. This is a common phenomenon with repeated measures data. Moreover, magnitudes of correlations between FEV1 repeated measures are similar for pairs of hours with the same interval between hours. Scatter plots of FEV1 for each hour versus FEV1 at each

other hour are presented in Figure 3. These are similar to the `draftsman's' plots as described by Dawson et al. [6]. The trends of decreasing correlations with increasing interval between measurement times is apparent in the plots. That is, points are more tightly packed in plots for two measures close in time than for measures far apart in time.

As a consequence of the patterns of correlations, a standard analysis of variance as prescribed in Milliken and Johnson [7] is likely not appropriate for this data set. Thus, another type of analysis must be used.

MODELLING COVARIANCE STRUCTURE FOR REPEATED MEASURES DATA

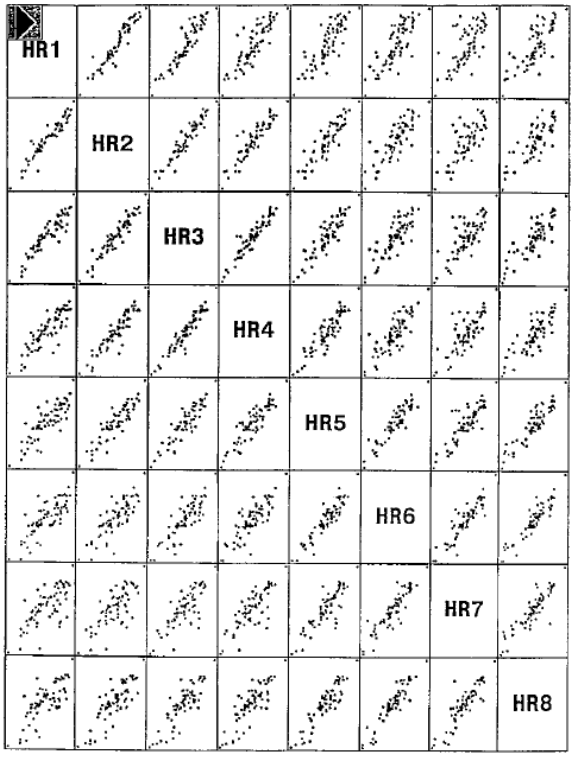


Figure 3. Scatter plots of FEV1 repeated measures at each hour versus each other hour.

3. LINEAR MIXED MODEL FOR REPEATED MEASURES

In this section we develop the general linear mixed model to a minimally sufficient level that will allow the reader to effectively begin using PROC MIXED of the SAS System. The development here is consistent and somewhat overlapping with that of Cnann et al. [2], but is needed for completeness. We assume a completely randomized design for patients in g treatment groups, with ni subjects assigned to group i. Thus, we assume data on different subjects are independent. For simplicity, we assume there are t measurements at the same equally spaced times on each subject. We choose to work in this nicely balanced situation so that we can illustrate the basic issues of modelling covariance structure without complications introduced by unbalanced data.

Let Y_{ijk} denote the value of the response measured at time k on subject j in group i; i=1,…., g, j=1,…., n_i, and k =1,…., t. Throughout this paper, we assume all random effects are normally distributed. The fixed effect portion of the general linear mixed model specifies the expected value of Yijk to be $E(Y_{ijk})= \mu_{ijk}$. The expected value, μ_{ink} , usually is modelled as a function of treatment, time, and other fixed effects covariates. The random effect portion of the model specifies the covariance structure of the observations. We assume that observations on different subjects are independent, which is legitimate as a result of the completely randomized design. Thus, $cov(Y_{ijk}; Y_{i\,j'l})=0$ if $i' \neq i'$ or $j \neq j'$. Also, we assume that variances and covariances of measures on a single subject are the same within each of the groups. However, we allow for the possibility that variances are not

Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijt})'$ denote the vector of data at times $1, 2, \ldots, t$ on subject j in group

homogeneous at all times, and that covariance between observations at different times on the same subject are not the same at all pairs of times. A general covariance structure is denoted as cov(Y_{ijk}; Y_{ijl})=σk,l, where σk,l is the covariance between measures at times k and l on the same subject, and σk,k =σ^2_k denotes the variance at time k. This is sometimes called `unstructured' covariance, because there are no mathematical structural conditions on the variances and covariances.

i. Then, in matrix notation, the model can be written

where is the vector of means $\mathbf{Y}_{ij} = \mu_{ij} + \varepsilon_{ij}$ and ε_{ij} =($\varepsilon_{ij}1$,$\varepsilon_{ij}2$,….., ε_{ij} t)′ is the vector of errors, respectively, for subject j in group i. Matrix representations of the expectation and variance of Y_{ij} are E(Y_{ij})=μ_{ij} and V(Y_{ij})=V_{ij} , where V_{ij} is the t ×t matrix with σk,l in row k, column l. We assume that V_{ij} is the same for all subjects (that is, for all i and j), but we continue to use the subscripts ij to emphasize that we are referring to the covariance matrix for a single subject.

We represent the vector of data for all subjects as Y=(Y′$_{11}$,…..,Y′ $_{1n,}$Y′$_{21}$,….,Y′ 2n,…..,Y′$_{g1}$,….., Y′$_{gn}$) ′, and similarly for the vectors of expected values and errors to get E(Y)=μ=(μ′$_{11}$,……,μ_{1n}, μ′$_{21}$,…, μ′$_{2n}$,….., μ′$_{g1}$, ….., μ′$_{gn}$)′ and ε =(ε′$_{11}$, ….. , ε′$_{1n}$; ε′$_{21}$,…… , ε′$_{2n}$,……, ε′$_{g1}$,……, ε′$_{gn}$) ′. Then we have the model

$$\mathbf{Y} = \mu + \varepsilon \qquad (1)$$

and

$$V(\mathbf{Y}) = \mathbf{V} = \mathrm{diag}\{\mathbf{V}_{ij}\}$$

where diag{V_{ij}} refers to a block-diagonal matrix with V_{ij} in each block.

A univariate linear mixed model for the FEV1 repeated measures data is

$$Y_{ijk} = \mu + \lambda x_{ij} + \alpha_i + d_{ij} + \tau_k + (\alpha\tau)_{ik} + e_{ijk} \qquad (2)$$

where μ is a constant common to all observations, λ is a fixed coefficient on the covariate x_{ij} =BASEFEV1 for patient j in drug group i, α_i is a parameter corres ponding to drug i,τ_k is a parameter corresponding to hour k, and $(\alpha\tau)_{ik}$ is an interaction parameter corresponding to drug i and hour k; d_{ij} is a normally distributed random variable with mean zero and variance σ_{2d} cor- responding to patient j in drug group i, and e_{ijk} is a normally distributed random variable with me an zero and variance, independent of d_{ij} , corresponding to patient j in drug group i at hour k. Then

$$\begin{aligned} E(Y_{ijk}) &= \mu_{ijk} = \mu + \lambda x_{ij} + \alpha_i + \tau_k + (\alpha\tau_{ik}) \\ V(Y_{ijk}) &= \sigma_d^2 + \sigma_e^2 \end{aligned} \qquad (3)$$

and

$$\operatorname{cov}(Y_{ijk}, Y_{ijl}) = \sigma_d^2 + \operatorname{cov}(e_{ijk}, e_{ijl})$$

The model (2), written in matrix rotation, is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \mathbf{e} \qquad (4)$$

$\mu, \lambda, \alpha_i, \tau_k,$ and $(\alpha\tau_{ik}), \beta$ is where X is a matrix of known coefficients of the fixed effect parameters the vector of fixed effect parameters, Z is a matrix of (1), $\mu = \mathbf{X}\beta$ and $\varepsilon = \mathbf{Z}\mathbf{U} + \mathbf{e}$. coefficients (zeros and ones) of the random patient effects d_{ij} ;U is the vector of random effects d_{ij} , and e is the vector of the errors e_{ijk}. In relation to model

Model (4) for the FEV1 data is a special case of the general linear mixed model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZU} + \mathbf{e} \qquad (5)$$

in which no restrictions are necessarily imposed on the structures of G=V(U) and R=V(e). We assume only that U and e are independent, and obtain Equation (6) expresses the structure of V(Y) as a function of G and R. In many repeated measures applications, ZGZ*0* represents the between-patient portion of the covariance structure, and R rep-resents the within-patient portion. By way of notation, sub-matrices of X;Z;R and e corresponding to subject j in drug group i will be denoted by X_{ij} ;Z_{ij} ;R_{ij} and e_{ij} , respectively.

More details on implementation of the model for statistical inference are presented in the

$$V(\mathbf{Y}) = \mathbf{ZGZ}' + \mathbf{R}. \qquad (6)$$

In order to apply the general linear mixed model (5) using PROC MIXED in the SAS System, the user must specify the three parts of $\mathbf{X}\beta, \mathbf{ZU}$ the model and e. Specifying X_ is done in the same manner as with PROC GLM, and presents no new challenges to PROC MIXED users who are familiar with GLM. However, specifying ZU and e entails de_ning covariance structures, which may be less familiar concepts. Several covariance structures are discussed in Section 4.

4. COVARIANCE STRUCTURES FOR REPEATED MEASURES

Modelling covariance structure refers to representing V(Y) in (6) as a function of a relatively small number of parameters. Functional specification of the covariance structure for the mixed model is done through G and R of (5), often only in terms of Rij . We present six covariance structures that will be fitted to the FEV1 data. Since observations on different patients are assumed independent, the structure refers to the covariance pattern of measurements on the same subject. For most of these structures, the covariance between two observations on the same subject depends only on the length of the time interval between measurements (called the lag), and the variance is constant over time. We assume the repeated measurements are equally spaced so we may define the lag for the observations Y_{ijk} and Y_{ijl} to be the absolute value of k − l, that is |k − l|. For these structures, the covariance can be characterized in terms of the variance and the correlations expressed as a function of the lag. We generically denote the correlation function $corr_{XXX}$(lag), where XXX is an abbreviation for the name of a covariance structure.

4.1. Simple (SIM)

$$\mathrm{cov}(Y_{ijk}, Y_{ijl}) = 0 \text{ if } k \neq l, \quad V(Y_{ijk}) = \sigma^2_{\mathrm{SIM}}$$

Simple structure specifies that the observations are independent, even on the same patient, and have homogeneous variance $V(Y_{ijk})= \sigma^2_{SIM}$. The correlation function is $corr_{SIM}(lag)=0$. Simple structure is not realistic for most repeated

measures data because it specifies that observations on the same patient are independent. In terms of model (5), G=0 and Rij = σ^2_{SIM}I, where I is an identity matrix. For the model (3), simple structure would be obtained with d_{ij} =0 (equivalently,

$$\sigma_d^2 = 0),\ \mathrm{cov}(e_{ijk}, e_{ijl}) = 0 \text{ for } k \neq l, \text{ and } V(e_{ijk}) = \sigma^2_{\mathrm{SIM}}.$$

4.2. Compound Symmetric (CS)

$$\mathrm{cov}(Y_{ijk}, Y_{ijl}) = \sigma^2_{\mathrm{CS,b}} \text{ if } k \neq l, \quad V(Y_{ijk}) = \sigma^2_{\mathrm{CS,b}} + \sigma^2_{\mathrm{CS,w}}$$

Compound symmetric structure specifies that observations on the same patient have $\sigma^2_{\mathrm{CS,b}}$, homogeneous covariance and homogeneous variance $V(Y_{ijk})$=_2 CS; b + _2 CS;w. The correlation function is

$$\mathrm{corr}_{\mathrm{CS}}(\mathrm{lag}) = \sigma^2_{\mathrm{CS,b}}/(\sigma^2_{\mathrm{CS,b}} + \sigma^2_{\mathrm{CS,w}})$$

Notice that the correlation does not depend on the value of lag, in the sense that the correlations between two observations are equal for all pairs of observations on the same subject. Compound symmetric structure is sometimes called `variance components' structure, because the two parameters $\sigma^2_{\mathrm{CS,b}}$ and $\sigma^2_{\mathrm{CS,w}}$ represent between-subjects and within-subjects variances, respectively. This mix of between- and within-subject variances logically motivates the form of $V(Y_{ij})$ in many situations and implies a non-negative correlation between pairs of within-subject observations. It can be specified in one of two ways through G and R in (5). One way is to define $\mathbf{G} = \sigma^2_{\mathrm{CS,b}}\mathbf{I},$ and $\mathbf{R} = \sigma^2_{\mathrm{CS,w}}\mathbf{I}.$. In terms of the univariate

model (3), we would have $\sigma_d^2 = \sigma_{\mathrm{CS,d}}^2$, $\mathrm{cov}(e_{ijk}, e_{ijl}) = 0$ for $k \neq l$, and $V(e_{ijk}) = \sigma_{\mathrm{CS,w}}^2$. The other way to specify compound symmetric structure is to define G=0, and define R_{ij} to be compound symmetric; for example, $\mathbf{R}_{ij} = \sigma_{\mathrm{CS,w}}^2 \mathbf{I} + \sigma_{\mathrm{CS,b}}^2 \mathbf{J}$, where J is a matrix of ones. In terms of the univariate model (3), we would have $\sigma_d^2 = 0$, $\mathrm{cov}(e_{ijk}, e_{ijl}) = \sigma_{\mathrm{CS,b}}^2$ for k ≠l, and V(e_{ijk})= $\sigma_{\mathrm{CS,b}}^2 + \sigma_{\mathrm{CS,w}}^2$. The second formulation using only the R matrix is more general, since it can be defined with negative within-subject correlation as well.

4.3. Autoregressive, order 1 (AR(1))

$$\mathrm{cov}(Y_{ijk}, Y_{ijl}) = \sigma_{\mathrm{AR(1)}}^2 \rho_{\mathrm{AR(1)}}^{|k-l|}$$

Thus, observations on the same patient far apart in time would be essentially independent, which may not be realistic. Autoregressive structure is de_ned in model (5) entirely in terms of R, with G=0. The element in row k, column l of Rij is denoted to be $\sigma_{2\mathrm{AR(1)}}\, \rho^{|k-l|}_{\mathrm{AR(1).}}$ In terms of the univariate model (3), we would have σ_{2d} =0, and cov(e_{ijk},e_{ijl})= $\sigma_{2\mathrm{AR(1)}}\, \rho^{|k-l|}_{\mathrm{AR(1)}}$.

4.4. Autoregressive with random e_ect for patient (AR(1)+RE)

$$\mathrm{cov}(Y_{ijk}, Y_{ijl}) = \sigma_{\mathrm{AR(1)+RE,b}}^2 + \sigma_{\mathrm{AR(1)+RE,w}}^2 \rho_{\mathrm{AR(1)+RE}}^{|k-l|}$$

$$V(Y_{ijk}) = \sigma_{\mathrm{AR(1)+RE,b}}^2 + \sigma_{\mathrm{AR(1)+RE,w}}^2.$$

$$\mathrm{corr}_{\mathrm{AR(1)+RE}}(\mathrm{lag}) = (\sigma_{\mathrm{AR(1)+RE,b}}^2 + \sigma_{\mathrm{AR(1)+RE,w}}^2 \rho_{\mathrm{AR(1)+RE}}^{\mathrm{lag}}) / (\sigma_{\mathrm{AR(1)+RE,b}}^2 + \sigma_{\mathrm{AR(1)+RE,w}}^2)$$

Autoregressive with random effect for patient covariance structure specifies homogeneous variance The correlation fun Autoregressive $\sigma^2_{AR(1)+RE,b}$ plus random effects structure specifies that covariance between observations on the same patient comes from two sources. First, any two observations share a $\sigma^2_{AR(1)+RE,b}$. common contribution simply because they are on the same subject. This is the portion of the covariance, and results from defining a random e_ect for patients. Second, the covariance between observations decreases exponentially with lag, but decreases only to This is the autoregressive contribution to the covariance $\sigma^2_{AR(1)+RE,w}\rho^{|k-l|}$. In terms of model (5), AR(1)+RE is represented with $\mathbf{G} = \sigma^2_{AR(1)+RE,b}\mathbf{I}$ and autoregressive Rij . In terms of the univariate model (3), we would have $\sigma^2_d = \sigma^2_{AR(1)+RE,b}$, and cov($e_{ijk}$; e_{ijl})= $\sigma^2_{AR(1)+RE,w}\rho^{|k-l|}$. The AR(1)+RE covariance structure actually results from a special case of the model proposed by Diggle [4].

4.5. Toeplitz (TOEP)

$$\text{cov}(Y_{ijk}; Y_{ijl}) = \sigma_{TOEP,|k-l|},\ V(Y_{ijk}) = \sigma^2_{TOEP}$$

Toeplitz structure, sometimes called `banded', specifies that covariance depends only on lag, but not as a mathematical function with a smaller number of parameters. The correlation function is corr(lag)= $\sigma_{TOEP\ |lag|} = \sigma^2_{TOEP}$. In terms of model (5), TOEP structure is given with G=0. The elements of the main diagonal of R are _2 TOEP. All elements in a sub-diagonal |k−l|=lag are $\sigma_{TOEP\ \ |k-l|}$, where k is the row number and l is the column number.

4.6. Unstructured (UN)

$$\mathrm{cov}(Y_{ijk}, Y_{ijl}) = \sigma_{\mathrm{UN},kl}$$

The `unstructured' structure specifies no patterns in the covariance matrix, and is completely general, but the generality brings the disadvantage of having a very large number of parameters. In terms of model (5), it is given with G=0 and a completely general R_{ij} .

5. USING THE MIXED PROCEDURE TO FIT LINEAR MIXED MODELS

We now turn to PROC MIXED for analyses of the FEV1 data which _t the mean model (3) and accommodate structures defined on the covariance matrix. We assume the reader has some familiarity with the SAS System, and knows how to construct SAS data sets and call SAS procedures.

The general linear mixed model (5) may be fit by using the MODEL, CLASS, RANDOM and REPEATED statements in the MIXED procedure. The MODEL statement consists of an equation which specifies the response variable on the left side of the equal sign and terms on the right side to specify the fixed effects portion of the model, X β. Readers familiar with the GLM procedure in SAS will recognize the RANDOM and REPEATED statements as being available in GLM, but their purposes are quite different in MIXED. The RANDOM statement in MIXED is used to specify the random effects portion, ZU, including the structure of V(U)=G. The REPEATED statement in MIXED is used to specify

the structure of V(e)=R. Also, the MODEL statement in MIXED contains only fixed effects, but in GLM it contains both fixed and random effects. The CLASS statement, however, has a similar purpose in MIXED as in GLM, which is to specify classification variables, that is, variables for which indicator variables are needed in either X or Z. The CLASS statement in MIXED also is used to identify grouping variables, for example, variables that delineate the submatrices of block diagonal G or R.

In the FEV1 data, PATIENT and DRUG are clearly classification variables, and must be listed in the CLASS statement. The variable HR (hour) could be treated as either a continuous or a classification variable. In the first stage of implementing the linear mixed model, the mean structure E(Y)=X β usually should be fully parameterized, as emphasized by Diggle [4]. Underspecifying the mean structure can result in biased estimates of the variance and covariance parameters, and thus lead to an incorrect assessment of covariance structure. Therefore, unless there are a very large number of levels of the repeated measures factor, we usually specify the repeated measures factor as a classification variable. Thus, we include the variable HR in the CLASS statement class drug patient hr;

On the right side of the MODEL statement, we list terms to specify the mean structure (3) model fev1=basefev1 drug hr drug _ hr Executing the statements proc mixed data=fev1uni; class drug patient hr; (7) model fev1= basefev1 drug hr drug _ hr; would provide an ordinary least squares _t of the model (3). Results would be equivalent to those obtained by executing the CLASS and MODEL statements in (7) using PROC GLM. All tests of hypotheses,

standard errors, and con_dence intervals for estimable functions would be computed with an implicit assumption that $V(Y)= \sigma_2 I$, that is, that G=0 and that $R= \sigma_2 I$.

Specifying the MODEL statement in (7) is basically stage 1 of our four-stage process. Stage 2 is to select an appropriate covariance structure. The covariance structures described in Section 3 may be implemented in PROC MIXED by using RANDOM and=or REPEATED statements in conjunction with the statements (7). These statements cause PROC MIXED to compute Residual Maximum Likelihood (REML, also known as restricted maximum likelihood) or Maximum Likelihood (ML) (Searle et al. reference [8], chapter 6) estimates of covariance parameters for the specified structures.

Several options are available with the REPEATED and RANDOM statements, and would be specified following a slash (=). Following is a list of some of the options, and a brief description

of their functions:

TYPE = *structure type*.	Specifies the type of structure for **G** or **R**. Structure options are given in SAS Institute Inc. [3].
R and RCORR (REPEATED).	Requests printing of **R** matrix in covariance or correlation form.
G and GCORR (RANDOM).	Requests printing of **G** matrix in covariance or correlation form.
V and VCORR (RANDOM).	Requests printing of $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ matrix in covariance or correlation form
SUBJECT = *variable name*.	Specifies variables whose levels are used to identify block diagonal structure in **G** or **R**. When used in conjunction with R, RCORR, G, GCORR, V, or VCORR options, only a sub-matrix for a single value of the variable is printed.

We now present statements to produce each of the covariance structures of Section 3. Basic output from these statements would include a table of estimates of parameters in the specified covariance structure and a table of tests of fixed effects, similar to an analysis of variance table. In each of the REPEATED statements, there is a designation `SUBJECT=PATIENT (DRUG)'. This specifies that R is a block diagonal matrix with a sub-matrix for each patient. In this example, it is necessary to designate PATIENT (DRUG) because patients are numbered 1{24 in each drug. If patients were numbered 1{72, with no common numberings in different drugs, it would be sufficient to designate only `PATIENT'. The options R and RCORR are used with the REPEATED statement and V and VCORR are used with the RANDOM statement to request printing of covariance and correlation matrices.

5.1 Simple

This is the default structure when no RANDOM or REPEATED statement is used, as in statements (7), or when no TYPE option is specified in a RANDOM or REPEATED statement. It can be specified explicitly with a REPEATED statement using a TYPE option:

```
proc mixed data=fev1uni;
class drug patient hr;
model fev1=basefev1 drug hr drug * hr; (8)
repeated/type=vc subject=patient(drug) r corr;
```

Note that in SAS version 6.12, the option `simple' can replace `vc' in the REPEATED statement.

5.2. Compound Symmetric

As noted in the previous section, compound symmetric covariance structure can be specified two different $\mathbf{G} = \sigma^2_{CS,b}\mathbf{I}$ and $\mathbf{R} = \sigma^2_{CS,w}\mathbf{I}$, ways using G or R. Correspondingly, it can be implemented two different ways in the MIXED procedure, which would give identical results for non-negative within-subject correlation, except for labelling. The first way, setting is implemented with the RANDOM statement:

```
proc mixed data = fev1uni; class drug patient hr;
        model fev1 = basefev1 drug hr drug * hr;          (9)
        random patient(drug);
```

The RANDOM statement defines $\mathbf{G} = \sigma^2_{CS,b}\mathbf{I}$ and the absence of a REPEATED statement (by default) defines $\mathbf{R} = \sigma^2_{CS,w}\mathbf{I}$. The second way, setting G=0 and $R_{ij} = \sigma^2_{CS,w}\mathbf{I} + \sigma^2_{CS,b}\mathbf{J}$, is implemented with a REPEATED statement using a SUBJECT and TYPE options. The following statements would specify compound symmetric structure for each individual patient, and print the Rij submatrix for one patient in both covariance and correlation forms:

The PROC MIXED output from statements (10) is shown in Figure 4, so that the reader can relate it to the parts we summarize in tables.

```
proc mixed data = fev1uni; class drug patient hr;
        model fev1 = basefev1 drug hr drug * hr;          (10)
        repeated/type = cs subject = patient(drug) r rcorr;
```

5.3. Autoregressive, order 1

This covariance structure would be specified for each patient using a REPEATED statement:

```
proc mixed data = fev1uni; class drug patient hr;
     model fev1 = basefev1 drug hr drug * hr;                    (11)
     repeated/type = ar(1) subject = patient(drug) r rcorr;
```

5.4. Autoregressive with random effect for patient

This covariance structure involves both G and R, and therefore requires both a RANDOM and a REPEATED statement:

```
proc mixed data = fev1uni; class drug patient hr;
     model fev1 = basefev1 drug hr drug * hr;
     random patient(drug);                                       (12)
     repeated/type = ar(1) subject = patient(drug);
```

The RANDOM statement defines $\mathbf{G} = \sigma^2_{\mathrm{AR(1)+RE:b}}\mathbf{I}$; and the REPEATED statement defines Rij to be autoregressive, with parameters $\sigma^2_{\mathrm{AR(1)+RE,w}}$ and $\rho_{\mathrm{AR(1)+RE}}$. Notice that we have no R and RCORR options in the REPEATED statement in (12). Covariance and correlation estimates that would be printed by R and RCORR options in (12) would not be directly comparable with the other covariance's and correlations for other structures that are defined by REPEATED statements without a RANDOM statement. Covariance and correlation estimates that would be printed by R and RCORR options in the REPEATED statement in (12) would pertain only to the R matrix. Estimates for AR(1)+RE structure which are comparable to covariances and correlations for other structures must be based on covariances of the observation vector Y, that is,

on V(Y)=ZGZ′ + R. This could be printed by using V and VCORR options in the RANDOM statement in (12). However, the entire ZGZ′ + R matrix, of dimension 576×576, would be printed. Alternatively, the statements (13) could be used, which are the same as (12) except for the RANDOM statement, but would print only $Z_{ij}GZ'_{ij} + R_{ij}$, of dimension 8×8.

```
proc mixed data = fev1uni; class drug patient hr;
        model fev1 = basefev1 drug hr drug * hr;
        random int/subject = patient(drug) v vcorr;                    (13)
        repeated/type = ar(1) subject = patient(drug);
```

Effects of Three Drugs on FEV1
Compound Symmetric with BaseFEV1 Covariable

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
CS	PATIENT(DRUG)	0.20625696
Residual		0.06312683

Model Fitting Information for FEV1

Description	Value
Observations	576.0000
Res Log Likelihood	-173.645
Akaike's Information Criterion	-175.645
Schwarz's Bayesian Criterion	-179.957
-2 Res Log Likelihood	347.2902
Null Model LRT Chi-Square	569.6449
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0000

Effects of Three Drugs on FEV1
Compound Symmetric with BaseFEV1 Covariable

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
BASEFEV1	1	68	76.42	0.0001
DRUG	2	68	7.24	0.0014
HR	7	483	38.86	0.0001
DRUG*HR	14	483	7.11	0.0001

Figure 4. Basic PROC MIXED output for compound symmetric covariance structure.

Executing statements (13) results in the covariance and corresponding correlation estimates for AR(1)+RE structure shown in Table II. The RANDOM statement in (13) defense ZU in (5) equivalent to the RANDOM statement in (12), but from an `individual subject' perspective rather than a `sample of subjects' perspective. The RANDOM statement in (12) basically defines column soft Z as indicator variables for different patients. The RANDOM statement in (13), with the `int/sub=patient(drug)' designation, defines a set of ones as `intercept' coefficients for each patient.

Table II. REML variance, covariance and correlation estimates for five covariance structures for FEV1 repeated measures.

Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8
1. *Simple*							
0.267	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2. *Compound Symmetric*							
0.269	0.206	0.206	0.206	0.206	0.206	0.206	0.206
1.0	0.766	0.766	0.766	0.766	0.766	0.766	0.766
3. *Autoregressive* (1)							
0.266	0.228	0.195	0.167	0.143	0.123	0.105	0.090
1.0	0.856	0.733	0.629	0.538	0.461	0.394	0.338
4. *Autoregressive* (1) *with random effect for patient*							
0.268	0.230	0.209	0.198	0.192	0.189	0.187	0.186
1.0	0.858	0.780	0.739	0.716	0.705	0.698	0.694
5. *Toeplitz* (*banded*)							
0.266	0.228	0.216	0.207	0.191	0.183	0.169	0.158
1.0	0.858	0.811	0.777	0.716	0.686	0.635	0.593

Variances and covariances in top line; correlations in bottom line.

5.5. Toeplitz

This structure can be specified in terms of **R** with $\mathbf{G} = \mathbf{0}$, and therefore requires only a REPEATED statement:

$$\begin{aligned} &\text{proc mixed data} = \text{fev1uni; class drug patient hr;} \\ &\quad \text{model fev1} = \text{basefev1 drug hr drug} * \text{hr;} \\ &\quad \text{repeated/type} = \text{toep subject} = \text{patient(drug) r rcorr;} \end{aligned} \tag{14}$$

5.6. Unstructured

This structure can be specified in terms of **R** with $\mathbf{G} = \mathbf{0}$, and therefore requires only a REPEATED statement:

$$\begin{aligned} &\text{proc mixed data} = \text{fev1uni; class drug patient hr;} \\ &\quad \text{model fev1} = \text{basefev1 drug hr drug} * \text{hr;} \\ &\quad \text{repeated/type} = \text{un subject} = \text{patient(drug) r rcorr;} \end{aligned} \tag{15}$$

Parameter estimates in the covariance and correlation matrices for the various structures (excepting 'unstructured') are:

SIM $\quad \hat{\sigma}^2_{\text{SIM}} = 0.267$

CS
$\hat{\sigma}^2_{\text{CS,b}} = 0.206$
$\hat{\sigma}^2_{\text{CS,w}} = 0.063$

AR(1)
$\hat{\sigma}^2_{\text{AR(1)}} = 0.266$
$\hat{\rho}\hat{\sigma}_{\text{AR(1)}} = 0.856$

AR(1) + RE
$\hat{\sigma}^2_{\text{AR(1)+RE,b}} = 0.185$
$\hat{\sigma}^2_{\text{AR(1)+RE,w}} = 0.083$
$\hat{\rho}_{\text{AR(1)+RE}} = 0.540$

TOEP
$\hat{\sigma}^2_{\text{TOEP}} = 0.266$
$\hat{\sigma}_{\text{TOEP,1}} = 0.228$
$\hat{\sigma}_{\text{TOEP,2}} = 0.216$
$\hat{\sigma}_{\text{TOEP,3}} = 0.207$
$\hat{\sigma}_{\text{TOEP,4}} = 0.191$
$\hat{\sigma}_{\text{TOEP,5}} = 0.183$
$\hat{\sigma}_{\text{TOEP,6}} = 0.169$
$\hat{\sigma}_{\text{TOEP,7}} = 0.158$

UN (parameter estimates shown in Table I).

The covariance and correlation matrices resulting from statements (8), (10), (11), (13) and (14) are summarized in Table II. Rather than printing the entire matrices, covariances and correlations are displayed as a function of lag for SIM, CS, AR(1), AR(1)+RE and TOEP structures. Covariances and correlations resulting from (15) are printed in Table I.

6. COMPARISON OF FITS OF COVARIANCE STRUCTURES

We discuss covariance and correlation estimates in Table II for the structured covariances in comparison with those in Table I for the unstructured covariances. First, simple and compound symmetric estimates in Table II clearly do not reflect the trends in Table I. Autoregressive estimates in Table II show the general trend of correlations decreasing with length of time interval, but the values of the correlations in the autoregressive structure are too small, especially for long intervals. Thus, none of SIM, CS or AR(1) structures appears to adequately model the correlation pattern of the data. The AR(1)+RE correlations in Table II show good agreement with TOEP estimates in Table II and UN estimates in Table I. Generally, we prefer a covariance model which provides a good _t to the UN estimates, and has a small number of parameters. On this principle, AR(1)+RE is preferable.

The correlogram (Cressie, reference [9], p. 67) is a graphical device for assessing correlation structure. It is basically a plot of the correlation function. Correlation plots are shown in Figure 5 based on estimates assuming UN, CS, AR(1), AR(1)+RE and TOEP structures. Plots for CS, AR(1), AR(1)+RE and TOEP may be considered correlogram estimates assuming these structures. Of these correlations which are a function only of lag, the TOEP structure is the most general, and thus is used as the reference type in Figure 5. These plots clearly show that the plot of the AR(1)+RE structure agrees with TOEP and is superior to the plots of CS and AR(1).

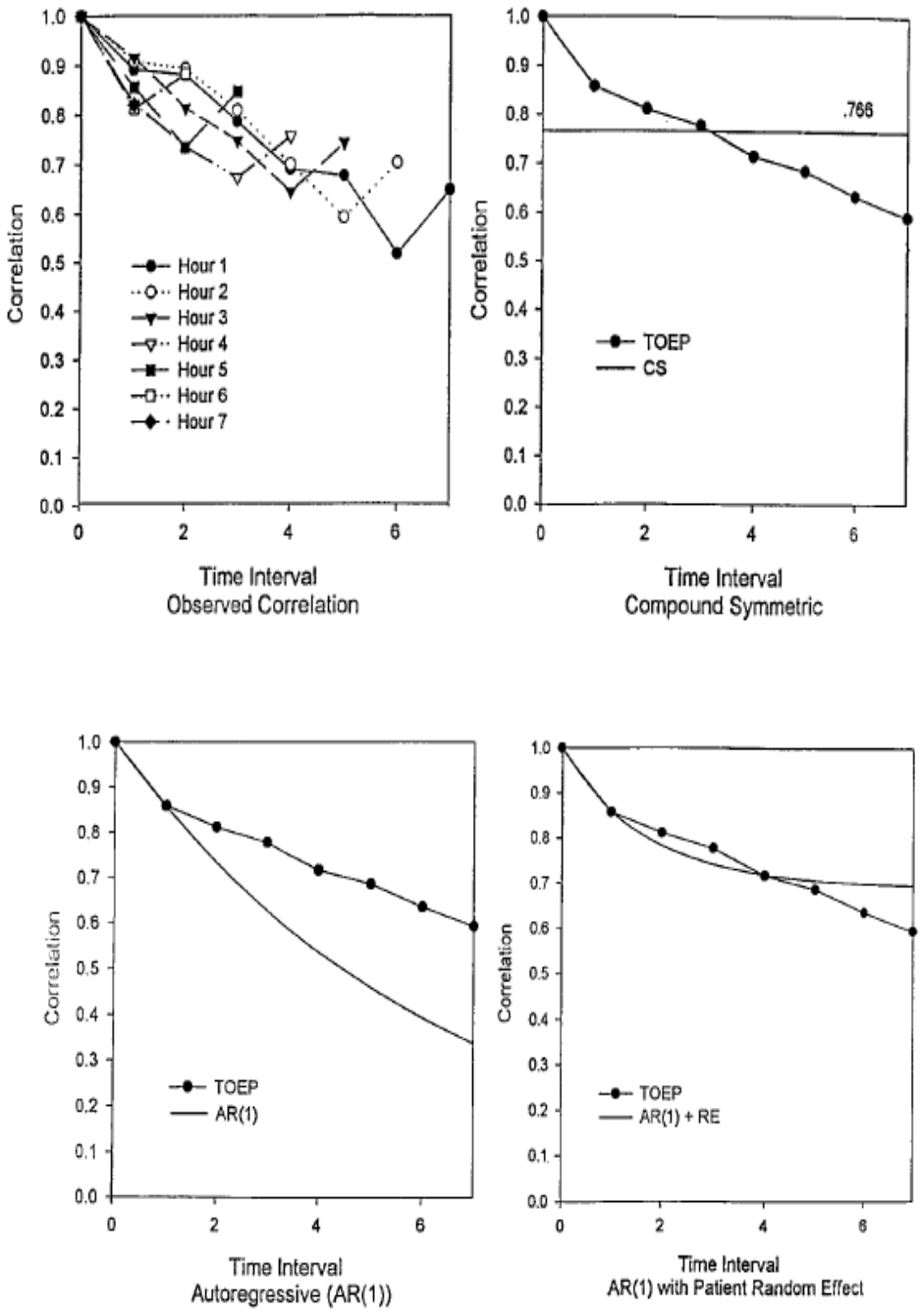


Figure 5. Plots of correlation estimates and correlograms.

Akaike's information criterion (AIC) [10] and Schwarz's Bayesian criterion (SBC) [11] are indices of relative goodness-of-_t and may be used to compare models with the same fixed effects but different covariance structures. Both of these criteria apply rather generally for purposes of model selection and hypothesis testing. For instance, Kass and Wassermann [12] have shown that the SBC provides an approximate Bayes factor in large samples. Formulae for their computation are

$$\mathrm{AIC} = L(\hat{\theta}) - q$$
$$\mathrm{SBC} = L(\hat{\theta}) - (q/2)\log(N^*)$$

Table III. Akaike's information criterion (AIC) and Schwarz's Bayesian criterion (SBC) for six covariance structures.

	Structure name	AIC	SBC
1.	Simple	−459.5	−461.6
2.	Compound symmetric	−175.6	−179.9
3.	Autoregressive (1)	−139.5	−143.8
4.	Autogressive (1) with random effect for patients	−126.5	−132.9
5.	Toeplitz (banded)	−121.9	−139.2
6.	Unstructured	−110.1	−187.7

where L(oˆ) is the maximized log-likelihood or restricted log-likelihood (REML), q is the number of parameters in the covariance matrix, p is the number of fixed effect parameters and N* is the total number of `observations' (N for ML and N − p for REML, where N is the number of subjects).

Models with large AIC or SBC values indicate a better fit. However, it is important to note that the SBC criterion penalizes models more severely for the number of estimated parameters than does AIC. Hence the two criteria will not always agree on the choice of `best' model. Since our objective is parsimonious modelling of the covariance structure, we will rely more on the SBC than the AIC criterion.

AIC and SBC values for the six covariance structures are shown in Table III. `Unstructured', has the largest AIC, but AR(1)+RE, `autoregressive with random effect for patient', has the largest SBC. Toeplitz ranks second in both AIC and SBC. The discrepancy between AIC and SBC for the UN structure reflects the penalty for the large number of parameters in the UN covariance matrix. Based on inspection of the correlation estimates in Tables I and III, the graphs of Figure 5, and the relative values of SBC, we conclude that AR(1)+RE, `autoregressive with random effect for patient', is the best choice of covariance structure.

7. EFFECTS OF COVARIANCE STRUCTURE ON TESTS OF FIXED EFFECTS, ESTIMATES OF FIXED EFFECTS AND STANDARD ERRORS OF ESTIMATES

In Section 6 we compared the correlation and covariance matrices produced by five choices of covariance structure. In this section we examine the effects of choices of covariance structure on tests and estimates of fixed effects. First, we examine the table of tests for fixed effects specified in the MODEL statements. Then we select a set of 15 comparisons among means and use the ES-

TIMATE statement to illustrate effects of covariance structure on estimates of linear combinations of fixed effects.

Table IV contains values of F tests for fixed effects that are computed by the MIXED procedure for each of the covariance structures specified in (8), (10), (11), (13), (14) and (15). The F values differ substantially for SIM, CS and AR(1) structures. These are the structures that did not provide good fits in Section 6. Failure of SIM to recognize between-patient variation results in the excessively large F values for BASEFEV1 and DRUG, which are between patient effects. Using CS structure produces essentially the same results that would be obtained by using a univariate split-plot type analysis of variance (Milliken and Johnson, reference [7], chapter 26). It results in excessively large F values for HR and DRUG_ HR. This is a well-known phenomenon of

Table IV. Values of *F* tests for fixed effects for six covariance structures.

Structure name	BaseFEV1	DRUG	HR	DRUG * HR
1. Simple	490.76	46.50	9.20	1.69
2. Compound symmetric	76.42	7.24	38.86	7.11
3. Autoregressive (1)	90.39	8.40	7.39	2.46
4. Autoregressive (1) with random effect for patient	75.93	7.28	17.10	3.94
5. Toeplitz (banded)	76.31	7.30	13.75	3.82
6. Unstructured	92.58	7.25	13.72	4.06

performing univariate analysis of variance when CS (actually, Hyunh{Feldt [13]) assumptions are not met. It is basically the reason for making the so-called Hyunh{Feldt [13] and Greenhouse {Geisser [14] adjustments to ANOVA p-values as done by the REPEATED statement in PROC GLM [15]. F values for tests of HR and DRUG_HR using AR(1) structure are excessively small due to the fact that AR(1) underestimates covariances between observations far apart in time, and

thereby overestimates variances of differences between these observations. Results of F tests based on AR(1)+RE, TOEP and UN covariance are similar for all fixed effects. All of these structures are adequate for modelling the covariance, and therefore produce valid estimates of error.

Now, we investigate effects of covariance structure on 15 linear combinations of fixed effects, which are comparisons of means. The first seven comparisons are differences between hour 1 and subsequent hours in drug A; these are within-subject comparisons. In terms of the univariate model (2), they are estimates of

$$\mu_{\mathrm{A}.1} - \mu_{\mathrm{A}.k} = \tau_1 - \tau_k + (\alpha\tau)_{\mathrm{A}1} - (\alpha\tau)_{\mathrm{A}k} \qquad (16)$$

for k =2; : : : ; 8.

The next eight comparisons are differences between drugs A and B at hours 1 to 8; these are between-subject comparisons at particular times. In terms of the univariate model (3), they are estimates of

$$\mu_{\mathrm{A}.k} - \mu_{\mathrm{B}.k} = \lambda(\bar{X}_{\mathrm{A}} - \bar{X}_{\mathrm{B}}) + \alpha_{\mathrm{A}} - \alpha_{\mathrm{B}} + (\alpha\tau)_{\mathrm{A}k} - (\alpha\tau)_{\mathrm{B}k} \qquad (17)$$

for k =1; : : : ; 8.

The ESTIMATE statement in the MIXED procedure can be used to compute estimates of linear combinations of fixed effect parameters. It is used for this purpose in essentially the same manner as with the GLM procedure. With MIXED, the ESTIMATE statement can be used for the more general purpose of computing estimates of linear combinations of fixed and random effectes, known as Best Linear Unbiased Predictors (BLUPs) [16].

The following ESTIMATE statements in (18) can be run in conjunction with the PROC MIXED statements (7)-(15) to obtain estimates of the differences (16). Coeffcients following `hr' in (18) specify coeffcients of Tk parameters in (16), and coeffcients following `drug * hr' in (18) specify coeffcients of (α T)ik parameters in (16):

estimate `hr1{hr2-drgA' hr 1 −1 0 0 0 0 0 0 drug _ hr 1 −1 0 0 0 0 0 0;

estimate `hr1{hr3-drgA' hr 1 0 −1 0 0 0 0 0 drug _ hr 1 0 −1 0 0 0 0 0;

Table V. Estimates and standard errors for six covariance structures: within-subject comparisons across time.

		Standard errors					
Parameter	Estimate*	Simple	CS	AR(1)	AR(1)+RE	Toeplitz (banded)	Unstructured
hr1–hr2 drug A	0.0767	0.1491	0.0725	0.0564	0.0564	0.0562	0.0470
hr1–hr3 drug A	0.2896	0.1491	0.0725	0.0769	0.0700	0.0647	0.0492
hr1–hr4 drug A	0.4271	0.1491	0.0725	0.0908	0.0764	0.0704	0.0698
hr1–hr5 drug A	0.4200	0.1491	0.0725	0.1012	0.0796	0.0794	0.0822
hr1–hr6 drug A	0.4942	0.1491	0.0725	0.1093	0.0813	0.0836	0.0811
hr1–hr7 drug A	0.6050	0.1491	0.0725	0.1158	0.0822	0.0900	0.1002
hr1–hr8 drug A	0.6154	0.1491	0.0725	0.1211	0.0827	0.0951	0.0888

*Parameter estimates are the same regardless of variance structure for these contrasts.

estimate `hr1{hr4-drgA' hr 1 0 0 −1 0 0 0 0 drug _hr 1 0 0 −1 0 0 0 0;

estimate `hr1{hr8-drgA' hr 1 0 0 0 −1 0 0 0 drug _hr 1 0 0 0 −1 0 0 0;

estimate `hr1{hr5-drgA' hr 1 0 0 0 0 − 1 0 0 drug _hr 1 0 0 0 0 −1 0 0;

estimate `hr1{hr6-drgA' hr 1 0 0 0 0 0 − 1 0 drug _hr 1 0 0 0 0 0 −1 0;

estimate `hr1{hr7-drgA' hr 1 0 0 0 0 0 0 − 1 drug _hr 1 0 0 0 0 0 0 − 1;

Results from running these ESTIMATE statements with each of the six covariance structures in (18) appear in Table V. The estimates obtained from (18) are simply differences between the two drug A means for each pair of hours, that is, the estimate labelled `hr1-hrk drgA' is $\bar{Y}_{A1} - \bar{Y}_{Ak}$, , or in terms of the model (2), $\tau_1 - \tau_k + (\alpha\tau)_{A1} - (\alpha\tau)_{Ak} + \bar{e}_{A1} - \bar{e}_{Ak}$, for $k = 2, \ldots, 8$. Because the covariable BASEFEV1 is a subject-level covariate, it cancels in this

comparison. Consequently, the estimates are all the same for any covariance structure due to the equivalence of generalized least squares (GLS) and ordinary least squares (OLS) in this setting. This will not happen in all cases, such as when the data are unbalanced, when the covariate is time-varying, or when polynomial trends are used to model time effects. In this example, the data are balanced and hour is treated as a discrete factor. See Puntanen and Styan [17] for general conditions when GLS estimates are equal to OLS estimates.

Even though all estimates of differences from statements (18) are equal, each of the six co-variance structures results in a different standard error estimate (Table V). Note that the `simple' standard error estimates are always larger than those from the mixed model. The general expression for the variance of the standard error estimate is

$$V(Y_{\mathrm{A}.1} - Y_{\mathrm{A}.k}) = [\sigma_{1,1} + \sigma_{k,k} - 2\sigma_{1,k}]/24 \qquad (19)$$

Where $\sigma_{k,l} = \mathrm{cov}(Y_{ijk}, Y_{ijl})$. For structured covariances, $\sigma_{k,l}$ will be a function of k,l, and a small number of parameters.

Standard error estimates printed by PROC MIXED are square roots of (19), with _k; l expressions replaced by their respective estimates, assuming a particular covariance structure. We now discuss effects of the assumed covariance structure on the standard error estimates.

Structure number 1, `simple', treats the data as if all observations are independent with the same variance. This results in equal standard error estimates of

$$0.14909825 = (2\,\hat{\sigma}^2_{\mathrm{SIM}}/24)^{1/2}$$
$$= (2(0.267/24))^{1/2}$$

for all differences between time means in the same drug. These are incorrect because SIM structure clearly is inappropriate for two reasons. First, the SIM structure does not accommodate between- patient variation, and second, it does not recognize that measures close together in time are more highly correlated than measures far apart in time.

Structure number 2, `compound symmetric', acknowledges variation as coming from two sources, between- and within-patient. This results in standard error estimates of

$$0.07252978 = (2\,\hat{\sigma}^2_{\mathrm{CS,w}}/24)^{1/2}$$
$$= (2(0.063/24))^{1/2}$$

being functions only of the within-patient varience component estimate. However, compound sym-metry does not accommodate different standard errors of differences between times as being de- pendent on the length of the time interval. Consequently, the standard error estimates based on the compound symmetric structure also are invalid.

Structure number 3, `autoregressive', results in standard errors of estimates of differences between times which depend on the length of the time interval. For example, the standard error estimate for the difference between hours 1 and 8 (lag=7) is

$$0.121 = (2\,\hat{\sigma}^2_{\mathrm{AR(1)}}(1 - \hat{\rho}^7_{\mathrm{AR(1)}})/24))^{1/2}$$
$$= (2(0.266(1 - 0.856^7)/24))^{1/2}$$

and similarly for other lags. The standard error estimates are 0.121 for the difference between hours 1 and 8 etc., down to 0.056 for the difference between hours 1 and 2. If the autoregressive structure were correct, then these estimates of standard errors should be in good agreement with those produced by TOEP covariance. The TOEP standard error estimates range from 0.095 for the difference between hours 1 and 8 down to 0.056 for the difference between hours 1 and 2. Thus the autoregressive estimates are too large by approximately 30 per cent for long time intervals (for example, hours 1 to 8). This is because the autoregressive structure underestimates the correlation between observations far apart in time by forcing the correlation to decrease exponentially toward zero.

Next, we examine the standard errors provided by structure 4, `autoregressive with random effect for patient'. The standard error estimate for the difference between hours 1 and 8 (lag = 7) is

$$\begin{aligned} 0.083 &= (2(\hat{\sigma}^2_{\mathrm{AR(1)+RE,w}}(1 - \hat{\rho}^7_{\mathrm{AR(1)+RE}})/24))^{1/2} \\ &= (2(0.083(1 - 0.540^7)/24))^{1/2} \end{aligned}$$

and similarly for other lags. We see that these standard error estimates generally provide good agreement with the TOEP and UN standard error estimates. These three structures (TOEP, UN and AR(1)+RE) are all potential candidates, because they accommodate between-subject variance and decreasing correlation as the lag increases. The intuitive advantage of the AR(1)+RE estimates over the TOEP and UN estimates in this setting is that the standard errors of the AR(1)+RE estimates follow a smooth trend as a function of lag, whereas the TOEP and UN standard error estimates are more erratic, particularly so for the UN estimates. In all three

strucutres, the standard errors for the larger time lags are larger than those for the smaller lags, reacting the pattern seen in the data.

The following ESTIMATE statements can be run in conjunction with PROC MIXED statements (7){(15) to obtain estimates of the differences between drugs A and B at each hour, defined

Table VI. Estimates and standard errors for six covariance structures: between-subject comparisons.

Parameter	Simple	CS	AR(1)	AR(1)+RE	Toeplitz (Banded)	Unstructured
Estimates						
drg B–drg A hr1	0.2184	0.2184	0.2179	0.2182	0.2180	0.2188
drg B–drg A hr2	0.2305	0.2305	0.2300	0.2303	0.2301	0.2308
drg B–drg A hr3	0.3943	0.3943	0.3938	0.3941	0.3939	0.3946
drg B–drg A hr4	0.3980	0.3981	0.3975	0.3978	0.3976	0.3984
drg B–drg A hr5	0.1968	0.1968	0.1963	0.1966	0.1964	0.1971
drg B–drg A hr6	0.1068	0.1068	0.1063	0.1066	0.1064	0.1071
drg B–drg A hr7	0.1093	0.1093	0.1088	0.1091	0.1088	0.1096
drg B–drg A hr8	0.1530	0.1530	0.1525	0.1528	0.1526	0.1534
Standard errors						
drg B–drg A hr1	0.1491	0.1499	0.1489	0.1494	0.1490	0.1374
drg B–drg A hr2	0.1491	0.1499	0.1489	0.1494	0.1490	0.1471
drg B–drg A hr3	0.1491	0.1499	0.1489	0.1494	0.1490	0.1454
drg B–drg A hr4	0.1491	0.1499	0.1489	0.1494	0.1490	0.1578
drg B–drg A hr5	0.1491	0.1499	0.1489	0.1494	0.1490	0.1544
drg B–drg A hr6	0.1491	0.1499	0.1489	0.1494	0.1490	0.1465
drg B–drg A hr7	0.1491	0.1499	0.1489	0.1494	0.1490	0.1501
drg B–drg A hr8	0.1491	0.1499	0.1489	0.1494	0.1490	0.1579

In (17):

```
estimate 'drgA–drgB hr1' drug 1 –1 0 drug*hr 1 0 0 0 0 0 0 0 –1 0 0 0 0 0 0 0;
estimate 'drgA–drgB hr2' drug 1 –1 0 drug*hr 0 1 0 0 0 0 0 0 0 –1 0 0 0 0 0 0;
estimate 'drgA–drgB hr3' drug 1 –1 0 drug*hr 0 0 1 0 0 0 0 0 0 0 –1 0 0 0 0 0;
estimate 'drgA–drgB hr4' drug 1 –1 0 drug*hr 0 0 0 1 0 0 0 0 0 0 0 –1 0 0 0 0;
estimate 'drgA–drgB hr5' drug 1 –1 0 drug*hr 0 0 0 0 1 0 0 0 0 0 0 0 –1 0 0 0;
estimate 'drgA–drgB hr6' drug 1 –1 0 drug*hr 0 0 0 0 0 1 0 0 0 0 0 0 0 –1 0 0;
estimate 'drgA–drgB hr7' drug 1 –1 0 drug*hr 0 0 0 0 0 0 1 0 0 0 0 0 0 0 –1 0;
estimate 'drgA–drgB hr8' drug 1 –1 0 drug*hr 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 –1;
```

(20)

Results appear in Table VI. These estimates are the same for structures SIM and CS. They are simply differences between ordinary least squares means, adjusted for the covariable BASEFEV1. However, a simple expression for the variance of the estimates is not easily available. The standard errors differ for the two covariance structures, because simple structure does not recognize between-patient variation.

Estimates of drug differences for the four covariance structures other than `Simple' and `Com-pound symmetric' are all numerically different, though similar. Also, standard errors of the drug differences are not the same for covariance structures AR(1), AR(1)+RE, TOEP and UN, but the standard errors for the AR(1)+RE and TOEP structure are constant over the hours. This is de-sirable, because data variance are homogeneous over hours, and the adjustment for the covariable BASEFEV1 would be the same at each hour. However, the standard errors of drug differences for UN covariance vacillate between 0.137 and 0.158, a range of approximately 16 per cent. The standard errors are not constant because UN does not assume homogeneous variances. In the present example, it is reasonable to assume homogeneous variances, and this should be exploited. Not doing so results in variable and inefficient standard error estimates.

The purpose of this section was to illustrate the practical effects of choosing a covariance structure. The results show that SIM, CS and AR(1) covariance structures are inadequate for the example data. These structure models basically provide ill-fitting estimates of the true covariance matrix of the data. In turn, the ill-fitting estimates of data covariance result in poor estimates of standard

errors of certain differences between means, even if estimates of differences between means are equal across covariance structures. The structures AR(1)+RE, TOEP and UN are adequate, in the sense that they provide good fits to the data covariance. (This is always true of UN because there are no constraints to impose lack of fit.) These adequate structures incorporate the two essential features of the data covariance. One, observations on the same patient are correlated, and two, observations on the same patient taken close in time are more highly correlated than observations taken far apart in time. As a result, standard error estimates based on assumptions of AR(1)+RE, TOEP or UN covariance structures are valid, but because UN imposes no constraints or patterns, the standard error estimates are somewhat unstable.

8. MODELLING POLYNOMIAL TRENDS OVER TIME

Previous analyses have treated hour as a classification variable and not modelled FEV1 trends as a continuous function of hour. In Section 6, we fitted six covariance structures to the FEV1 data, and determined that AR(1)+RE provided the best _t. In Section 7, we examined effects of covariance structure on estimates of fixed effect parameters and standard errors. In this section, we treat hour as a continuous variable and model hour effects in polynomials to refine the fixed effects portion of the model. Then we use the polynomial model to compute estimates of differences analogous to those in Section 7.

Statements (21) fit the general linear mixed model using AR(1)+RE covariance structure to model random effects and third degree polynomials to

model fixed effects of drug and hour. A previous analysis (not shown) that fitted fourth degree polynomials using PROC MIXED showed no significant evidence of fourth degree terms.

```
proc mixed data=fev1uni; class drug patient;
model fev1=basefev1 drug drug * hr drug * hr * hr drug * hr * hr *
hr/htype=1 3
solution noint;                                                    (21)
random patient(drug);
repeated/type=ar(1) sub=patient(drug);
```

The MODEL statement in (21) is specified so that parameter estimates obtained from the SOLUTION option directly provide the coefficients of the third degree polynomials for each drug. The fitted polynomial equations, after inserting the overall average value of 2.6493 for BASEFEV1, are

A: FEV1=3:6187 − 0:1475 HR + 0:0034 HR2 + 0:0004 HR3

B: FEV1=3:5793 + 0:1806 HR − 0:0802 HR2 + 0:0061 HR3

P: FEV1=2:7355 + 0:1214 HR − 0:0289 HR2 + 0:0017 HR3

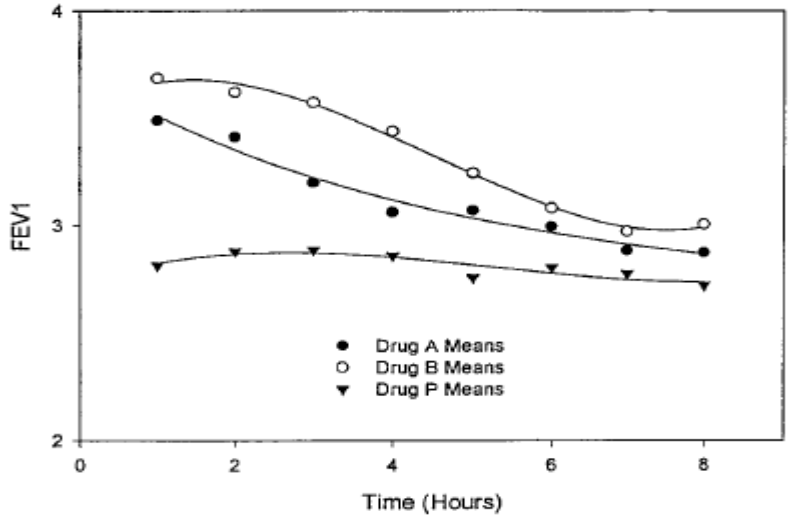


Figure 6. Plots of polynomial trends over hours for each drug.

The polynomial curves for the drugs are plotted in Figure 6.

Estimates of differences between hour 1 and subsequent hours in drug A based on the fitted polynomials may be obtained from the ESTIMATE statements (22):

```
estimate 'hr1–hr2 drga'drug * hr –1 drug * hr * hr –03 drug * hr * hr * hr –007;
estimate 'hr1–hr3 drga'drug * hr –2 drug * hr * hr –08 drug * hr * hr * hr –026;
estimate 'hr1–hr4 drga'drug * hr –3 drug * hr * hr –15 drug * hr * hr * hr –063;
estimate 'hr1–hr5 drga'drug * hr –4 drug * hr * hr –24 drug * hr * hr * hr –124;
estimate 'hr1–hr6 drga'drug * hr –5 drug * hr * hr –35 drug * hr * hr * hr –215;
estimate 'hr1–hr7 drga'drug * hr –6 drug * hr * hr –48 drug * hr * hr * hr –342;
estimate 'hr1–hr8 drga'drug * hr –7 drug * hr * hr –63 drug * hr * hr * hr –511;
```

(22)

Results from statements (22) appear in Table VII.

We see that standard errors of differences between hour 1 and subsequent hours in drug A using AR(1)+RE covariance and polynomial trends for hour are smaller than corresponding standard errors in Table V using AR(1)+RE covariance and hour as a classification variable. This is due to the use of the polynomial model which exploits the continuous trend over hours. If the polynomial model yields very different results, one would conclude it does not adequately represent the trend over time.

```
estimate 'drga – drgb hr1' drug 1 –1 0
      drug * hr 1 –1 0 drug * hr * hr 1 –1 0 drug * hr * hr * hr 1 –1 0;
estimate 'drga – drgb hr2' drug 1 –1 0
      drug * hr 2 –2 0 drug * hr * hr 4 –4 0 drug * hr * hr * hr 8 –8 0;
```

Table VII. Estimates and standard errors for AR(1) + RE covariance structure and third degree polynomial model for hour.m

Parameter	Estimate	Standard error
Within-subject comparisons		
hr1–hr2 drug A	0.1346	0.0453
hr1–hr3 drug A	0.2577	0.0634
hr1–hr4 drug A	0.3669	0.0686
hr1–hr5 drug A	0.4599	0.0720
hr1–hr6 drug A	0.5344	0.0754
hr1–hr7 drug A	0.5880	0.0753
hr1–hr8 drug A	0.6183	0.0828
Between-subject comparisons		
drg B–drg A hr1	0.2108	0.1494
drg B–drg A hr2	0.3280	0.1429
drg B–drg A hr3	0.3463	0.1434
drg B–drg A hr4	0.2998	0.1408
drg B–drg A hr5	0.2228	0.1408
drg B–drg A hr6	0.1492	0.1434
drg B–drg A hr7	0.1132	0.1429
drg B–drg A hr8	0.1489	0.1494

```
estimate 'drga – drgb hr3' drug 1 – 1 0
      drug * hr 3 – 3 0 drug * hr * hr 9 – 9 0 drug * hr * hr * hr 27 – 27 0;
estimate 'drga – drgb hr4' drug 1 – 1 0
      drug * hr 4 – 4 0 drug * hr * hr 16 – 16 0 drug * hr * hr * hr 64 – 64 0;
estimate 'drga – drgb hr5' drug 1 – 1 0
      drug * hr 5 – 5 0 drug * hr * hr 25 – 25 0 drug * hr * hr * hr 125 – 125 0;
estimate 'drga – drgb hr6' drug 1 – 1 0
      drug * hr 6 – 6 0 drug * hr * hr 36 – 36 0 drug * hr * hr * hr 216 – 216 0;
estimate 'drga – drgb hr7' drug 1 – 1 0
      drug * hr 7 – 7 0 drug * hr * hr 49 – 49 0 drug * hr * hr * hr 343 – 343 0;
estimate 'drga – drgb hr8' drug 1 – 1 0
      drug * hr 8 – 8 0 drug * hr * hr 64 – 64 0 drug * hr * hr * hr 512 – 512 0;
```
(23)

Results from statements (23) appear in Table VII.

Standard errors for differences between drug A and drug B at hours 1 and 8 using the polynomial model are similar to standard errors for these differences using the model with hour as a classification variable. The standard errors of

differences between drugs A and B at intermediate hours are less than the standard errors for respective differences using hour as a classification variable. Again, this is a phenomenon related to using regression models, and has very little to do with the covariance structure. It demonstrates that there is considerable advantage to refining the fixed effects portion of the model. We believe, however, that refining the fixed effects portion of the model should be done after arriving at a satisfactory covariance structure using a saturated fixed effects model.

9. SUMMARY AND CONCLUSIONS

One of the primary distinguishing features of analysis of repeated measures data is the need to accommodate the covariation of the measures on the same sampling unit. Modern statistical software enables the user to incorporate the covariance structure into the statistical model. This should be done at a stage prior to the inferential stage of the analysis. Choice of covariance structure can utilize graphical techniques, numerical comparisons of covariance estimates, and indices of goodness-of-_t. After covariance is satisfactorily modelled, the estimated covariance matrix is used to compute generalized least squares estimates of fixed effects of treatments and time.

In most repeated measures settings there are two aspects to the covariance structure. First is the covariance structure induced by the subject experimental design, that is, the manner in which subjects are assigned to treatment groups. The design typically induces covariance due to contribution of random effects. In the example of this paper, the design was completely randomized which results in

covariance of observations on the same subject due to between-subject variation. If the design were randomized blocks, then there would be additional covariance due to block variation. When using SAS PROC MIXED, the covariance structure induced by the subject experimental design is usually specified in the RANDOM statement. Second is the covariance structure induced by the phenomenon that measures close in time are more highly correlated than measures far apart in time. In many cases this can be described by a mathematical function of time lag between measures. This aspect of covariance structure must be modelled using the REPEATED statement in PROC MIXED.

Estimates of fixed effects, such as differences between treatment means, may be the same for different covariance structures, but standard errors of these estimates can still be substantially different. Thus, it is important to model the covariance structure even in conditions when estimates of fixed effects do not depend on the covariance structure. Likewise, tests of significance may depend on covariance structure even when estimates of fixed effects do not.

The example in the present paper has equal numbers of subjects per treatment and no missing data for any subject. Having equal numbers of subjects per treatment is not particularly important as far as implementation of data analysis is concerned using mixed model technology. However, missing data within subjects can present serious problems depending on the amount, cause and pattern of missing data. In some cases, missing data can cause non-estimability of fixed effect parameters. This would occur in the extreme situation of all subjects in a particular treatment having missing data at the same time point. Missing data can

also result in unstable estimates of variance and covariance parameters, though non-estimability is unlikely. The analyst must also address the underlying causes of missing data to assess the potential for introducing bias into the estimates. If the treatment is so toxic as to cause elimination of study subjects, ignoring that cause of missingness would lead to erroneous conclusions about the efficacy of the treatment. For more information on this topic, the reader is referred to Little and Rubin [18], who describe different severity levels of missingness and modelling approaches to address it.

Unequal spacing of observation times presents no conceptual problems in data analysis, but computation may be more complex. In terms of PROC MIXED, the user may have to resort to the class of covariance structures for spatial data to implement autoregressive covariance. See Littell et al. [15] for illustration.

Using regression curves to model mean response as functions of time can greatly decrease standard errors of estimators of treatment means and differences between treatment means at particular times. This is true in any modelling situation involving a continuous variable, and is not related particularly to repeated measures data. This was demonstrated in Section 8 using polynomials to model FEV1 trends over time. In an actual data analysis application, pharmacokinetic models could be used instead. Such models usually are non-linear in the parameters, and thus PROC MIXED could not be used in its usual form. However, the NLINMIX macro or the new NLMIXED procedure could be used.

The general linear mixed model specifies that the data vector Y is represented by the equation…

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZU} + \mathbf{e} \tag{24}$$

Where $E(\mathbf{U}) = 0$, $E(\mathbf{e}) = 0$, $V(\mathbf{U}) = \mathbf{G}$ and $V(\mathbf{e}) = \mathbf{R}$. Thus

$$E(\mathbf{Y}) = \mathbf{X}\beta \tag{25}$$

We assume that U and e are independent, and obtain

$$V(\mathbf{Y}) = \mathbf{ZGZ}' + \mathbf{R} \tag{26}$$

Thus, the general linear mixed model specifies that the data vector Y has a multivariate normal distribution with mean vector $\mu = \mathbf{X}\beta$ and covariance metrix $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$.

Generalized least squares theory (Graybill, Reference [19], Chapter 6) states that the best linear unbiased estimate of β is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{27}$$

and the covariance matrix of the sampling distribution of b is

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{28}$$

The BLUE of a linear combination $\mathbf{a}'\beta$ is $\mathbf{a}'\mathbf{b}$, and its variance is $\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}$. More generally, the BLUE of a set of linear combinations $\mathbf{A}'\beta$ is $\mathbf{A}'\mathbf{b}$, and its sampling distribution covariance matrix is $\mathbf{A}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}$. Thus, the sampling distribution of $\mathbf{A}'\mathbf{b}$ is multivariate normal with mean vector $E(\mathbf{A}'\mathbf{b}) = \mathbf{A}'\beta$ and covariance matrix $\mathbf{A}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}$. Inference procedures for the general linear mixed model are based on these principles. However, the estimate $\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ and its covariance matrix $V(\mathbf{b}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ both are

functions, of $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, and in most all cases **V** will contain unknown parameters. Thus, an estimate of **V** must be used in its place.

Usually, elements of **G** will be functions of one set of parameters, and elements of **R** will be functions of another set. The MIXED procedure estimates the parameters of **G** and **R**, using by default the REML method, or the ML method, if requested by the user. Estimates of the parameters are then inserted into **G** and **R** in place of the true parameter values to obtain $\hat{\mathbf{V}}$. In turn, $\hat{\mathbf{V}}$ is used in place of **V** to compute **B** and $\hat{\mathbf{V}}(\hat{\mathbf{b}})$.

Standard errors of estimates of linear combinations are computed as $(\hat{\mathbf{V}}(\mathbf{a}'\hat{\mathbf{b}}))^{1/2} = (\mathbf{a}'(\hat{\mathbf{V}}(\hat{\mathbf{b}}))\mathbf{a})^{1/2}$. Statistics for tests of fixed effects are computed as $F = \mathbf{b}'\mathbf{A}(\hat{\mathbf{V}}(\hat{\mathbf{b}}))^{-1}\mathbf{A}'\mathbf{b}/\mathrm{rank}(\mathbf{A})$. In some cases, the distributions of **F** are, in fact, **F** distributions, and in other cases they are only approximate. Degrees of freedom for the numerator of the F statistic are given by the rank of A, but computation of degrees of freedom for the denominator is a much more difficult problem. One possibility is a generalized Satterthwaite approximation as given by Fai and Cornelius [20]. The interested reader is also referred to McLean and Sanders [21] for further discussion on approximating degrees of freedom, and to Hulting and Harville [22] for some Bayesian and non-Bayesian perspectives on this issue. For more information on analysis of repeated measures data, see Diggle et al. [23] and Verbeke and Molenberghs [24].

Chapter – 5

Covariance Models for Latent Structure in Longitudinal Data

Chapter – 5
Covariance Models for Latent Structure in Longitudinal Data

I present several approaches to modeling latent structure in longitudinal studies when the covariance itself is the primary focus of the analysis. This is a departure from much of the work on longitudinal data analysis, in which attention is focused solely on the cross-sectional mean and the influence of covariates on the mean. Such analyses are particularly important in policy-related studies, in which the heterogeneity of the population is of interest. We describe several traditional approaches to this modeling and introduce a flexible, parsimonious class of covariance models appropriate to such analyses. This class, while rooted in the tradition of mixed effects and random coefficient models, merges several disparate modeling philosophies into what we view as a hybrid approach to longitudinal data modeling. We discuss the implications of this approach and its alternatives especially on model interpretation. We compare several implementations of this class to more commonly employed mixed effects models to describe the strengths and limitations of each. These alternatives are compared in an application to long-term trends in wage inequality for young workers. The findings provide additional guidance for the model formulation process in both statistical and substantive senses.

5.1 Introduction and Motivation

An increasing number of social and behavioral science studies collect information from subjects at several points in time. These longitudinal studies enable researchers to study changes in the phenomena of interest over the life-course of the subjects. At each observation time, at least one response, such as wages earned or the occurrence of a meaningful event, such as graduation from college, is recorded. As with regression, one may collect explanatory covariates in the hope that differences in these inputs will be associated with different levels of response. Each subject is thus associated with their own time series of responses and a corresponding set of potentially time-varying explanatory covariates. Models for longitudinal data attempt to relate those individual time series to an overall group process.

The focus on either individual or group processes plays a key role in how one models longitudinal data. For example, if we are modeling a continuous response, Yi, in terms of explanatory covariates, Xi, then the familiar linear model, for individual i,

$$Y_i = X_i\beta + \epsilon_i \qquad (1)$$

could be adopted, but Yi and would be n_i-vectors, where n_i is the number of observations on individual i. Similarly, Xi would be of dimension n_i * p, where p is the number of explanatory covariates. Note that we are modeling a response vector, yet this distinction is not made explicitly with our notation. Alternatively, the model may be writte.

$$Y_i(t) = X_i(t)\beta + \epsilon_i(t),$$

where the index t identifies a specific element of the response vector Yi. If we were to proceed with a classical multiple regression, stacking the responses by individual and then by observation within individual, we would obscure an important feature of longitudinal data; namely, we know that some set of observations come from the same individual. And observations within the same individual may be correlated due to unobserved individual characteristics. To see this, let us return to the notation in (1), but now

$$Y_i = X_i\beta + \alpha_i + \epsilon_i^*, \qquad (2)$$

where i is an unobserved scalar trait for subject i, and the residual variation is mean zero and uncorrelated with the unobserved process, and 0 for t 6= t0. Since we do not observe i, we tend to use (1) when the underlying process is accurately described by (2), so the residual variation structure i is really . The unobserved trait induces a correlation within individual i, since

$$E(\epsilon_i(t)\epsilon_i(t')|\alpha_i) = E((\alpha_i + \epsilon_i^*(t))(\alpha_i + \epsilon_i^*(t'))|\alpha_i) = \alpha_i^2,$$

and in general. Note that the unobserved may not correspond to a single measurable characteristic; instead it proxies for all unobserved characteristics.

There are several different ways to think about the correlation structure in longitudinal data. The different perspectives are induced by the nature of the unobserved trait and its relationship to the covariates and residual variation. If substantive interest is on the effects of the covariates on the response averaged over the population then models are usually formulated for the mean response averaged over the unobserved traits. Broadly speaking, the correlation structure is

modeled as a nuisance parameter, and regression coefficients represent population-average effects (Liang and Zeger , Zeger and Liang , Prentice). Alternatively, individual differences may be of interest, and can be modeled directly as latent variables; for these individual-specific models, regression parameters are to be interpreted conditionally on the value of the subject's latent variable. We will discuss these different approaches, certain variations thereof, and their implications in subsequent sections. Along the way, we will introduce a class of models for longitudinal data that merges these two approaches in a new hybrid form, which is conceptually linked to principal components and factor analysis. First, we introduce the substantive problem that motivates this new formulation.

In labor market economics, a rise in cross-sectional measures of wage inequality that began in the 1970s and has persisted into the 1990s is well-documented (Levy and Murnane 1992; Danziger and Gottschalk 1993; McMurrer and Sawhill 1998). This means that there are greater numbers of workers making more and less than ever before. And for many groups of workers, wages have remained stagnant over time. This stagnation is due in part to a disproportionate growth in the lower tail of the wage distribution. Using data from two young adult cohorts in the National Longitudinal Survey (NLS), we find, for example, that 30-35 year old white men have a mean wage of $17.78 in 1979, while this figure is $14.27 per hour for a similarly aged group in 1992 (inflation adjusted, 1999 dollars). A measure of inequality is the variance in outcomes; the variance of the logged wages increased 44% over the same period.

This dramatic rise has prompted researchers to look more closely at trends in inequality over the life course of a worker. Cross-sectional data can document a rise in inequality, but since each cross-section is a random sample from the population, one cannot conclude that the same people are making the higher wages in each period. Statements such as the rich are getting richer while the poor are getting poorer" cannot be definitively made. But longitudinal data can be used to address this type of question. To couch this in labor economic terms, we would like to examine two competing hypotheses that explain the growth in inequality:

I. Wages have become more volatile.

II. Wages have become more stratified over time, indicating a reduction in economic mobility.

The scenarios are illustrated in Figures 1 - 3 below. Figure 1, represents an economy in which individual "profiles" fan out over time, but not excessively. This is our stylized image of a past economy; in Figures 2 and 3, we changed the covariance structure to reflect at least a doubling of process variance, but we do this in very different ways. In Figure 2, the structured variation has become more stratified, but the residual process is left unchanged. In Figure 3, the structured variation is identical to that used in Figure 1, but the residual variance of the process has been greatly increased. This last figure may seem exaggerated, but the average variation between individuals is actually a bit smaller than in Figure 2.

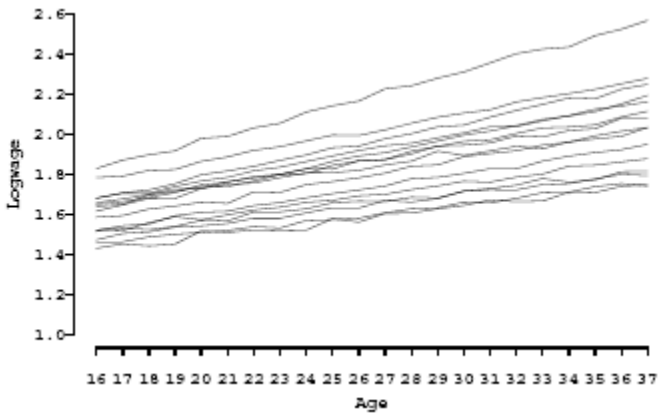


Figure 1: Stylized wage trajectories for a less stratified economy

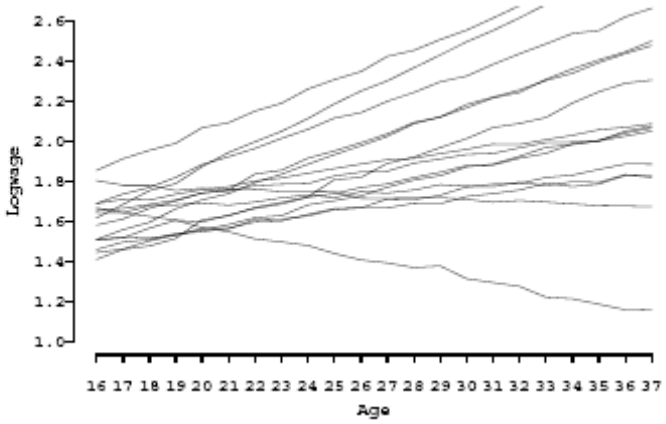


Figure 2: Stylized wage trajectories for a more stratified economy

Based on the figures, the difference between increased stratification (Figure 2) and increased Volatility (Figure 3) seems transparent, but in a real application, both hypotheses

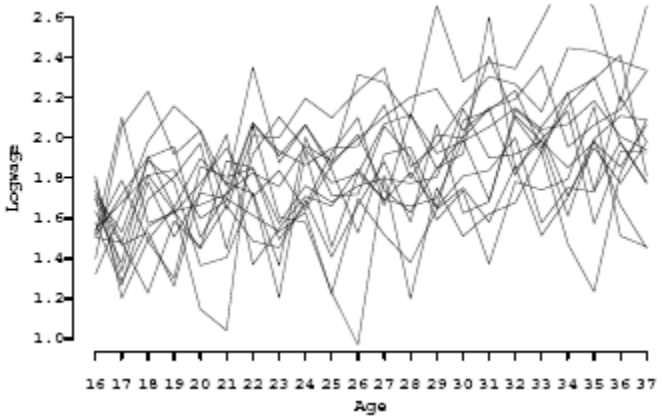


Figure 3: Stylized wage trajectories for a more volatile economy

may be true and the differences may be more subtle. Each possibility has adifferent substantive interpretation, so sorting out the extent to which each hypothesis describes the changes in wage structure is very important and will have different implications in terms of policy.

To investigate the two hypotheses, Bernhardt, et al. (1997), Gottshalk and Moffitt (1994), Haider (1996) and Baker (1997) decompose the wage into permanent and transient components as follows. Let

$$w(t) = p(t) + u(t); \quad \text{.....(3)}$$

where w is the wage, p is its permanent portion, and u is a residual variation term, capturing short-term, or transient variation. For a specific worker, p(t) can be thought of as their mean wage at time t, with residual variation u(t). Assuming independence of p(t) and u(t), we have that

$$Var(w(t)) = Var(p(t)) + Var(u(t));$$

and the two hypotheses can be differentiated through this variance decomposition: a rise in wage variance must involve a rise in at least one of the two variance

components. Greater stratification implies an increase in the first term, while greater volatility involves the second. If we had a substantial number of observations for each individual, we could estimate p(t) using separate regressions for each. The distribution of these predicted curves would represent permanent variation, while the residuals represent transient variation. The hypotheses of interest describe differences in wage trajectories without any socioeconomic controls, so the only explanatory covariates we include in this analysis are functions of time.

This last point warrants further explanation. Socioeconomic variables such as level of schooling, parent's education, and industry of employment, capture expected returns to individual (supply-side) and employer (demand-side) characteristics. For example, there may be changes in the mean return to obtaining a high school degree which reflects the value of that set of skills in the labor market. Including socioeconomic covariates also controls for compositional shifts in the labor market. The growth in a specific sector of the economy could induce growing inequality if that sector is typically associated with lower wages. But all of these explanations are necessarily focused on the permanent portion of the wage trajectory, since volatility is associated with residual, rather than mean effects. The first stage in any analysis of wage inequality is the accurate documentation of the growth in inequality, and how it is apportioned with regard to permanent and transient components. Thus, our focus is first on covariance structure, and not on the socioeconomic covariates that might "explain" the structure.

In many longitudinal studies, there are a relatively small number of observations per individual, so estimating separate regressions to assess wage inequality is infeasible. A variance components model (Searle, et al. 1992) using (3) can partition the variance into long-term, permanent variation and short-term, transitory variation. We note that the distinction between long- and short-term trends is fundamentally about economic mobility. The variation between the individuals' permanent components is a measure of mobility relative to one's peers, and a variance components analysis allows one to evaluate this important economic issue.

In matters with such strong policy implications, proper specification of a model that describes the components of variation is crucial. What may be less apparent is the role of the covariance structure in such an analysis. If we generalize the basic model (2) for longitudinal data to allow for more complex individual characteristics, we get the standard mixed effects model (Diggle, Liang and Zeger 1994),

$$Y_i = X_i\beta + Z_i\delta_i + \epsilon_i. \qquad (4)$$

We have introduced a random effects component, Zi in which Zi is an ni_q known design matrix, and is a q-vector of unknown (latent) variables. It is often assumed that are mean zero multivariate Gaussian. Under this assumption, it is seen that E(Yi) = Xi_; while E(Yij) = Xi_ + Zi. This distinction is important. The latter approach asserts that individuals differ from the population average response in a systematic manner, which is dependent on some latent characteristics. In our application, the growth in wage inequality and evidence for the two competing

hypotheses are features of the latent process, Zi, not the population-average process Xi. If Nq(0;G) and Nni(0;R), independently, then

$$\mathrm{Cov}(Y_i(t), Y_i(t')) = \{Z_i G Z_i' + R\}_{tt'} .$$

Note that the Xi_ term is not included. In other words, the covariance structure of the responses, rather than the mean structure, captures the nature of individual differences, and thus inequality, in the labor market. Strictly speaking, the covariance structure captures all extra-mean variation. Since for the mean we employ only time as an explanatory covariate, all of the contributions to inequality are expressed in the covariance. This establishes a baseline level of inequality that we can later use additional covariates (such as education) to explain.

Models for this covariance structure that can differentiate between hypotheses 1 and 2 will be developed in subsequent sections.

Data

As alluded to above, we will anchor our presentation by using an example from labor market economics, where proper modeling of covariance structure is of paramount importance. We will be investigating two datasets from the NLS. The first, or original cohort, is a representative sample of young men aged 14-21 first interviewed in 1966 and interviewed annually for the next fifteen years (with the exception of 1972, 1974, 1977 and 1979). The second dataset began with a comparable sample of young men in 1979 who have been interviewed yearly since then for fifteen additional years. For comparability between cohorts, we selected only non-Hispanic whites, with resulting sample sizes of 2,614 and 2,373

respectively for the original and recent cohorts. For a detailed description of these datasets and their comparability, see Bernhardt, et al. (1997). According to Topel and Ward (1992), the first 10 years of a career will account for 66 percent of lifetime wage growth for male high school graduates and almost exactly the same fraction of lifetime job changes," so it is important to understand trends manifesting themselves in this early period.

In this paper, we present several different ways to model covariance structure with the ultimate goal of addressing questions such as those posed in hypotheses 1 and 2. One of these methods is novel in the literature, so it is be developed in some depth. We begin by discussing several different philosophical perspectives to longitudinal data modeling in Section 2. To address hypotheses like the ones just presented, we argue that a different modeling philosophy is necessary; we develop a hybrid framework with this in mind in Section 3 and illustrate it in Section 4. We apply more traditional models to our labor market data in Section 5 and discuss the strengths and weaknesses of each approach, including the substantive implications of each choice. Section 6 summarizes the discussion and suggests future directions of research.

5.2 Alternative Modeling Philosophies

The choice of modeling framework should depend on the substantive question of interest. For example, in many medical applications, one may be focused on how a treatment affects the population as a whole. However, if there are potentially serious risks involved in treatment, the distribution of outcomes,

including information about the extremes, may be of importance. Along with many modeling paradigms comes a modeling philosophy, focused on the primary goals of the research. We now describe several philosophies in longitudinal data modeling.

Population-average analysis

Population average models focus on describing the population, rather than individuals within it. Much as in classical regression, the mean response is modeled conditional on the observed covariates. In a linear model, E(Y jX) = X_, the parameter _ describes how changes in the components of X affect the overall population. With longitudinal data, we have seen that the covariance of the responses within an individual influences the response trajectory. For some problems, that covariance structure is effectively a nuisance parameter it must be included in the model but is of no intrinsic interest in and of itself. Generalized Estimating Equations (GEE) is a methodology that produces consistent estimates of population-average parameters even when the covariance structure is misspecified (Liang and Zeger 1986, Zeger and Liang 1986, Prentice 1988). This technique allows one to pursue the population-average approach to modeling, while accounting for the dependencies due to the longitudinal nature of the data. Since the covariance is viewed as secondary, the method does not yield a variance components analysis, which one might use to address our labor market hypotheses, for example.

Individual-specific analysis

Individual-specific effect models consist of two key components: the fixed effects, which capture gross differences between individuals based on differences in their explanatory covariates; and the random effects, which reflect the influence of unobserved covariates. These so-called "unobserved" covariates are just a device to capture unexplained but systematic variation in outcomes. Typically there is no single covariate, such as "motivation," that would replace the individual effects in our model, were we able to measure it. Rather, after controlling for what was measured, some systematic differences between individuals are likely to exist for a variety of reasons. Because we are looking at longitudinal data, we can verify that some differences seem to persist throughout an individual's life course, and that these are not simply random disturbances.

Under model (4), the fixed effects are captured by the Xi_ term, while the random effects are modeled via Zi. The vector is indexed with an i to reflect the fact that every individual is expected to have their own value for this “parameter." These models are also referred to as random coefficient models (Longford 1993) because the coefficient on the Zi terms is allowed to vary. These coefficients introduce extra-mean variation into the response in a systematic manner mediated by the design matrix Zi.

We interpret these models conditional on the individual specific effects, so we are modeling E(YijXi; Zi;), rather than E(YijXi). Using model (4), the interpretation of the fixed effects parameters shifts to the following. Given the individual specific effect i, the expected response for individual i is Xi + Zi We are

making statements about individuals, not populations; the regression coefficients reflect this distinction and should be interpreted in this conditional manner. In the standard linear mixed effects model in which all random components are assumed Gaussian, the distinction between population average and individual specific modeling is more philosophical, as the models and their parameter estimates are identical. This is not the case, for a generalized linear mixed model (GLMM, McCulloch 1997, Hu, et al. 1998, Crouchley and Davies 1999).

What may be less immediately apparent about this shift toward an individual-specific perspective is that the parameters that define the distribution of the effects often represent meaningful components of variation. For example, if is a scalar and Zi is a column of ones, then the individual differences are being modeled as shifts in the intercept. This implies that the differences between individuals are constant over the life course. The variance of the random effect is an important model parameter. If it is large, then large differences between individuals exist and persist throughout the life course; if it is small, they do not. The ability to interpret a variance component in terms of a substantive question is a key feature of individual-specific modeling.

Note that not all mixed effects models are oriented toward meaningful variance components analyses. Beyond the fixed effects, the variation is modeled in the random effects and in the residual variation structure. In model (4), and the residual variation structure, R, can be made arbitrarily complex. There is often a tension, in terms of modeling, between these two components. ARMA models (Box and Jenkins 1976) can capture a substantial portion of the within-individual

correlation, but they do so via parameters that do not take on individual-specific values. For example, the correlation between observations may be given by , but does not vary between individuals. So we know how variation occurs, but we cannot directly use it to position a curve above or below the mean trajectory.

Jones (1990) discusses this model formulation issue by comparing a classical random growth curve model to an AR(1) model for that same structure. He finds that these two approaches typically compete with each other in terms of explaining the variation in the data. We tend to favor models that emphasize structured variation in the Zi term, because these provide direct summaries of differences between individuals.

In sum, mixed effects models may be based on an individual-specific philosophy, but they are not required to do so. Thus, care must be given in the model formulation process as to which philosophical perspective to adopt.

Latent Curve Models

A related, but philosophically different approach to modeling longitudinal trajectories was developed by Meredith and Tisak (1990).2 They outline a framework in which each response is a weighted average of a fixed set of curves:

$$Y_i(t) = \sum_k \omega_{ik}\phi_k(t) + \epsilon_i(t), \qquad (5)$$

where Yi(t) represents the response for the ith individual, !ik is the individual-specific coefffcient associated with the kth latent curve k(t), and i(t) is the residual process. The k capture the shape and magnitude of the variation, and the !ik allow individuals to differ systematically, much in the same way that

random coefficients do in mixed effects models. In the above formulation, the mean process will be a specific weighted sum of the latent curves, but it could just as well be parameterized separately, as in the fixed effects portion of a mixed model. For the remainder of this discussion, we will ignore the mean of the process, or assume it is identically zero.

If the latent curves are known, then the formulation is similar to a mixed effects model. If we stack the _k as columns of a design matrix Z(t) = f(t); then we can estimate a model:

$$Y_i(t) = Z(t)\delta_i + \epsilon_i(t),$$

again, ignoring the mean process. But when model (5) was originally presented, it was assumed that the latent curves were not known and would be estimated directly from the data. With a few additional assumptions, this would be a factor analysis, which is a particular decomposition of the covariance into structured and residual variation. The former are captured in the factor loadings, while the latter are summarized by the specific variances. The large variability inherent in covariance estimation prompted researchers to impose smoothness constraints on the curves (Rice and Silverman 1991). A basic premise of the new model 2 We also refer the reader to Raykov (2000), in which latent curve modeling is developed using the Structural Equation Modeling (SEM) approach. SEM emphasizes covariance structure in the model formulation that we will propose is that smooth latent curves can go a long way toward describing systematic variation in longitudinal data.

In practice, the latent curve model described above cannot be estimated without further assumptions. If the _k are known, then this can be estimated as a random growth curve mixed effects model. If they are left completely unspecified, we have a factor analysis formulation. Both of these model-based approaches avoid some technical problems that arise when an estimate of the full unspecified covariance matrix is required.3 Moving beyond these traditional models will actually open up a whole new way to think about longitudinal data modeling, and we develop this alternative approach at length in Section 3.

Latent Class Models

So far, we have discussed models for which differences between individuals are expressed as an offset from the mean value in shifts that come from a continuous distribution. For example, the random coefficients in mixed effects models come from a multivariate Gaussian distribution, yielding a wide (actually, infinite) variety of outcomes. If the differences "clump" together in a natural way, then it might make sense to restrict the variation to a finite set of possibilities, in which each represents a clump or cluster of similar outcomes in the population. This is the approach taken by latent class analysis (Clogg 1995).4 The analyst divides the population into K distinct classes, and typically any variation that exists within a class is of secondary interest.5 The model can be represented as:

$$Y_i(t) = X_i(t)\beta_k + \epsilon_i(t), \text{ when } C_i = k, \qquad (6)$$

where the random variable Ci captures the latent class membership, and k represents the regression coefficient for class k.6 By allowing the regression coefficients to take on several different values, a set of distinct trajectories can be captured, assuming that the data support them. In formulating such models, one typically models membership in one of the K classes as a random process following a multinomial distribution, so individuals are members of exactly one class. Several features of the population are documented in this approach: the shape of the different trajectories, and the probability of membership in each. Extensions of this approach by Muthen and Shedden (1999) and Roeder, et al. (1999) involve estimating a multinomial “choice" model for class membership. For example, we might assume that membership is based on a multinomial logit model in which some subset of the explanatory covariates play a role:

$$\text{logit}\left[P(C_i = k|X_i)\right] = \theta_k + X_i\beta_k, \quad k = 1, \ldots, K,$$

where Xi are explanatory variables influencing membership and _k are scalars (for identify- ability, we would x 1 = 0 . The full model combines this choice model with a model for the response, conditional on the class membership and the explanatory covariates.

This modeling philosophy focuses on identifying subgroups in the data with similar mean structures. However there are close links to approaches that model the covariance. To see this consider the model as the number of latent classes increases. If there is only one latent class, then this approach is equivalent to regression and is not modeling covariance at all. As the number of classes increases more of the co variation in profiles is attributed to the classification. If

there are a large number of classes then the co variation in the profiles is largely explained by the classification and the within-class variation will be reduced. There is a clear tradeoff between mean and covariance modeling as the number of classes increases. However, the present latent class models do not attempt to identify features shared by the entire population (the features are by definition disjoint), and we do not consider them here.

Discussion

In sum, there are several different ways to think about modeling longitudinal data. One can concentrate on the population average and represent the covariance structure as a foil. Or, one can model individual differences directly by imposing a strict structure on how these differences arise. A relatively general framework is to decompose variation into the sum of curves with different weights. This could be in the form of a factor analysis, but a model based approach to this is preferred over analyses based on the directly estimated covariance matrix. In some instances, one can separate the variation into similar clusters, with an explicit model for how these are determined by explanatory covariates.

All of these are good ideas, depending on the substantive issues to be addressed. We would use the second and third to explicitly model variation in populations that is quite general in form and consider the fourth when natural clusters are apparent.

5.3 A hybrid model

The theoretical approaches of Section 2 each have their value and place. In our labor economic example, we wish to extract permanent and transient variance components. We want the permanent component to reflect features of the whole population while still allowing the expression of individual differences. A modeling class that identifies a common, population level pattern as distinct from short-term effects requires a hybrid modeling philosophy, since both population-average and individual-specific approaches are being employed. When used to address hypotheses 1 and 2, such an approach will provide a highly interpretable and novel variance components decomposition.

The proto-spline model class

In Scott (1998) and Scott and Handcock (2000), we introduced the hybrid proto -spline class of heterogeneity models. Motivated by a longitudinal study of wage growth, we formulated a class of models that capture long- and short-term features of the covariance structure. The models use a latent curve formulation to identify long-term patterns of variation, and they yield a meaningful variance components decomposition. The proto-spline class is distinguished by the data-adaptive manner in which the curves are estimated. We will now describe this class in detail.

The proto-splines class is derived from the model class (5) of Meredith and Tisak (1990),

$$Y_i(t) = \mu_i(t) + \sum_{k=1}^{K} \omega_{ik}\phi_k(t) + \epsilon_i(t). \quad (7)$$

We have added a general mean process and changed the approach to modeling !ik as follows. What was formerly an unconstrained individual specific weight, we will now view as a random coefficient from some known distribution. In addition, we will specify a functional form for the k(t). However, only a functional form and not a specific function is necessary for estimation of the proto-spline class.

We restrict the _k to be orthonormal.7 This parallels the orthogonality employed in principal components analysis and allows us to interpret each curve's contribution as mathematically distinct from the others. We assume that the random coefficients, wk are independent (for different k), which further uncouples the latent curves. This formulation mirrors many psychological and behavioral models in which a response is the combination of several orthogonal shocks to the system. For the proto-spline class, the way the orthogonality of the _k is maintained is a departure from techniques used in principal component analysis and in Rice and Silverman (1991), in that no external constraints are placed on the estimation procedure.

Consider first the case where the stochastic variation can be described by one curve, _1 (this is a single latent curve model). We must specify the functional form of _1, and this is done by choosing an appropriate functional space.8 For example, we can assume that _1 is a cubic spline with knots at four equispaced time points. Cubic splines are smooth functions that have a tremendous degree of

flexibility in terms of the possible set of shapes that they describe (see Green and Silverman 1994 for details). In our theoretical development (Scott and Handcock 2000), we employed the cubic spline function space because of its smoothness features and flexibility, and this is where the “spline" portion of the proto-spline class is derived. We are not restricted to the class of cubic splines; for example, we can specify that 1 has the form of a jump process, well-described by wavelets (Ogden 1996). To keep the discussion on familiar ground, 1 will come from the function space of all quadratic curves for most of the remainder of this paper.

We denote the chosen function space by H, and proceed with the specification of the proto-spline model class. Let 1(t); : : : ; T (t) be an orthogonal basis for H. Then H is a specific linear combination of those bases, just as is an element of a vector space:

$$\phi_1(t) = \sum_{j=1}^{T} \eta_j \psi_j(t), \tag{8}$$

where the j are T non-random parameters that define the curve. If H is smooth, then so is (t). Extending this to the response variable, for the full model, we have

$$\begin{aligned} Y_i(t) &= \mu_i(t) + \omega_{i1}\phi_1(t) + \epsilon_i(t) && (9) \\ &= \mu_i(t) + \omega_{i1}\sum_{j=1}^{T} \eta_j \psi_j(t) + \epsilon_i(t), && (10) \end{aligned}$$

where is a mean zero random coefficient with variance one and Gaussian random variables with variance.

Our model has two variance components, the variance of and the residual variance, and it is a parametric covariance model that defers specification of the

curve 1(t) to the estimation phase. The uncertainty in the form of (t) (until estimation) is a distinguishing feature of proto-spline models. Note that the parameters define the shape of , which is fixed. These parameters do not represent the variance of a random coefficient, but taken as a whole, they determine the magnitude of the random curve and thus the variance in the process.

The uncertainty allowed in the proto-spline class deserves further attention. In a standard mixed effects model, the random effects design matrix is fixed. Systematic variation takes a known form mediated by that design. The proto-spline class is a departure from this paradigm because it allows the shape of the design, given by in our example, to be determined from all of the information in the data. In this sense, the curve is a population-average value the whole population influences its shape, and it can be considered a population “feature." Individual-specific differences are directly modeled using the random coefficient !i1. So this model is a hybrid between population-average and individual specific philosophies and it belongs to the latent curve class of models.

In fact, the only philosophy not employed here is that of latent class modeling. As previously discussed, there is always a tension between modeling the covariance and modeling the mean, and since our emphasis is on covariance modeling, we do not utilize the latent class modeling philosophy, in which mean processes dominate.

Extensions

In this section we extend the development of the single proto-spline model in (7)-(10) to the general multiple proto-spline model. We note that the choice of our function space implies that has a nonparametric interpretation, since it is a curve lying in a potentially smooth, continuous function space. Note that the model is not restricted to the space of any particular functions. Any finite-dimensional function space (or vector space) can be employed. An advantage to the proto-spline class of models is that the bases may be chosen to reflect the form expected in the substantive process without knowing which specific version of that form is present. If we choose models that result in latent curve estimates with a functional interpretation, features such as the derivative become available.

We define the full proto-spline class by extending the single curve example to more than one curve without introducing additional parameters. The main idea is to use only a subset of the T bases j to construct each curve k. For this development, we let H be a basis for cubic splines with an appropriate set of knots. Let be an indexing function defined on the integers 1; : : : ; T, which selects the basis functions used to construct the kth curve. In our simple example, uses all T basis functions, so Tg. We construct latent curves as a deterministically weighted sum of the basis functions specified by the indexing function I so

$$\phi_k(t) = \sum_{j \in \mathcal{I}_k} \eta_j \psi_j(t). \qquad (11)$$

In order to insure orthogonality of the k, the index sets given by Ik must be disjoint. This restriction implies that once we decide to estimate more than one latent curve, the curves are highly constrained elements from the class H, using

only a subset of the bases for each. Since in the theoretical development H was chosen to be the natural cubic splines, we named the resulting proto-splines, because they are partial versions of a full spline t. This method requires T parameters to build all K curves; if we do not normalize the curves, then for identfiability the random coefficients !ik are all presumed to have variance one.

To place this model in the context of those previously developed, we examine it for two extreme cases. First, if K = T, then each proto-spline is just a rescaled version of the basis function. This is essentially the model proposed in Brumback (1996) and Brumback and Rice (1998), although the form of their model was chosen to produce cubic spline predictions for individual curves. If K = 1, then we are estimating a smooth principal component in the presence of noise, and it is constrained to be a natural cubic spline.

A more useful approach is to choose K to be small in relation to T, so that for equalized index sets, T=K bases are available for each latent curve. Equations (7) and (11) still apply, but the “proto-spline" nature of the curve estimates becomes more apparent. This intermediate case is similar to a principal functions analysis (Ramsay and Silverman 1997), in which we expect that most of the variation in the process is captured in a few of the largest principal functions. We are enforcing a small number of these by our choice of K, and we maintain the orthogonality requirement by the way the model is constructed. Note that this model differs from a principal functions analysis in that we can choose our function spaces with substantive features in mind, rather than simple smoothness constraints. We then build our model directly around these structures.

Link to Mixed Effects models

The standard mixed effects model can be expressed in the following form:

$$Y_i(t) = X_i(t)\beta + Z_i(t)\delta_i + \epsilon_i(t). \qquad (12)$$

A key feature of this model is that Xi(t) and Zi(t) are per specified designs. The single latent curve proto-spline model is precisely the above model, with Zi(t) = (t) and only Zi(t) is specified from the data. This illustrates a conceptual distinction between proto-spline models and other mixed effects models.

To explore the conceptual difference, we will consider three random quadratic models. For Model I, we assume that we know the exact quadratic curve that describes the structured co variation about the mean. so Zi(t) is a scalar-valued function describing a particular growth structure. Further, let the random effect, , be a Gaussian random variable with unknown variance (the variance is one of the model's variance components). For a specific individual,

$$Y_i(t) = X_i(t)\beta + (t + \frac{1}{2}t^2)\delta_i + \epsilon_i(t). \qquad (13)$$

Every subject gets some random multiple of the fixed curve t + 1 2 t2.

For Model II we consider a mixed effects model in which each individual has their own quadratic perturbation as follows. Let the elements forming the three columns of Zi(t) be given by the vector , and let be a vector of random coefficients, with a multivariate Gaussian distribution. Then for an individual specific curve,

$$Y_i(t) = X_i(t)\beta + \delta_{1i} + \delta_{2i}t + \delta_{3i}t^2 + \epsilon_i(t). \qquad (14)$$

While this is quite flexible, the variance components analysis requires a full description of the estimated covariance structure of the random effects, which is contained in a 3*3 matrix that includes important covariance as well as variance components. We must use all of this information when describing any variance partitioning.

Model I is highly inflexible in that we must impose an exact form for growth beyond the mean. However, the variance component for is highly interpretable it is the variance of the coefficient of precisely determined shocks to the system, so a larger variance means there is greater dispersion in individual growth, and that all structured growth follows the same form. It would be difficult to make a similar statement about Model II.

For Model III we consider a single latent curve proto-spline model, which offers the interpretability of the simpler model (I), and the flexibility of the more complex model (II). While this might resemble model II, the vector is common to each individual and does not represent individual-specific random effects. Every individual curve has the following form:

$$Y_i(t) = X_i(t)\beta + \delta_i(\eta_1 + \eta_2 t + \eta_3 t^2) + \epsilon_i(t), \qquad (15)$$

with the parameters fixed and identical across individuals; this is a reparameterization of (10) that keeps the notation consistent. Each of these models is different, and we claim that the proto-splines offer an effective compromise between the rigidity and flexibility of Models I and II, respectively, while remaining highly interpretable from a variance components perspective.

To understand the link between proto-splines and other mixed effects models it is important to understand their technical distinctions. Scott and Handcock (1999) discuss estimation for the proto-spline model and show that there is a likelihood-equivalent but non-standard mixed effects model corresponding to the proto-spline class. In this section we describe the ways in which the proto-spline model is non-standard.

For notational convenience we suppress reference to time in the functions, representing j(t) and Zi(t) as j and Zi; respectively. Let Zi = f 1; 2 : : : ; T g be a design matrix constructed using the basis functions for the space H. The coefficients are assumed to be ordered so that if there are K different groups used in the model, with the kth group given as k = (_k1; : : : ; _knk)T , then the coefficients can be stacked into a T*K matrix. The difference between these two models can be understood by examining their representations. Our proto-spline model class is:

$$Y_i = X_i\beta + Z_i\Gamma\delta_i + \epsilon_i, \qquad (16)$$

The likelihood-equivalent mixed effects model is

$$Y_i = X_i\beta + Z_i\delta_i^* + \epsilon_i, \qquad (17)$$

The proto-spline formulation (16) has K random effects, while (17) has T. In (16), the random effect distribution is completely known (N(0; 1)), while in (17) the parameters governing the effects (the k's), must be estimated for us to know the structure of the random effects. In (16), the design represents the latent curve and is estimated, while in (17), the design Zi is prespecified. These distinctions are convenient ways to interpret the components of the models; they have the same

likelihood and set of unknown parameters. In principal, the likelihood equivalence means that any software that can estimate a mixed effects model can be used to estimate the parameters of the proto-spline class. However, the covariance structure associated with model (17) would not be implemented in standard statistical software for mixed effects models, such as SAS PROC MIXED.

While we have some evidence that these are different models, is this really the case? Restricting our attention to the single proto-spline model, formulation (16) contains a scalar random effect, while the vector defined in (17) contains T effects. It is interesting to note that the covariance structure governing is degenerate, since is not positive definite. This does not introduce problems with estimation, however, because the degeneracy is removed in the full likelihood, once residual variation is included. If one examines the structure more closely, it is apparent that each element of the vector is linearly dependent on each of the others, so in essence only one random effect is generated by this covariance structure.12 So the likelihood-equivalent model (17) is a non-standard mixed effects model, which is equivalent to our proto-spline model, even in terms of the observations that would be generated from it, if we consider the limit of its degenerate covariance matrix.

What this means is that our proto-spline formulation effectively “corrects" the degeneracy in (17) by modeling the random effects in a simpler manner, without direct reference to the relationships indicated by the latter model's covariance structure. The k parameters contained in the matrix are essential to each formulation of the model, but should not be confused with what is actually

random in the process. By viewing the design as estimated, rather than prespecified, our formulation (16) correctly separates the model into a portion driven by population features contained in the k and individual features represented.

In sum, the proto-spline class provides an interpretation for an interesting class of nonstandard mixed effect models.13 This interpretation is a philosophically distinct, hybrid modeling approach, and thus not only generates new knowledge with its use, but also establishes a new way to \allocate" the information provided in longitudinal data. The parameters contained in k partition the variance as follows: they set the overall level of variation, since this is given by the sum of the components' squared values; and they describe the correlation structure because they define a shape which relates observations at different points in time. This formulation thus captures two things simultaneously in a full modeling class| orienting a modeling class to have these philosophical properties is to our knowledge novel in the literature.14 By developing this class using likelihood-based procedures, a complete set of inferential tools is at the analyst's disposal. Scott and Handcock (2000) establish the asymptotic properties of this class and discuss inferential techniques. Being able to disorientate between population and individual effects is crucial to the formation of comparative statements in the policy domain.

5.4 Illustration

For ease of exposition, we illustrate our model class by fitting a single latent quadratic curve proto-spline model to longitudinal wage data from the NLS. For the fixed effects, Xi(t), we use a simple quadratic in age; this yields Model III of Section 3.3. In Figure 4, we display the cross-sectional mean of the process.15 It provides the center from which the curves deviate. In Figure 5, we superimpose 4 simulated realizations from the proto-spline model , with the residual process suppressed. The fitted curve ^_1 used in that simulation is presented in Figure 6. >From this figure, one can see that the growth of wages near the college years of 18 to 22 sets the extent of growth for the later years as well. The shape of the single latent curve describes the long-term trend in variation| strong growth in the 20s, followed by steady but diminished growth in the 30s. Each realization is simply the mean curve plus some random multiple (positive or negative) of the latent curve .

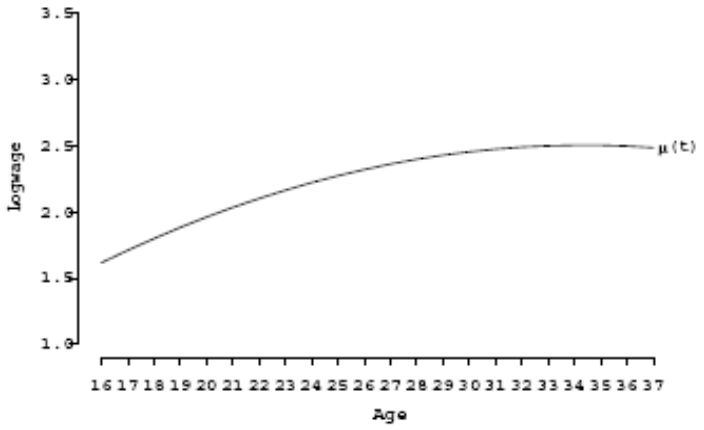


Figure 4: Mean curve for single proto-spline model

One might be concerned that imposing a quadratic latent curve is overly restrictive and essentially forces the decomposition into the shape indicated above. However, within the class of quadratic curves there are pure linear and constant curves, so if there were no change in the growth rate at early and later ages, then we would expect a different fitted latent curve. By forming a single latent curve spanning all ages, we are specifying that we want this curve to represent long-term structure, within the quadratic class. This is in part how we can model the covariance structure for the entire age span even though only segments of the full trajectory are observed for a specific individual.16 More

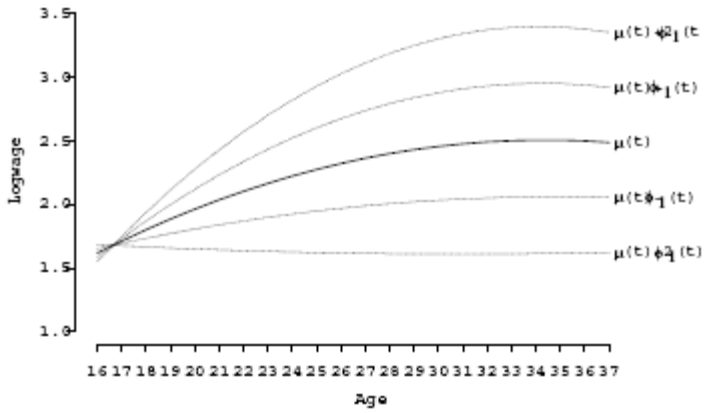


Figure 5: Curves for random coefficients one and two standard deviation from the mean for single proto-spline model.

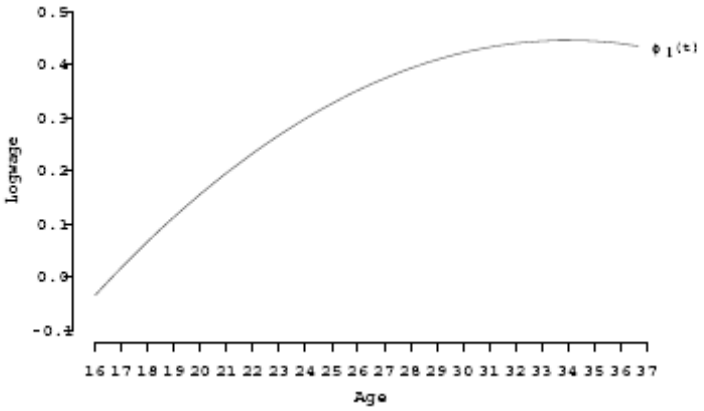


Figure 6: Fitted latent curve for single proto-spline model

complex spaces may reveal more complicated dependencies, should they exist, and they should be considered in the model formulation process.

In Figure 5, we see that the effect of the single latent curve crosses the mean at age 17. The zero value for near this age corresponds to a negligible amount of permanent variance, but even this is not predetermined by our choice of basis. If below-mean wages at those ages were to lead to much larger gains later on, the “crossover" would be at some later age. Random quadratics do limit us to a single change in the direction of growth (positive or negative), while higher order polynomials would not. Finally, note an important difference between this model and Model II. In Model II, each individual has a uniquely shaped quadratic curve, so it may rise quickly and not level o, or it may level o quickly. In our model, which is basically Model III, every individual's variation beyond the mean has the same shape, given by only the magnitude of that variation is allowed to vary.

5.5 Application and comparison of models

To illustrate how the models differ the in practice, we apply several different covariance models to labor market data from NLS. After a preliminary analysis, we found that the mean structure in this data resembles a quadratic curve, so we set the columns of the fixed effects design matrix to correspond to constant, linear and quadratic growth over time.17 More complex mean structures can describe the influence of additional covariates on aggregate wage growth; our goal in this study is to understand the degree of long-term wage stratification, so the overall divergence of these curves over time in comparable samples yields important substantive information. Next, we must select a form for the structured portion of the variation. For wage data, the structured portion consists of long-term, or permanent, differences between wage trajectories.

We will compare three models. The first is a random quadratic mixed effects model similar to Model II and to that used by Bernhardt, et al. (1997) in their analyses. The second is a single latent curve proto-spline model, similar to Model III, and the third is an extension of proto-spline models that includes a second non-orthogonal latent curve. Beyond the structural variation just described, the residual variation is modeled simply as independent with constant variance.

Random Quadratics

The strength of a random quadratic model, such as that given by (14) is the flexibility provided by the three random coefficients. Note that the quadratic basis we use is an orthogonal zed and normalized version of (1; t; t2), which is also the

fixed effects basis. The random coefficients are globally constrained to come from a multivariate Gaussian density. This choice yields a broad range of curves of various shapes and intensities. This distributional form does, however, require that there are no clusters of curves, or other multimodalities.

We assume that are distributed as N(0;G), where G is a completely unspecified 3 *3 covariance matrix defined by six distinct parameters. We the model on data from column is a vector consisting entirely, based on the
the recent NLS cohort18 using maximum likelihood estimation, and that

$$\hat{G} = \begin{pmatrix} +2.019 & +0.899 & -0.265 \\ +0.899 & +1.312 & +0.096 \\ -0.265 & +0.096 & +0.529 \end{pmatrix}$$

and 0:0719.19 Unfortunately, these results are somewhat hard to interpret. The structured portion of the covariance is given by Zi ,T i ; where Zi is the random effects design matrix. Since the rows of Zi correspond to the subject's age, this matrix product describes individual wage differences at each age and how they relate to each other. For example, the diagonal of Zi ,T i represents the structured, or permanent, wage variance at each age. These values are plotted against age in Figure 7 below. The initially larger variance at the earliest ages indicates some initial stratification between individual trajectories that seems to diminish by age 20, only to increase substantially from that point forward, with a dramatic rise after age 32. Had permanent differences in trajectories been limited to an intercept shift, this graph would have consisted of a horizontal line some distance above the axis. The result above indicates that wages fan out quite dramatically as individuals age,

and gives some indication of how the process accelerates. We can infer that the trajectory

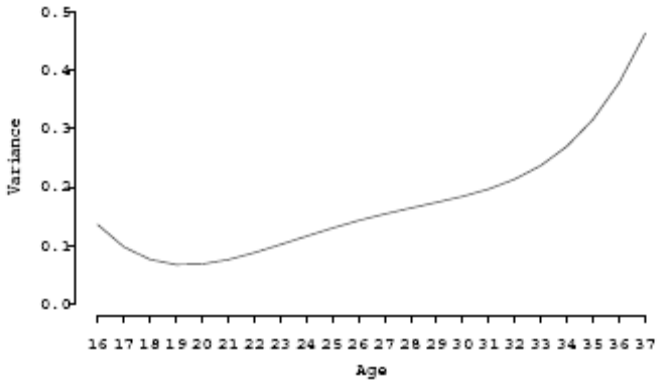


Figure 7: Permanent wage variance for random quadratic model

fans out as a whole because the partition is based on a model employing a continuous curve for the permanent portion of the trajectory.

Single latent curve proto-spline

In this model, we assume that most of the structured variation takes a specific form, but we let the exact shape be determined by the data. The explicit model is

$$Y_i(t) = X_i(t)\beta + \delta_i\phi_1(t) + \epsilon_i(t), \qquad (18)$$

where is the single latent curve, assumed to lie in the space of quadratic curves, and is the random coefficient for the ith individual. This is the same model as the one used for our illustrative example in Section 4. A look at Figure 5 (prior page) reveals the strength of this model. A wide range of outcomes are easily represented

by the mean plus a random multiple of the single latent curve, Figure 6 (prior page) displays this curve for the recent cohort.

In this figure, the interpretability of this class of longitudinal data models becomes apparent. The single latent curve reveals most of what we need to know about structured variation. Contrast this to the covariance matrix , which along with design matrix Zi provides the equivalent information in a less accessible form. The random coefficient on our proto-spline model is standard Gaussian, so we have an immediate sense of the range of impact of the single latent curve.

The restriction to a single latent curve does limit our ability to model more complex structured variation. In Figure 8 below, we see that the permanent variation, the squared version of , describes a very simple growth structure.20 Two features stand out in comparison to random quadratic models: the permanent variation starts out lower at the youngest ages and it does not grow as dramatically as individuals age. We believe that the initial variation is less important from a likelihood perspective, so it is effectively being ignored in the estimation process. Had we used a higher order polynomial, we might have discovered persistent initial wage differences. If this were the case, we would expect to begin higher, possibly decrease somewhat and then increase again, in a shape similar to the permanent variance graph from the random quadratic model.

Double latent curve model

The limitations of a single curve model prompts us to explore a model with two latent curves. Fitting such a model under the pure proto-spline formulation

would require the fitted curves to be orthogonal, and this restricts the function spaces in which each may lie. We propose a new model that effectively “reuses" the basis for each latent curve. The model, abstractly, is given by

$$Y_i(t) = X_i(t)\beta + \delta_{1i}\phi_1(t) + \delta_{2i}\phi_2(t) + \epsilon_i(t), \tag{19}$$

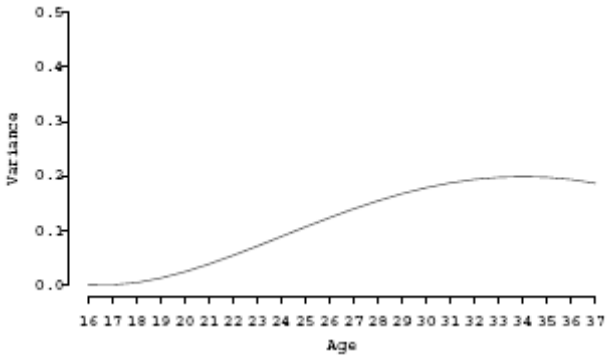


Figure 8: Permanent wage variance for single latent curve model

where latent curves. standard Gaussian random coefficients. We construct each curve from the same basis.

$$\phi_k(t) = \sum_{j=1}^{T} \eta_{kj}\psi_j(t), \tag{20}$$

with k = 1 or 2, effectively doubling the number of parameters used by the single latent curve model. After adding some identifiably constraints to our estimation procedure, we were able to _t this more complex model.

The double latent curve model can best be understood as the combination of a common mean process and two independent “shocks" taking some functional form. We choose to continue to employ the space of quadratic polynomials for ease of exposition. Looking at Figure 9, we find that the fitted curves are quite different from each other. These are the forms for the two shocks, We see that is

quite similar to its counterpart in the single latent curve model, although it starts out further below the origin. The latter feature will induce greater permanent variation at the youngest ages, and then this will subside, as the curve crosses the origin between ages 18 and 19 (contrast this to the crossing at age 17 in the single curve model). The second curve introduces a whole new feature to the co variation. It appears that individuals who start out earning more are penalized as they age. This is indicated by initially positive level of about 0.2 at age 16, which sinks to -0.3 by age 37. Of course, negative random coefficients are just as likely as positive ones, so this curve could also represent later growth for young workers who initially accept lower wages. There is mild evidence that this is capturing an "education effect," in which individuals who defer fully entering the labor market (and possibly pursue education or training) benefit with larger wage growth in the long run.

Note also the similarities and differences of our fitted model to a principal components analysis (PCA). The proto-spline restriction to a smooth function space means that short term variability is definitely removed, and each curve represents a permanent component of variation. With a model-based approach, we can precisely describe how the latent curves are added to the response process. This is less immediate with the components in a PCA, because the PC scores have no predetermined distributional form. Further, the proto-spline process is well-defined under the entire age range of interest without either the use of an ad hoc procedure or requiring a balanced design.

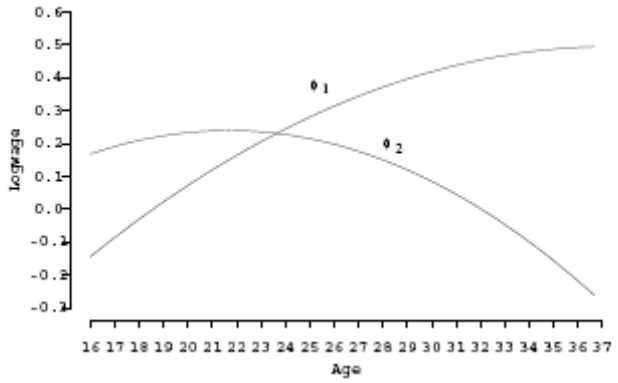


Figure 9: Fitted proto-splines for double latent curve model

The permanent variance partitioning for this model is given in Figure 10, below. By

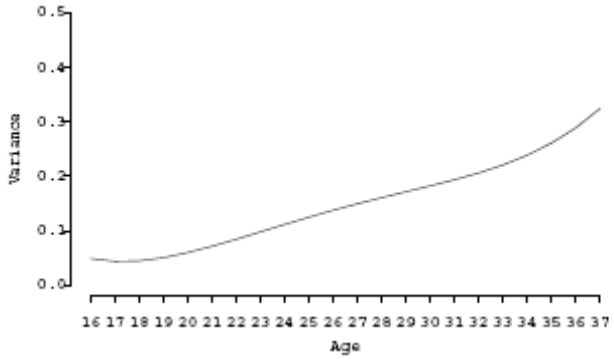


Figure 10: Permanent wage variance for double latent curve model

including two curves additively and independently, this model allows for larger early and later year variation. The effects are permanent, in that they persist over the lifetime of a worker, but their independence points to a subtlety of these variance decompositions. Two curves, along with their coefficients describe the systematic portion of a trajectory, but the independence of the coefficients severs any link between the two. In terms of generating mechanisms, this only makes sense if two different features of the wage growth process are being captured, such

as an overall growth (often attributed to returns to job tenure and experience) and an education effect.

The above comment also points to a limitation of the random quadratic model. Namely, it is hard to describe an underlying process (often thought of as a latent characteristic) that is driving the three coefficients forming the curves. It is hard to imagine that a social or economic generating mechanism involves intercept, slope and acceleration components. The latent curve models provide simpler explanations, which is an advantage in this case.

Comparing variance partitions

While related, these three models provide different variance decompositions. We display the permanent variation plots in Figure 11, below, and include 95% confidence intervals at each age. We discuss the construction of those intervals in Section A.1 of the appendix. Notable differences exist for the youngest and oldest ages, with strong agreement in the middle range. The single latent curve model does not pick up much structured wage variation at the youngest ages. If initial differences in wages persist during the youngest ages, but then diminish, then this model will have to choose between the initial and later year effects, and since the latter are larger, they tend to dominate. The double proto-spline model picks up this extra variation in , and this is reflected in larger permanent variance for the younger ages. The random quadratic model picks up more variation in both younger and older ages and labels it permanent. We contend that the additional flexibility of the random quadratic model allows it to follow the raw data more

closely, capturing less rigid forms of variation. This is indirectly confirmed by examining the residual variation, which is 0.072 for random quadratics, and 0.078 and 0.098 for double and single latent curve models, respectively.

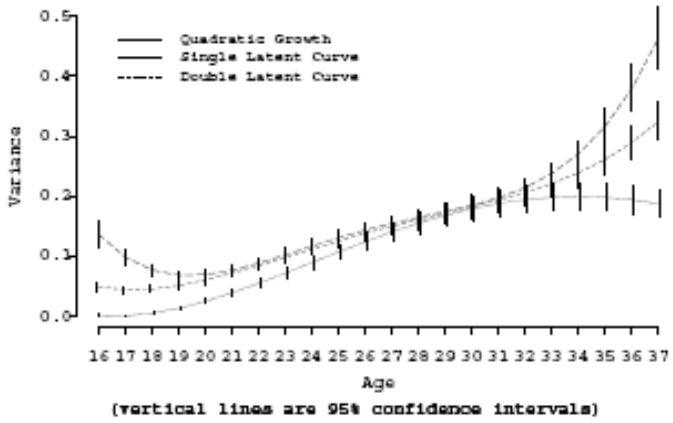


Figure 11: Permanent wage variance for all three models

Below we present the variance decomposition for each cohort to address an important question. While each model partitions the variance differently, do these differences have substantive impact? That is, how sensitive are the answers to the substantive questions to the choice of model? In our application, the question of interest is whether or not the permanent wage variance between the cohorts differs, and if so, by how much. Any model we use will only be an approximation, but if the answer to our question is consistent across models, we can have more confidence in any conclusions we draw.

In Figures 12 through 14 below, we make a cross-cohort comparison and display the model-based permanent variance for each model along with 95% confidence intervals at each age. All of the models indicate a significantly larger permanent variance in the recent cohort, starting sometime in the mid-twenties.

The difference is most dramatic in the random quadratic model and least so in the single latent curve model. There is some between-model discrepancy in what portion of the variance is permanent at the youngest ages, and in how the cohorts differ. Both latent curve models contain a crossover, in which the original cohort starts out more stratified until the early twenties, at which point the opposite it true. In contrast, the random quadratic model posits that both cohorts are more permanently stratified initially and to a comparable extent. If we are interested in the absolute magnitude of permanent wage stratification, we must look more closely at all of these models and determine which is more justified on substantive grounds. If we were concerned about wage stratification at the younger ages, a deeper understanding of each model's characteristics is warranted, since these models tell three different stories

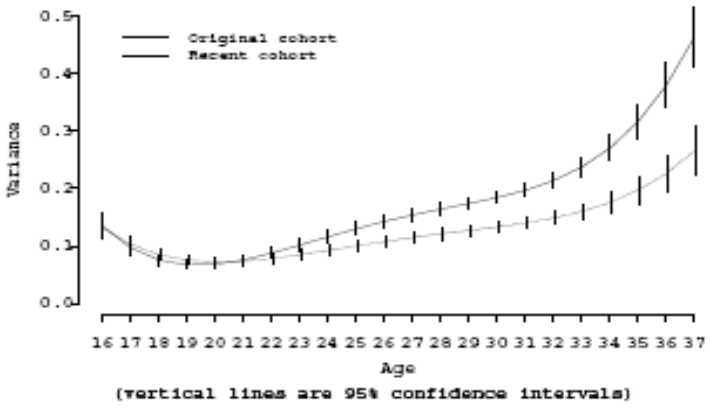


Figure 12: Permanent wage variance for random quadratic model

Discussion of findings

All three of these models indicate a significant increase in permanent wage variation in the recent cohort for the older ages. But the magnitude of these differences varies greatly between models, and strong differences in the partitions exist at the younger ages.

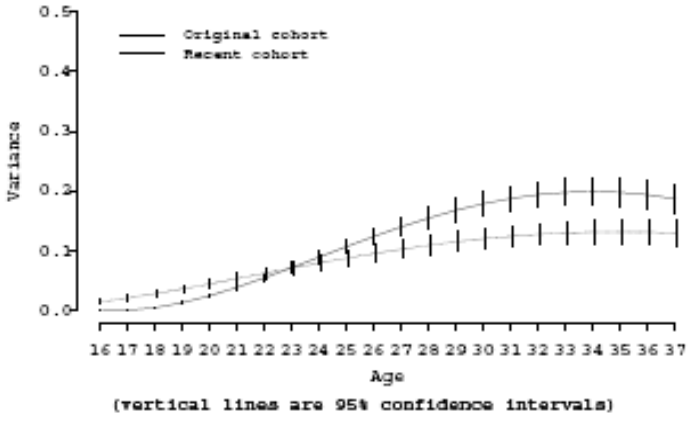


Figure 13: Permanent wage variance for single latent curve model

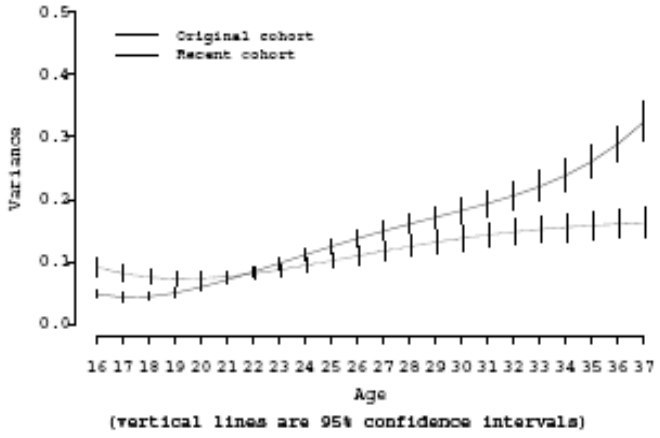


Figure 14: Permanent wage variance for double latent curve model

Since the random quadratic model labeled more variation as permanent, it may be over fitting that feature, in some sense. The flexibility of random quadratics admits even a U-shaped curve, but is it desirable to use such a shape to describe permanent wage gains? U-shaped curves, in which initial and final wages are nearly identical, involve a shift in the direction of wage change, from loss to gain, so in what sense is this indicative of a permanent, or lasting, trend? We must understand how the choice of model is reflected in the variance partition if we intend to make an informed assessment of social phenomena.

Latent curve models stand out as a philosophically mixed approach to creating variance partitions. They are highly interpretable, with independent components acting as shocks to the process. The shocks may be readily interpretable in the context of the generating mechanisms for the social processes under study. They offer a handy form of rigidity compared to random quadratic models, yet are inherently adaptive to overall patterns of structured variation. The hybrid nature of this model class provides a new type of analysis to which results from other classes can be compared. Thus, these models can be viewed as excellent foils to the classical random quadratic model.

All three of the models describe the structured portion of variance in such a way that ”permanent" is a reasonable label to apply. That is, the model describes smooth versions of the curves in space that are reasonable attempts to separate the analyst-defined signal from noise; and the signal is non-stochastic, conditional on the parameters that describe it. The differences in these definitions allow different aspects of variation to be identified.

5.6 Conclusion

We embarked on this analysis to determine how different models for covariance affect variance component partitioning. Along the way, we introduced a new, hybrid class of latent curve models, proto-splines, that offer an interpretable paradigm for describing co variation, which is well-suited to formulating substantive questions directly. These models locate population level persistent covariance structure and reflect it in the shape and size of latent curves. We view proto-splines as covariance function smoothers; they are non-parametric in the sense that the estimated curve lies in a function space, yet the model formulation provides a straightforward interpretation of the curves that is often missing in other non-parametric techniques. In the model formulation, the researcher imposes a class of functions to capture substantively meaningful structure. The restriction to a particular class of functions forces proto-spline models to be conservative in the way they t the data-they are less susceptible to outliers, which in other models may influence both prediction and t. This makes them invaluable in comparisons with more traditional models; the ways in which they differ point out characteristics of each, with the clearly defined behavior of our models acting as a foil for the others.

In future work, we will consider relaxing the independence assumption for the protospline model class. For example, our double latent curve model could include a term for the correlation between curves. This extension would open up the possibility of very different latent curves, since the independence constraint ultimately lowers the likelihood of certain shapes for the fitted curves. Including

more complex residual structures, such as age-specific variances, as a check on the homogeneous variance assumption could prove useful.

In many socio-economic processes, there are jump points that are not smooth, but have important substantive meaning. Adapting the proto-spline class to allows for uncertainty in the timing of the change-point could prove useful. Raftery (1994) explores this issue; integrating his approaches with ours is a research direction of interest.

Relaxing the Gaussianality assumption is worth investigating, but we would limit this to forms that remain interpretable, such as parametric forms. One approach that has been suggested by several researchers is a latent class, or mixture formulation (see Clogg 1995, Baneld and Raftery 1993, Muthen and Shedden 1999, Roeder 1999, Verbeke and Lesafire 1997, Xu, etal. 1996). Under this paradigm described earlier, individuals belong to one latent class, and then conditional on class membership they follow a certain structure. An important point is that the remaining structure could be flexibly captured in the proto-spline models just introduced; most models currently in use do not offer such directly interpretable covariance formulations.

In work in progress, we are examining diagnostics for these models in greater detail. Model selection criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978) can be applied here. These are discussed in Vonesh and Chinchilli (1998) and Pinheiro et al. (1994). Recent extensions to the AIC discussed in Simono_ and Tsai (1999) appear to be especially promising in in the context of these variance component models. An alternative to model selection is the use of

Bayesian model averaging (Hoeting, etal. 1998). A developed set of diagnostic techniques will add to our understanding of how each model captures and partitions variation.

Chapter – 6
Alternatives to Traditional Model Comparison Strategies for Covariance Structure Models

Chapter – 6
Alternatives to Traditional Model Comparison Strategies for Covariance Structure Models

6.1 Introduction

In this chapter I discuss two related issues relevant to traditional methods of comparing alternative covariance structure models (CSM) in the context of ecological research. Use of the traditional test of parametrically nested models in applications of CSM (the $\chi 2$ difference or likelihood ratio [LR] test) suffers from several limitations, as discussed by numerous methodologists (MacCallum, Browne, & Cai, 2005). Our primary objection is that the traditional approach to comparing models is predicated on the assumption that it is possible for two models to have identical fit in the population. We argue instead that any method of model comparison which assumes that a point hypothesis of equal fit can hold exactly in the population (e.g., the LR test) is fundamentally flawed. We discuss two alternative approaches to the LR test which avoid the necessity of hypothesizing that two models share identical fit in the population. One approach concerns framing the hypothesis of interest differently, which naturally leads to questions of how to assess statistical power and appropriate sample size. The other approach concerns a radical realignment of how researchers approach model evaluation, avoiding traditional null hypothesis testing altogether in favor of identifying the model that maximizes generalizability.

Power presents a recurrent problem to those familiar with null hypothesis significance testing (NHST). How large should a sample be in order to have

adequate probability of rejecting a false null hypothesis? What is the probability of rejecting a false null if our sample is of size *N*? These questions present special challenges in the context of CSM because the relative status of null and alternative hypotheses are interchanged from their familiar positions — the null hypothesis in CSM represents the theory under scrutiny, and power is framed in terms of the sample size necessary to reject a false model. Traditional goodness of- fit tests deal with the null hypothesis under which the model fits exactly in the population (exact fit test). Point hypotheses tested by the exact fit test are likely never true in practice, so how should power be conceptualized? We present an alternative strategy extending earlier work on power for tests of close fit (rather than exact fit) of single models to tests of *small difference* (rather than no difference) in comparisons of nested models. The null hypothesis in a test of small difference states that the model fits nearly as well, but not the same, as a less constrained model.

Another alternative to traditional methods of model assessment is to avoid the hypothesis-testing framework altogether, instead adopting a model selection approach that uses comparative reliability as the criterion for selecting a model as superior to its rivals (Weakliem, 2004). Specifically, we argue that the evaluation of models against arbitrary benchmarks of fit gets the researcher nowhere — only in the context of model comparison can science advance meaningfully (Burnham & Anderson, 2004). Maximizing generalizability involves ranking competing models against one another in terms of their ability to fit present and future data. Adopting this model selection strategy, however, necessitates proper quantification

of *model complexity* — the average ability of a model to fit any given data. Most model fit indices include an adjustment for complexity that is a simple function of the number of free model parameters. We argue that this adjustment is insufficient; the average ability of a model to fit data is not completely governed by the number of parameters. Consequently, we present and illustrate the use of a new information-theoretic selection criterion that quantifies complexity in a more appropriate manner. This, in turn, permits the adoption of an appropriate model selection strategy that avoids pitfalls associated with LR tests.

I begin by providing a review of the traditional representation of the covariance structure model (with mean structure), with an emphasis on its application to multiple groups. We then describe advantages granted by adopting a model comparison perspective in CSM. One way around the problems with traditional approaches is to change the hypothesis under scrutiny to a more realistic one. In describing this alternative approach, we describe an approach to power analysis in CSM involving an extension of recently introduced methods to nested model scenarios. Following our discussion of power, we further explore the potential value of adopting a model selection approach that avoids hypothesis testing — and thus most problems associated with LR tests—altogether. In the process, we introduce the topic of model complexity, suggesting and illustrating the use of a new selection criterion that permits appropriate model comparison even for no nested models.

6.2 COVARIANCE STRUCTURE MODELING

Covariance structure modeling (CSM) is an application of the general linear model combining aspects of factor analysis and path analysis. In CSM, the model expresses a pattern of relationships among a collection of observed (manifest) and unobserved (latent) variables. These relationships are expressed as free parameters representing path coefficients, variances, and covariance's , as well as other parameters constrained to specific, theory-implied values or to functions of other parameters. For simplicity, we restrict attention to the *ally* model (LISREL Submodel 3B; Joreskog & Sorbom, 1996), which involves only four parameter matrices, although the points we discuss later apply more broadly.

6.3 The Importance of CSM to Ecological Research

There are several advantages associated with CSM that make it especially appropriate for addressing hypotheses in the context of ecological models. First, CSM permits the specification and testing of complex causal and co relational hypotheses. Sets of hypotheses can be tested simultaneously by constraining model parameters to particular values, or equal to one another within or across multiple groups or occasions of measurement, in ways consistent with theoretical predictions. Second, by permitting several measured variables to serve as indicators of unobserved latent variables, CSM separates meaningful variance from variance specific to items, allowing researchers to test structural hypotheses relating constructs that are not directly observed. Third, CSM is appropriate for testing co relational or causal hypotheses using either (or both) experimental or

observational data. One of the central ideas behind ecological modeling is that there is much knowledge to be gained by collecting data observed in context that would be difficult or impossible to learn under artificial conditions. Finally, CSM is a flexible modeling approach that can easily accommodate many novel modeling problems.

6.4 The Importance of Adopting a Model Comparison Perspective

In practice, CSMs are typically evaluated against benchmark criteria of good fit. Based on how well a model fits data relative to these criteria, the model is 0usually said to fit well or poorly in an absolute sense. The reasoning underlying this strategy of gauging a model's potential usefulness is predicated on an approach to science termed falsifications, which holds that evidence accumulates for theories when their predictions are subjected to, and pass, realistic "risky" tests. If a model passes such a test under conditions where it would be expected to fail if false (i.e., if it shows good fit), evidence accumulates in favor of the theory whose predictions the model represents. If it fails, the model is either rejected or modified, with implications for the revision or abandonment of the theory. Ideally, a model is subjected to repeated risky tests to give a better idea of its long-term performance, but replication is unfortunately rare in the social sciences.

An alternative philosophical perspective maintains that the evaluation of models in isolation tells us very little, and that the fit of a model to a particular data set is nearly uninformative. Rather, science progresses more rapidly if

competing theories are compared to one another in terms of their abilities to fit existing data and, as we will discuss, their abilities to fit *future* data arising from the same latent process (Lakatos, 1970; MacCallum, 2003). This approach is sometimes termed *strong inference* (Platt, 1964), and involves model comparison as a signature feature. We know from the outset that no model can be literally true in all of its particulars, unless one is extraordinarily lucky or possesses divinely inspired theory-designing skills. But it stands to reason that, given a set of alternative models, one of those models probably represents the objectively true data-generating process better than other models do. It is the researcher's task to identify this model and use it as the best working hypothesis until an even more appropriate model is identified (which, by design, inevitably happens). Every time a model is selected as the optimal one from a pool of rivals, evidence accumulates in its favor. This process of rejecting alternative explanations and modifying and re-testing models against new data continues *ad infinitum*, permitting scientists to constantly update their best working hypotheses about the unobserved processes underlying human behavior.

Because no model is literally true, there is an obvious logical problem in testing the null hypothesis that a model fits data perfectly in the population. Yet, this is precisely the hypothesis tested by the popular LR test of model fit. Moreover, most fit indices require the researcher to choose arbitrary values to represent benchmarks of good fit. A model comparison approach goes far in avoiding these problems, although it cannot avoid them altogether. Most damning, it is possible to assert apriori that the hypothesis tested with the $\chi 2$ statistic — that

a model fits exactly in the population or that two models share exactly the same fit — is false in virtually every setting (Bentler & Bonett, 1980; Tucker & Lewis, 1973). A model selection approach avoids the pitfalls inherent in hypothesis testing by avoiding such tests altogether.

In addition to adhering more closely to scientific ideals and circumventing logical problems inherent in testing isolated models, the practice of model comparison avoids some problems associated with confirmation bias. Confirmation bias reflects the tendency for scientists unconsciously to increase the odds of supporting a preferred hypothesis (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Regardless of why or how much the deck is stacked in favor of the researcher's preferred model in terms of absolute fit, one model is virtually guaranteed to outperform its rivals. Model comparison does not entirely eliminate confirmation bias, but it certainly has the potential to improve the researcher's objectivity.

In the foregoing we have explained that the popular LR test is fundamentally flawed in that the hypothesis it tests is rarely or never true in practice; thus, persistent and frequent use of the LR test is of questionable utility. We have also explained that adopting a model selection approach, in which at least two theory-inspired models are compared, has potentially greater scientific potential. In the following two broad sections, we outline some practical solutions to logical problems imposed by use of the traditional LR tests of model fit in ecological research. The first suggested approach emphasizes the utility of avoiding the hypothesis that two models have identical fit in favor of a hypothesis

that the difference is within tolerable limits. This approach recognizes that no model can realistically fit perfectly in the population, and points out that shifting the focus to a less stringent hypothesis is more logical, yet has consequences for statistical power and identifying the necessary sample size. We describe and discuss methods that can be used to address these problems. The second section focuses more closely on the model selection perspective just outlined, emphasizing that model fit is overrated as a criterion for the success or usefulness of a theory. Rather, more attention should be paid to a model's ability to cross-validate, or generalize, relative to competing models. Special attention is devoted to a new model selection criterion that considers aspects of model complexity beyond simply the number of free parameters.

6.5 Concluding Remarks

There are two broad issues that we wish to emphasize to close this section on power analysis and specification of the null hypothesis when performing comparisons of nested models. The first issue is the choice of pairs of RMSEA values. Essentially the results of any application of any of the methods we described are contingent on the particular RMSEA values that the user selects. Here we can offer only some general principles. For a more thorough discussion of this issue we refer the reader to MacCallum et al. (2006). For specifying RMSEA values for testing a null hypothesis of a small difference in fit, the user should regard the Good-Enough Principle (Serlin & Lapsley, 1985) as the objective, and pick RMSEA values for Models A and B that represent a difference so small that

the user is willing to ignore it. In the context of power analysis, the relevant general principle would be to choose values that represent a difference that the investigator would wish to have a high probability of detecting. In practice, users will need to rely on guidelines for the use of RMSEA as mentioned earlier (Browne & Cudeck, 1993; Steiger, 1994), as well as the characteristics of the models under comparison.

The second issue has to do with the assumptions involved in our developments. All of the methodological developments presented thus far rely on well known distribution theory and its assumptions. Specifically, we make extensive use of the assumptions that ensure the chi-squared ness of the LR test statistic T , for both the central and non central cases. These include multivariate normality, the standard set of regularity conditions on the likelihood to carryout asymptotic expansions, and the population drift assumption (Steiger et al., 1985). As always, however, such assumptions never hold exactly in the real world, so the user should always be cautious in the application of these methods in data analysis and should watch for potential pitfalls due to assumption violations. MacCallum et al. (2006) discuss the consequences of such violations.

6.6 MODEL SELECTION AND MODEL COMPLEXITY

Model Selection and Generalizability

In the preceding section we provide and illustrate methods for comparing rival models in terms of a noncentrality-based fit index, RMSEA. We suggest that this strategy is appropriate for statistically comparing the fit of rival,

parametrically nested models, but the procedure depends in part on the researcher's judgment of appropriate choices for $\varepsilon * A$ and $\varepsilon * B$, or what, in the researcher's judgment, constitutes the smallest difference in fit that it would be interesting to detect. In practice, a model can demonstrate good fit for any number of reasons, including a theory's proximity to the objective truth (or *verisimilitude*; Meehl, 1990), random chance, simply having many free parameters, or by possessing a structure allowing parameters to assume values which lead togood model fit for many different data patterns—even those generated by other processes not considered by the researcher. In other words, models can demonstrate close fit to data for reasons other than being "correct," even if one grants that true models are possible to specify (we do not), so good fit should represent only one criterion by which we judge a model's usefulness or quality.

Another criterion of model success that has found much support in mathematical psychology and the cognitive modeling literature is generalizability (or reliability). The idea here is that it is not sufficient for a model to show good fit to the data in hand. If a model is to be useful, it should *predict* other data generated by the same latent process, or capture the regularities underlying data consisting of signal and noise. If a model is highly complex, refitting the model to new data from scratch will not advance our knowledge by much; if a model's structure is complex enough to show good fit to one data set, it may be complex enough to show good fit to many other data sets simply by adjusting its parameters. In other words, pure goodness of fit represents fit to signal *plus*

fit to noise. However, if model parameters are fixed to values estimated in one setting, and the model still demonstrates good fit in a second sample (i.e., if the model *cross-validates* well), the model has gained considerable support. A model's potential to cross-validate well is its generalizability, and it is possible to quantify generalizability based only on knowledge of the model's formant of its fit to a given data set. By quantifying a model's potential to cross validate, generalizability avoids problems associated with good fit arising from fitting error or from a model's flexibility. It also does not rely on unsupportable assumptions regarding a model's absolute truth or falsity. Therefore, generalizability is arguably a better criterion for model retention than is goodness of fit per se (Pitt & Myung, 2002).

Earlier we stated that adopting a model selection perspective requires a fundamental shift in how researchers approach model evaluation. Traditional hypothesis testing based on LR tests results in a dichotomous accept–reject decision without quantifying how much confidence one should place in a model, or how much relative confidence one should place in each member of a set of rival models. In model comparison, on the other hand, no null hypothesis is tested(Burnham & Anderson, 2004). The appropriate sample size is not selected based on power to reject hypotheses of exact or close fit (obviously, since no such hypotheses are tested), but rather to attain acceptable levels of precision of parameter estimates. Rather than retaining or discarding models on a strict accept–reject basis, models are ranked in terms of their generalizability, a notion that

combines fit with parsimony, both of which are hallmark characteristics of a good model.

The model selection approach does not require that any of the rival models be correct, or even (counter intuitively) that any of the models fit well in an absolute sense. The process is designed in such a way that researchers will gravitate toward successively better models after repeated model comparisons. The more such comparisons a particular model survives, the better its track record becomes, and the more support it accrues. Therefore, it is incumbent upon scientists to devise models that are not only superior to competing models, but also perform well in an absolute sense. Such models will, in the long run, possess higher probabilities of surviving risky tests, facilitate substantive explanation, predict future data, and lead to the formulation of novel hypotheses. But, again, the model selection strategy we advocate does not require that any of the competing models be correct or even close to correct in the absolute sense.

Information-Theoretic Criteria

In contrast to model selection methods rooted in Bayesian or frequents traditions, much research points to information theory as a likely source for the optimal model selection criterion. Selection criteria based on information theory seek to locate the one model, out of a pool of rival models, which shows the optimal fidelity, or signal-to-noise ratio; this is the model that demonstrates the best balance between fit and parsimony. This balance was termed *generalizability* earlier. Several popular model selection criteria were either derived from, or are

closely related to, information theory. The most popular such criteria are the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978). Excellent treatments of AIC and BIC can be found elsewhere (e.g., Burnham & Anderson, 2002, 2004; Kuha, 2004).

Many information-based criteria may be construed as attempts to estimate the Kullback–Leibler (K–L) distance. The K-L distance is the (unknown) information lost by representing the true latent process with an approximating model (Burnham & Anderson, 2004). Even though we cannot compute the K–L distance directly because there is one term in the K–L distance definition that is not possible to estimate, we can approximate *relative* K–L distance in various ways by combining knowledge of the data with knowledge of the models under scrutiny. Of great importance for model comparison, the ability to approximate relative K–L distance permits the ranking of models in terms of their estimated verisimilitude, tempered by our uncertainty about the degree of approximation. In other words, using information-based criteria, models can be ranked in terms of estimated generalizability.

Minimum Description Length and the Normalized Maximum Likelihood

Information-based criteria such as AIC and BIC are used with great frequency in model comparisons and with increasing frequency in applications of CSM. However, they suffer from at least two major drawbacks. First, they employ complexity adjustments that are functions only of the number of free model parameters. Second, they implicitly require the strong assumption that a correct

model exists. We focus instead on a newer criterion that remains relatively unknown in the social sciences, yet we feel has great promise for application in model selection. This is the principle of *minimum description length* (MDL: Grunwald, 2000; Myung, Navarro, & Pitt, 2005; Rissanen, 1996, 2001; Stine, 2004). The MDL principle involves construing data as compressible strings, and conceiving of models as compression codes. If models are viewed as data compression codes, the optimal code would be one that compresses (or simply represents) the data with the greatest fidelity. With relevance to the limitations of criteria such as AIC and BIC, the MDL principle involves no assumption that a true model exists. If one accepts that a model's proximity to the truth is either undefined (i.e., that the notion of a true model is merely a convenience and bears no direct relation to reality) or is at any rate impossible to determine, then the MDL principle offers a viable alternative to traditional methods of model selection. Excellent discussions of the MDL principle can be found in Grunwald (2000), Grunwald, Myung, and Pitt (2005), Hansen and Yu (2001), and Markon and Krueger (2004). Three quantifications of the MDL principle are *normalized maximum likelihood* (NML), *Fisher information approximation* (FIA), and *stochastic information complexity* (SIC). NML is quantified as:

$$\mathrm{NML} = \frac{L(y|\hat{\theta})}{\int_S L(z|\hat{\theta}(z))dz}, \qquad (14)$$

or the likelihood of the data given the model divided by the sum of all such likelihoods. FIA is quantified as

$$\mathrm{FIA} = -\ln L\left(y|\hat{\theta}\right) + \frac{q}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int_{\Theta}\sqrt{|I\,(\theta)|}d\theta, \qquad (15)$$

an approximation to the negative logarithm of NML that makes use of the number of free parameters (q) and the determinant of the Fisher information matrix, $I(\theta)$. SIC, an approximation to FIA that is typically more tractable in practice, is quantified as:

$$\mathrm{SIC} = -\ln L\left(y|\hat{\theta}\right) + \frac{1}{2}\ln\left|nI\,(\theta)\right|. \qquad (16)$$

The Appendix (see Quant.KU.edu) contains more detailed discussion of these criteria. NML, FIA, and SIC all represent model fit penalized by the model's average ability to fit any given data.

NML is similar in spirit to selection criteria such as AIC and BIC in several respects, save that preferable models are associated with higher values of NML but with lower values of AIC or BIC.1 All of these criteria can be framed as functions of the likelihood value adjusted for model complexity, although the complexity correction assumes different forms for different criteria. NML differs from criteria like AIC and BIC mainly in that not every parameter is penalized to the same extent. NML imposes an adjustment commensurate with the degree to which each free parameter increases complexity, as reflected in the model's general data-fitting capacity. Consequently, NML does not assume (as do AIC and BIC) that each parameter contributes equally to goodness of fit. Therefore, both parametric and structural components of complexity are considered. A major additional advantage of NML (which it shares with AIC and BIC) is that it does not require rival models to be nested. Thus, if two competing theories posit different patterns

of constraints, such models can be directly compared using criteria derived from information theory.

6.7 Applying MDL in Practice

To illustrate how the MDL principle may be employed in practice, we present two brief examples from the applied literature. In both examples we compute NML; in the second, we supplement NML with computation of SIC because original data were available with which to compute the $|nI(\theta)|$ term. Neither the denominator term in NML (see Equation [A1]) nor the structural complexity term in FIA (see Equation [A2]) can be computed directly in the context of CSM. Numerical integration techniques are typically applied instead. To facilitate computation of NML, we simulated the data space by generating large numbers of random uniform correlation matrices (R) using Markov chain Monte Carlo (MCMC) methods.2 These matrices were uniform in the sense that all possible R matrices had equal *apriori* probabilities of being generated. All models were fit to all simulated matrices, and the likelihoods were averaged to form the denominator of the NML formula.3 The numerators were supplied by simply noting the likelihood value associated with the converged solution for each model applied to real data.

Example 1. Our second example draws on three covariance structure models compared by Larose, Guay, and Boivin (2002). The authors were primarily interested in comparing the Cognitive Bias Model and Social Network Model, two models proposed to explain variability in a Loneliness latent variable using

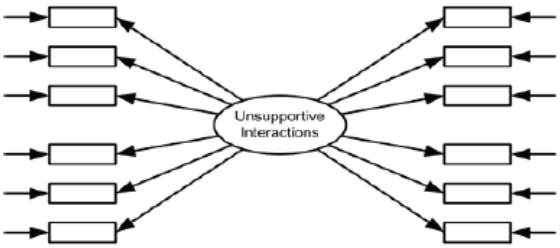
Attachment Security, Emotional Support, and Social Support. These two models (which we denote *L*1 and *L*2) are presented in the first two panels of Figure 3.3. Based on results indicating that both models fit the data well and were thus viable explanations for the observed pattern of effects, the authors devised a third model combining features of the first two, dubbed the Cognitive-Network Model (*L*3 in Figure 3.3).

All three models were found to fit the data well using self-report measures (N = 125), and to fit even better using friend-report measures. In both cases, the Cognitive-Network Model was found to fit the data significantly better than either the Cognitive Bias Model or the Social Network Model. Following procedures already described, we reevaluated Larose et al.'s models (fit to self report data) using NML. Results are reported in Table 3.2. Because raw data were available in their article, we are also able to provide estimates of SIC.

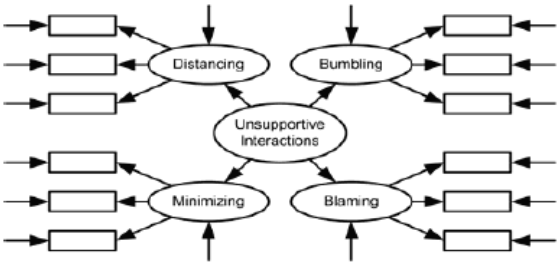
Contrary to the authors' findings, both NML and SIC indicate that the Cognitive Bias Model performs better than either the Social Networks Model or the proposed Cognitive-Network Model in terms of generalizability. Combining features of two already well-fitting models does not necessarily grant a scientific advantage when the resulting model is more complex than either of its

FIGURE 3.2
Rival models investigated by Ingram et al. (2001).

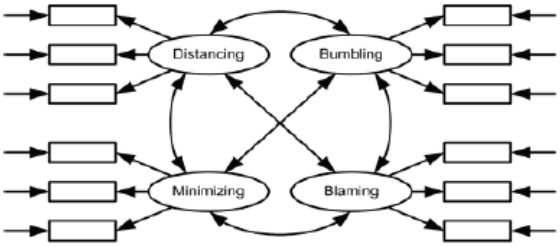
Model I1



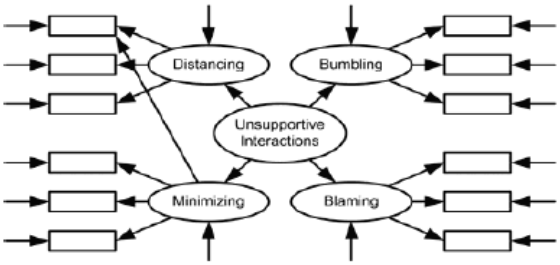
Model I2



Model I3



Model I4



Model I5

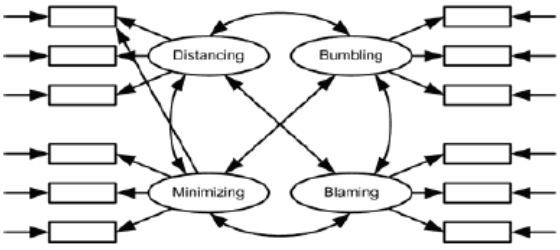
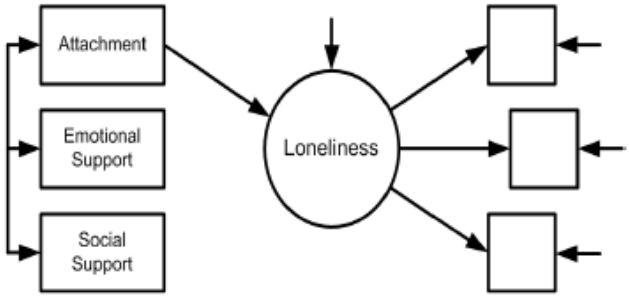
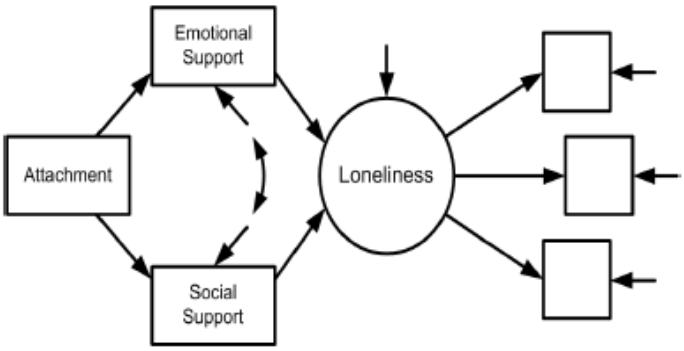


FIGURE 3.3
Rival models investigated by Larose et al. (2002).

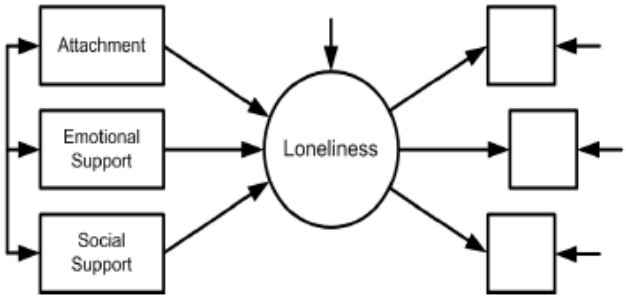
Model L1



Model L2



Model L3



competitors. In this instance, as in the previous example, the chosen model was selected primarily because it showed better absolute fit; this better fit was due in part to the fact that the Cognitive-Network Model was more complex than its competitors. An implication of this finding is that, whereas the Cognitive-Network Model may fit the given data set better than the Cognitive Bias Model and the Social Networks Model in absolute terms, it has a lower likelihood of generalizing well to future data.

6.8 Summary

Like other information-theoretic selection criteria, MDL does not require rival models to be parametrically nested. Nor does its use require the assumption that a true model exists. Furthermore, MDL considers more sources of complexity than simply a model's number of parameters. In sum, we feel that the MDL principle has great potential for use in model comparisons in CSM.

6.9 Limitations

Of course, NML is not a panacea. Three limitations of NML are that it is difficult to compute, it relies on the assumptions of maximum likelihood, and it involves often arbitrary bounds on the data space. The first limitation will be overcome as processor speeds increase and as NML becomes included in standard model estimation packages. In the meantime, the more tractable MDL approximation, SIC (Rissanen, 1989), can be used if the numerical integration

necessary for NML proves too time-intensive. As for the second limitation, it is unknown how robust MDL methods are to violations of ML assumptions. This would be a fruitful avenue for future research.

The third limitation is more challenging because it requires the researcher to make a subjective decision regarding boundaries on the data space. We restricted attention to correlation matrices for simplicity. We recognize that many modeling applications require covariance matrices rather than correlation matrices (and sometimes also mean vectors). For example, virtually any application in which models are fit to multiple groups simultaneously, such as in factorial invariance studies, requires the use of covariance matrices. Growth curve modeling requires covariance matrices and mean vectors. Lower and upper boundaries must be imposed on generated means and variances if such data are required, and these choices constitute even more subjective input. It is generally agreed that data generated for the purpose of quantifying model complexity should be uniformly representative of the data space (Dunn, 2000), yet choices regarding the range of data generation may exert great influence on the ranking of competing models. It is thus important that reasonable bounds be investigated to ensure reasonable and stable model rankings. A discussion of the implications for arbitrary integration ranges can be found in Lanterman (2005).

6.10 DISCUSSION

We have proposed two alternatives to traditional methods of comparing covariance structure models. Both alternatives were suggested in response to

limitations of the popular LR test; the most severe limitation is that the hypothesis tested by the LR test (that two models have identical fit) is never true in practice, so investigating its truth or falsity would seem to be a questionable undertaking (MacCallum et al., 2006). The first alternative procedure posits a modified null hypothesis such that the difference in fit between two nested models is within tolerable limits. The second alternative we discuss is to compare rival (not necessarily nested) models in terms of relative generalizability using selection indices based on the MDL principle. Both methods encourage a model comparison approach to science that is likely to move the field in the direction of successively better models.

There are interesting parallels between the strategies proposed here and a framework for model assessment proposed by Linhart and Zucchini (1986) and elaborated upon by Cudeck and Henly (1991) in the context of CSM. Because it relies on RMSEA to specify null and alternative hypotheses, the first approach (using RMSEA to specify hypotheses of close fit) can be seen as way to compare nested models in terms of their *approximation discrepancy*, or lack of fit in the population. In other words, this method is a way to gauge models' relative nearness to the objectively true data-generating process, or their relative verisimilitudes. The second method of model comparison makes use of the MDL principle to facilitate comparison of models in terms of their relative generalizabilities, or abilities to predict future data arising from the same generating process. This strategy can be seen as a way to compare models (nested or non-nested) in terms of their *overall discrepancy*, tempering information about

lack of fit with lack of confidence due to sampling error. When *N* is large, enough information is available to support highly complex models if such models are appropriate. When *N* is small, uncertainty obliges us to conservatively select less complex models until more information becomes available (Cudeck & Henly, 1991). Thus, NML and similar criteria are direct applications of the parsimony principle, or Occam's razor.

The parallels between the measures of verisimilitude and generalizability on one hand, and the Linhart–Zucchini and Cudeck–Henly frameworks on the other, perhaps deserve more attention in future research. High verisimilitude and high generalizability are both desirable characteristics for models to possess, but selecting the most generalizable model does not necessarily imply that the selected model is also closest to the objective truth. Therefore we do not advocate choosing one approach or the other, or even limiting attention to these two strategies. Rather, we suggest combining these strategies with existing model evaluation and selection techniques so that judgments may be based on as much information as possible. Regardless of what strategy the researcher chooses, the strongest recommendation we can make is that researchers should, whenever circumstances permit it, adopt a model selection strategy rather than to evaluate single models in isolation. The methods illustrated here are viable alternatives to the standard approach, and can be applied easily in many modeling settings involving longitudinal and/or ecological data.

REFERANCE

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

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

Asparouhov, T. (2005). Sampling weights in latent variable modeling. Structural Equation Modeling, 12, 411–434.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London, Series A, 160, 268–282.

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Psychology (Statistical Section), 3, 77–85.

Ogasawara, H. (2001). Approximations to the distributions of fit indexes for misspecified structural equation models. Structural Equation Modeling, 8, 556–574.

Ramaswami, S. N., & Singh, J. (2003). Antecedents and consequences of merit pay fairness for industrial salespeople. Journal of Marketing, 67, 46–66.

Saris, W. E., Den Ronden, J., & Satorra, A. (1987). Testing structural equation models. In P. Cuttance & J. Ecob (Eds.), Structural modeling by example: Applications in educational, sociological, and behavioral research (pp. 202–220). Cambridge, UK: Cambridge University Press.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for statistics in covariance structure analysis (UCLA Statistics Series, No. 2). Los Angeles: University of California, Department of Psychology.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variable analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. Psychometrika, 66, 507–514.

Satterthwaite, F. E. (1941). Synthesis of variance. Psychometrika, 6, 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 110–114.

Scargle, J. D. (2000). Publication bias: The "file drawer" problem in scientific inference. Journal of Scientific Exploration, 14, 91–106.

Sörbom, D. (1989). Model modification. Psychometrika, 54, 371–384.

Swain, A. J. (1975). Analysis of parametric structures for variance matrices. Unpublished doctoral dissertation, University of Adelaide, Australia.

Venables, W. N., & Smith, D. M. (2005). An introduction to R (Version 2.1.1). Retrieved August 15, 2005 from http://www.r-project.org.

Wakaki, H., Eguchi, S., & Fujikoshi, Y. (1990). A class of tests for a general covariance structure. Journal of Multivariate Analysis, 32, 313–325.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics, 9, 60–62.

Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. Statistics in Medicine 1997; 16:2349 {2380.

Dixon WJ, (ed.). BMDP Statistical Software Manual, Volume 2. University of California Press: Berkeley, CA, 1988.

SAS Institute Inc. SAS=STATJ Software: Changes and Enhancements through Release 6.12. SAS Institute Inc.: Cary, NC, 1997.

Diggle PJ. An approach to the analysis of repeated measures. Biometrics 1988; 44:959 {971.

Wol_nger RD. Covariance structure selection in general mixed models. Communications in Statistics, Simulation and Computation 1993; 22(4):1079 {1106.

Dawson KS, Gennings C, Carter WH. Two graphical techniques useful in detecting correlation structure in repeated measures data. American Statistician 1997; 45:275 {283.

Milliken GA, Johnson DE. Analysis of Messy Data, Volume 1: Designed Experiments. Chapman and Hall: New York, 1992.

Searle SR, Casella G, McCulloch CE. Variance Components. Wiley: New York, 1991.

Cressie NAC. Statistics for Spatial Data. Wiley: New York, 1991.

Akaike H. A new look at the statistical model identi_cation. IEEE Transaction on Automatic control 1974; AC- 19:716 {723.

Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978; 6:461{ 464.

Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association 1995; 90:928 {934.

Huynh H, Feldt LS. Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split plot designs. Journal of Educational Statistics 1976; 1:69 {82.

Greenhouse SS, Geisser S. On methods in the analysis of pro_le data. Psychometrika 1959; 32:95 {112.

Littell RC, Milliken GA, Stroup WW, Wol_nger RD. SASJ System for Mixed Models. SAS Institute Inc.: Cary, NC, 1996.

Robinson GK. That BLUP is a good thing-the estimation of random e_ect. Statistical Science 1991; 6:15 {51.

Puntanen S, Styan GPH. On the equality of the ordinary least squares estimator and the best linear unbiased estimator. American Statistician 1989; 43:153 {164.

Little RJ, Rubin DB. Statistical Analysis with Missing Data. Wiley: New York, 1987.

Graybill FA. Theory and Application of the Linear Model. Wadsworth & Brooks=Cole: Paci_c Grove, CA, 1976.

Fai AHT, Cornelius PL. Approximate F-tests of multiple degrees of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. Journal of Statistical Computing and Simulation 1996; 54:363 {378.

McLean RA, Sanders WL. Approximating degrees of freedom for standard errors in mixed linear models. In Proceedings of the Statistical Computing Section. American Statistical Association, 50 {59.

Hulting FL, Harville DA. Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small area estimation: computational aspects, frequentist properties, and relationships. Journal of the American Statistical Association 1991; 86:557 { 568.

Diggle PJ, Liang K-Y, Zeger SL. Analysis of Longitudinal Data. Oxford University Press: New York, 1994.

Verbeke G, Molenberghs B. Linear Mixed Models in Practice. Springer: New York, 1997.