Contributions to Statistics

W. G. Müller · H. P. Wynn
A. A. Zhigljavsky (Eds.)

Model-Oriented Data Analysis

Physica-Verlag
A Springer-Verlag Company

W. G. Müller · H. P. Wynn
A. A. Zhigljavsky (Eds.)

Model-Oriented Data Analysis

Proceedings of the 3rd International Workshop in Petrodvorets, Russia, May 25-30, 1992

With 56 Figures

Physica-Verlag
A Springer-Verlag Company

Editors
Dr. Werner G. Müller
University of Economics and Business Administration
Augasse 2-6
A-1090 Vienna, Austria

Professor Dr. Henry P. Wynn
Department of Mathematics
City University
Northampton Square
GB-London EC1V OHB, Great Britain

Professor Dr. Anatoli A. Zhigljavsky
Department of Mathematics and Mechanics
St. Petersburg University
Bibliotechnaja sq. 2
St. Petersburg, Petrodvorets, 198904, Russia

ISBN 3-7908-0711-7 Physica-Verlag Heidelberg
ISBN 0-387-91457-9 Springer-Verlag New York

CIP-Titelaufnahme der Deutschen Bibliothek
Model oriented data analysis: proceedings of the 3rd international workshop in Petrodvorets, Russia, May 25 - 30, 1992 / [International Workshop on Model Oriented Data Analysis]. Werner G. Müller ... (ed.). - Heidelberg: Physica-Verl., 1993
(Contributions to statistics)
ISBN 3-7908-0711-7
NE: Müller, Werner G. [Hrsg.]; International Workshop on Model Oriented Data Analysis <03, 1992, Petergof>

Printing: Weihert-Druck, Darmstadt
Bookbindung: T. Gansert GmbH, Weinheim-Sulzbach
88/7130-543210 - Printed on acid-free paper

PREFACE

This volume contains the majority of papers presented at the Third Model-Oriented Data Analysis Workshop/Conference (MODA3) in Petrodvorets, Russia at 25.-30. May 1992. The previous two MODA workshops were held in Eisenach, East Germany in 1987 and in St.Kyrik, Bulgaria in 1990. These conferences, including the present one, cover theoretical and applied statistics with a heavy emphasis on experimental design. Under these broad headings other specialised topics can be mentioned, particularly quality improvement and optimization. The decision to hold MODA3 in St.Petersburg achieved an unanimous vote at the MODA2 Workshop as it was considered that it would provide an opportunity for scientists from the former Soviet Union to attend more easily. As history has evolved it was fortuitous that the opening-up of opportunities in the East more than fulfilled the ambitions of the organizing committee. In the event, there was a strong participation both from the East and West. Excellent on-the-ground organisation produced a pleasant environment for debate. An additional contributing factor was the fine location close to the summer palace at Petrodvorets.

Acknowledgement should be made to the Institute of Applied Systems Analysis in Laxenburg, Austria for providing support for the publication of these proceedings and to the initiator of this series of events Professor Valery Fedorov. The Department of Statistics of St.Petersburg University provided the chairman, Professor Sergei Ermakov, and the organizing committee. The conference also received the constant support of the Dean of Mathematics Professor Gennady Leonov. The participants and organisers of the conference gratefully acknowledge the financial support of The Procter and Gamble Company. This support was facilitated by the attendance at the conference of Dr. Michael Meredith.

This proceedings volume consists of three main parts:

I Optimal Design, II Statistical Applications, III Stochastic Optimization

A constant theme at MODA conferences is the subject of optimal experimental design. This was well-represented at MODA3 and readers will find important contributions. In recent years the models investigated under this heading have become progressively more complex and adaptive.

Several papers deal with the problem of designing experiments involving nonlinear models. A description of these methods including a number of applications to biological experiments stressing sequential procedures is given by C.Kitsos. A detailed consideration of sequential techniques is presented by L.Pronzato, E.Walter and C.Kulcsar. They provide a comparison of the efficiencies of different approaches in classical examples. The paper by A.C. Ponce de Leon and A.C. Atkinson investigates the design properties of generalized linear models when the link function is to be estimated simultaneously with the unknown parameters of the predictor. Illustrations are given of locally optimal and Bayesian designs. Particular cases, when the design problem in generalized linear models can be reduced to the classical situation, are described by B.Torsney and A.K.Musrati. A two stage sequential design procedure for the nonlinear Behrens-Fischer problem is proposed by R.Schwabe.

The following papers in this section are to be considered as contributions to the classical theory of optimal design for regression experiments. Ch. Müller gives asymptotical results for models with contaminated errors. V.P. Kozlov completes results on optimal designs for polynomial abel inversion accomplished by numerical examples. J.Lopez-Fidalgo presents and investigates a new design criterion, which extends the classical maximum variance criterion.

One of the major computational problems in constructing mixture designs is due to the nonorthogonality of the regression functions, which R.D. Hilgers resolves for some special cases. An orthogonality condition for nonproportional row column designs is considered from different sides including the randomization theory viewpoint in the paper of J.Kunert. Some new approaches for planning simulation experiments in queueing theory are discussed by V.Melas.

The second section of the proceedings contains a broad collection of statistical applications ranging from econometrics to biometrics. S.M. Ermakov and J.N. Kashtanov consider the Monte-Carlo estimation of functionals of stationary distributions of Markov chains. V.Fedorov, P.Hackl and W.G.Müller demonstrate empirically the advantages of choosing the weight function in a nonparametric regression according to a specified criterion at hand of a forecasting task. A similar problem is considered by A.V.Makshanov, who uses adaptive polynomial smoothing for time series data. Spectral estimation is applied by V.N.Fomin for extrapolation of stationary time series. A main part of so-called statistical safety theory is the detection of change points of particular random processes which is discussed in the paper by A.E. Kraskovsky. Examples of solutions of inverse problems arising in biological data analysis are given by A.G. Bart, N.P. Clochkova and V.M. Kozhanov. Important tools in applied statistical analysis are nonparametric sign and variance component techniques. The former is considered by G.I.Simonova and Yu.N.Tyurin whilst J.Volaufova concludes the section by surveying the latter emphasizing linear approaches.

There exist strong connections between the philosophies and methodologies in experimental design and stochastic optimization. Therefore many of the contributions in the third section are devoted to illuminating this interference. An introduction to a new branch of search techniques based on the study of ergodic processes is presented in the paper by H.P.Wynn and A.Zhigljavsky. Optimization algorithms for some simulation experiments are considered in the following two papers. G.Yin, H.M.Yan and S.X.C. Lou improve stochastic approximation for some manufacturing models by using ideas from perturbation analysis. An extension of ordinary perturbation analysis technique to a more general class of problems is given by N.Krivulin.

Markov chain optimization algorithms are discussed in a pair of contributions. R.Zielinski studies the simulated annealing algorithm with the help of Borel-Cantelli arguments. Advanced results on a particular global random search algorithm are obtained by A.S. Tikhomirov. Environmental applications of genetic optimization algorithms including discussions on efficiency are provided by J.Kettunen and M.Jalava. M.V. Chekmasov and M.V. Kondratovich prove that stratified sampling dominates independent sampling in global random search algorithms. An attempt to apply the Bellman approach for average optimization of one-dimensional local search algorithms is described by L.Pronzato and A.Zhigljavsky. The concluding paper by T.Kulakovskaja and A. Shamon studies some features of $n-$person cooperative market games for nonbalanced models of economy.

The editors acknowledge the help of numerous persons in the publication of these proceedings. Our special thank goes to Maxim Chekmasov, E.P. Andreeva, Christine Beiglböck and all the referees.

Vienna, April 1993 W.G.Müller, H.P.Wynn,A.A.Zhigljavsky

TABLE OF CONTENTS

PART I. OPTIMAL DESIGN

Adopting Sequential Procedures for Biological Experiments

Christos P. Kitsos

1 Introduction

The main target of this paper is to construct designs which estimate the desirable unknown parameter as well as possible. The ingredients are:

(i) The underlying model which links the covariates u and the parameter θ with the reponse y, and is supposed nonlinear in θ. As the model comes from the biological field of applications the parameter can be either:

- the "velocity of reaction", or
- the $100p$ percentile, L_p, say.

(ii) The optimality criterion, φ say, used to choose the experimental procedure.

We shall face models from biochemistry and experimental carginogens adopting the sequential procedure to estimate the different types of parameters mentioned above.

2 Background

It is usually assumed that the response y going with the covariate $u \in U \subseteq \mathbf{R}^k$ is linked with the parameter of interest $\theta \in \Theta \subseteq \mathbf{R}^p$ through a deterministic part $f(u,\theta)$ and a stochastic part e, known as error, which gives the regression form

$$y = f(u,\theta) + e \tag{2.1}$$

with $$\eta = E(y|u) = f(u,\theta)$$

The function $f(u,\theta)$ is in general non-linear and in this paper only nonlinear cases will be considered. In case of a binary response $y = 0$ or 1 then the link with the covariate u is through a probability model $T(u;\theta)$

$$P(y=1) = T(u;\theta) = 1 - P(y=0) \tag{2.2}$$

If $I(\theta,u)$ is the Fisher information matrix and ξ the design measure, from a family of design measures Ξ, Silvey (1980), then the average per observation information matrix $M = M(\theta,\xi)$ can be defined, Kitsos (1989). In principle the nonlinear problem is distinguished from the linear one on the fact that the matrix M suffers on this θ - dependence, while in the linear

case $M = M(\xi)$, see e.g Chaloner (1986), Rash (1988) and the survey paper Walter and Pronzato (1990). Now:

Let $\text{Mat}(s,p)\, 1 \leq s \leq p$ be the set of $s \times p$ matrices and $\text{NMat}\ (s,p)$ be the set of $s \times p$ nonnegative definite matrices. If $Q \in \text{NMat}(s,p)$ interest might focus on estimating a linear transformation $Q\theta$. Then the following operator J_Q can be considered on $M = M(\theta,\xi)$

$$J_Q(M) = QM^{-}Q^T \tag{2.3}$$

with M^{-} a generalized inverse of M and $Q^T \in \text{Mat}\ (p,s)$. Choose φ to be a convex decreasing function on N Mat (s,s). Then the design measure ξ^* is called φ - optimal iff

$$\varphi\{J_Q[M(\theta,\xi)]\} = \min_{\xi \in \Xi}\{\varphi\{QM^{-}(\theta,\xi)Q^T\}\} \tag{2.4}$$

$$= +\infty, \text{ when } QM^{-}Q^T \text{ is singular}$$

Traditional definitions of φ and Q might lead to well known optimality criteria ($D(\theta)$, $A(\theta)$, $E(\theta)$, among others with the notation θ to emphasize this θ - dependence in the non-linear case). With θ taking its true value, following Pukelsheim and Titterington (1983), the general local optimum experimental design can be stated as:

$$\text{minimize: } \varphi \circ J_Q \tag{2.5}$$

$$\text{subject to: } \xi \in \Xi, M(\theta,\xi) \in \text{NMat}(p,p)$$

see Kitsos (1986) for details.

Obtaining the minimum, especially in a local non-linear problem, is not feasible in practice via direct calculation.

Sequential procedures have been adopted both in the linear (Fedorov (1972), Wu and Wynn (1978)) and nonlinear (Kitsos (1986, 1989, 1992a)) cases.

The general dichotomous convergence theorem for the sequence of design measures ξ_n is not easily to be extended to the non-linear case. In nonlinear problems, interest is focussed rather on the sequence of estimates, $\hat{\theta}_n$ at stage n, the target being to converge to θ a.s as $n \to \infty$ when the sequence of matrices $M_n = M(\hat{\theta}_n,\xi)$ converges a.s to $M = M(\theta,\xi)$. See Wu (1985 b).

Although, under strong conditions, the sequence $D_n = \det M(\theta,\xi_n)$ converges to $D = \det M(\theta,\xi^*)$ it is rather difficult to obtain results for the sequence $\hat{D}_{n,n} = \det M(\hat{\theta}_n,\xi_n)$ with det(.) or log det(.) being a particular case of a criterion function φ.

With Φ being the Frechet directional derivative, when rp is differentiable, the appropriate sequential approach for Biological Based Experiments would be the following (Wu, 1985b) At stage i choose that u_i which minimizes

$$d_\varphi(\hat{\theta}_{i-1},\xi_{i-1},u_i) = \Phi\{M(\hat{\theta}_{i-1},\xi_{i-1}), I(\hat{\theta}_{i-1},u_i)\} \tag{2.6}$$

This sequential procedure is a generalization of a typical D-algorithm; Fedorov (1972,p101), Wu and Wynn (1978). The sequential procedure (2.6) is reduced to a fully-sequential one when the batch size b is $b = \dim\theta$ at each stage. As it has been discussed (Kitsos (1989)) stochastic approximation schemes are fully sequential procedures which lead to D-optimal design applying "steepest ascent", in a parallel way of scheme (2.6). The virtue of stochastic approximation is that it is Markovian in the sense that the choice of the next run depends only on the current situation. The martingale structure, Lai and Robbins (1979), of the stochastic approximation which leads to certain limiting theorems, is destroyed when the sequence is truncated in a predefined interval in the way that the obtained values outside that interval are truncated to the bounds of the interval, see Kitsos (1989) for the dilution series assessment.

We comment here that in this truncated stochastic approximation scheme the bounded martingale structure is reduced to an amart, ie asymptotic martingale, Edgar and Sucheston (1976), and therefore, this fully sequential procedure, can be written, technically, as a sum of a martingale and a vanishing (in L^I) amart. Therefore the stochastic approximation scheme, even truncated, can provide limiting results within the class of nonlinear sequential designs. The construction of the sequential design will be discussed on models from Biological Based Experiments for cases (2.1) in paragraph 3 and for case (2.2) in paragraph 4. D-optimality as a φ criterion, appears an aesthetic appeal in these biological applications, as it will be discussed in what follows.

3 Application on enzyme - kinetic models

From biochemistry and especially in enzyme-kinetic studies two models

(i) The first order growth (decay) curve with

$$\eta = \theta_0 \exp(\theta_1 u), \qquad u \in [T_1, T_2] \tag{3.1}$$

$$\theta_1 > 0 (\theta_1 < 0)$$

(ii) Saturation of an enzyme by its substrate, following the Michaelis - Menten equation

$$\eta = \frac{\theta_0 u}{\theta_1 + u}, \qquad u \in [C_1, C_2] \tag{3.2}$$

with:

u the concentration of substrate

η the initial velocity of reaction

θ_0 the maximum initial velocity

θ_1 Michaelis constant

θ_0 corresponds to the velocity attained when enzyme has been "saturated" by an infinite concentration of substrate, while θ_1 is numerically equal to the concentration of substrate for half-maximal initial velocity. In biological bibliography $\theta = (\theta_0, \theta_1)$ is often denoted by (V, K) or $(V_{\max}, K_m)$ or (V, K^n) while the covariate u is denoted by C, or C^n or S.

Both models are linear with respect to θ_0 and therefore their design should depend only on θ_1, Kitsos (1986). Moreover even the design for Mithscherish equation of diminishing returns, $\eta = \theta_2 + \theta_0 \exp(\theta_1 u)$, will depend only on θ_1. Indeed it has been proved early by Box and Lucas (1959) that for the first order curve under D-optimality the support points are $T_2 - 1/\theta_1, T_2$ if $\theta_1 > 0$ and $T_1, T_1 - 1/\theta_1$ when $\theta_1 < 0$, with weight $1/2$ at both points.

The sequential design for the first order growth law has been discussed by Kitsos (1989). The initial design was built up on the optimum design points, functions of the unknown parameters. Different batch sizes were examined among which a fully sequential design was obtained when one observation is added at each stage for each re-estimated optimum point with stopping rule as usually, $|\hat{\theta}_{n+1} - \hat{\theta}_n| < \varepsilon$.

A simulation study was performed for 1000 runs with the true, θ_T, vector of parameters to be $\theta_{1T} = 10$ and $\theta_{2T} = 1, 2, 3, 4$. As starting values, "far" from the true were given $\theta_{2T} \pm 2 > 0$. The fully sequential design provided smaller estimated mean square errors than the static design, Kitsos (1989). What is also of interest is the evaluation of coverage probabilities for constructed approximated confidence intervals for both θ_1 and θ_2 and θ_1, θ_2 individually. The confidence intervals were constructed by "pretending" that the average information matrix, related asymptotically to the variance - covariance matrix $C = C(\theta)$ as

$$C^{-1}(\theta) \cong nM(\theta, \xi) \cong nM(\hat{\theta}_n, \xi) = C^{-1}(\hat{\theta}_n) \tag{3.3}$$

We assume that C (θ) is obtained through independent observations, with $\hat{\theta}_n$ being the estimate at the final stage. Table 1 summarizes the results providing also the evaluation of $\ln \det M(\hat{\theta}_n, \xi)$, with $\hat{\theta}_n$ evaluated at the final stage ie when $n = 40$. The corresponding values for the static design, i.e. the design which allocates at one stage half of observations at the optimal design points, are also presented, proving that the fully sequential procedure provided satisfactory results. Michaelis - Menten equation was considered by Dowd and Riggs (1965) for different values of the parameters, while Duggleby (1979) faced the experimental design problem and compared different design applying Box and Lucas argument. Endrenyi and Chan (1981) obtained a D-optimal design, allocating half observations at the optimal design points,

$$u_1 = C_2, \qquad u_2 = \frac{0.5\theta_1 C_2}{\theta_1 + 0.5C_2} \tag{3.4}$$

That is the optimal design allocates half of observation at the highest practically attainable concentration, C_2, yielding the maximum velocity, $\eta(\max)$. The other half of observations should yield a velocity of magnitude $\eta(\max)/2$ which is obtained at concentration u_2 as above.

Duggleby (1981) obtained different results on the model, while Bates and Watts (1981) applied their design criterion based on curvature effects.

Currie (1982) obtained D-optimal designs for n design points of the geometrical form $u_i = ar^{i-1}$, with different values of a and r. These design points only by accident can be optimal points and therefore this design is neither static or sequential, nor optimal.

What we propose is to construct a sequential design for the Michaelis - Menten biological based experiments. That is:

- Devote a proportion, P_0, say, of the observations to the first stage, allocating $nP_0/2$ observations at the optimal design points (3.4).

  Get an estimate $\hat{\theta}_1$.

- Redesign at the new optimal through (3.4) design points, with batch size $b = \frac{(1-P_0)n}{2k}$ at each design point, with $k+1$ the total number of stages.

Now, if $k = 1$ a quasi - sequential design has been obtained Kitsos (1992a). When $b = 2$ a fully sequential procedure has been constructed.

4 Carcinogenic experiments

Carcinogenic risk assessment may be based on experimental data. The experimental dose-response relationship wherever a saturation mechanism is assumed can be described by the Michaelis- Menton function discussed above. Different models for estimating low risk dose have been discussed by Hartley and Sielken (1977). The so called Multi-stage model in experimental carcinogenesis is of the form

$$T(u;\theta) = 1 - \exp\{-\sum_{i=0}^{k} \theta_i u_i\}, \quad u \in [D_1, D_2] \tag{4.1}$$

with u presenting the "dose" and $T(u;\theta)$ the predicted response as in (2.2). As the number of experimental points is generally very low, Zapponi et al (1989) model (4.1) is reduced to the so called "one-hit" of the form

$$T(u;\theta) = 1 - \exp\{-(\theta_0 + \theta_1 u)\} \tag{4.2}$$

The low dose effect is of interest and therefore the $100p$ percentile, L_p say, has to be estimated. We are adopting a fully sequential procedure converging in mean square to L_p, namely a stochastic approximation scheme. It is easy to see that for model (4.2) the 100p percentile is

$$L_p = \frac{-1}{\theta_1}(\theta_0 + \ln(1-p)) \tag{4.3}$$

The first derivative of $T(.)$ at L_p is

$$T'(L_p) = \theta_1(1-p) \tag{4.4}$$

Devoting n_0 observations at the first stage it would be $T'(L_{p,k}) \approx \hat{\theta}_{1,k}$ $(1-p)$ for $k \geq n_0$ and therefore the iterative scheme

$$L_{p,n+1} = L_{p,n} - (n\hat{\theta}_{1,n}(1-p))^{-1}(y_n - p), \qquad n = n_0, n_{0+1} \ldots \tag{4.5}$$

is a typical stochastic approximation scheme, Robbins and Monro (1951), of the form $x_{n+1} = x_n - \alpha_n(y_n - p)$ with y_n as in (2.2), the appropriate sequence an, $\alpha_n, \alpha_n = \frac{c}{n}$ with optimal choice of $c, C_{op} = (T'(L_p))^{-1}$ leading to D-optimal design, Kitsos (1989).

Therefore $L_{p,n+1}$ converges in mean square to L_p ie

$$L_{p,n+1} \rightarrow^{m.s} L_p, \; n \rightarrow \infty \tag{4.6}$$

Notice that the term θ_0 is not included in the iteration ie the design does not depend on θ_0 as it is partially nonlinear for θ_0.

Table 1. First order growth law. 95 % Coverages probabilities (CP): for θ_1 and θ_2, for $\theta_2; \theta_{1T} = 10.0, n = 40, V(\varepsilon) = \sigma^2 = 1$. (1) : Static Design, (2) : Fully Sequential.

True θ_2	starting θ_2	C.P. For θ_1, and θ_2		C.P. For θ_2		$\ln \det M(\theta, 1/2)$	
		(1)	(2)	(1)	(2)	(1)	(2)
1.0	1.0	.943	.944	.954	.953	5.470	5.471
	3.0	.954	.954	.953	.955	5.096	5.463
2.0	2.0	.942	.936	.954	.935	6.606	6.605
	4.0	.945	.946	.946	.951	6.437	6.601
3.0	1.0	.947	.946	.958	.946	7.201	7.973
	3.0	.961	.956	.954	.950	7.991	7.990
	5.0	.943	.946	.941	.946	7.895	7.987
4.0	2.0	.957	.957	.967	.945	9.211	9.468
	4.0	.947	.941	.940	.940	9.478	9.478
	6.0	.953	.958	.954	.971	9.415	9.476

5 Discussion

We tackled different Biological Based models usually used in biochemistry and experimental carcinogens under the sequential principle of design. In biochemical models we assumed constant variance. In pharmacokinetic models it is sometimes assumed that Var $(y_1) = \sigma^2/W_i$ with $W_i = W_i(u, \theta)$, when the standard deviation is desirable to be proportional to its mean $W_i = f(u_1, \theta)^{-2}$. In this case too, D-optimality criterion, is of use.

Fully sequential procedures as stochastic approximation schemes provide satisfactory limiting results both theoretically and practically as the limit is "approached" with not too many iterations. The fully sequential procedure scheme was suggested for the Michaelis - Menten model. A converging stochastic approximation scheme is suggested for experimental carcinogens to prove that fully sequential procedure can be applied to different Biological Models.

References

1. Bates, D.M., Watts, D.G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares, Ann. Stat. 9,1152-1167.

2. Box, G.E.P, Lucas, H.L. (1959). Design of experiments in nonlinear situation. Biometrika, 49, 77-90.

3. Currie, D.J. (1982). Estimating Michaelis - Menten parameters, bias, variance and experimentai design. Biometrics, 38, 907-919.

4. Chaloner K. (1986). Optimal Bayesian design for nonlinear estimation. Technical Report 468, University of Minnesota, School of Statistics.

5. Dowd, J.E, Riggs, D.S (1965). A comparison of estimates of Michaelis - Menten kinetic constants from various linear transformations. The J. of Biol. Chemistry, 240, 863 - 869.

6. Duggleby, R. G (1979). Experimental designs for estimating the kinetic parameters for enzyme - catalysed Reactions. J. theor. Biol., 81, 671-684.

7. Duggleby, R. G. (1981). A nonlinear regression program for small computers. Analytical Biochemistry, 110, 9-18.

8. Edgar, G.A., Sucheston, L. (1976). Amarts: A class of asymptotic martingales A. discrete parameter. J. of Mult. Anal., 6, 193-221.

9. Endrenyi, L., Chan F.Y. (1981). Optimal design of experiments for the estimation of precise hyperbolic kinetic and binding parameters. J. theor. Biol., 90, 241-263.

10. Fedorov, V.V. (1972). Theory of Optimal Experiments. Academic Press, New York.

11. Ford, I., Kitsos, C.P. Titterington, D.M. (1989) Recent advances in nonlinear experimental design. Technometrics, 31, 49-60.

12. Kitsos, C.P. (1986). Design and inference in nonlinear problems. PhD thesis U. of Glasgow.

13. Kitsos, C.P. (1989). Fully sequential procedures in nonlinear design problems. Comput. Statistics and Data Analysis, 8, 13-19.

14. Kitsos, C.P. (1992). Quasi-sequential procedures for the calibration problem. COMPSTAT 92, Neuchâtel, Aug. 92.

15. Lai, T.L., Robbins, H. (1979). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. Z. Wahrscheinlichkeitstheorie, verw. Gebiete, 56, 329-360.

16. Pukelsheim, F, Titterington, D.M. (1983). General differential and lagrangian theory for optimal experiment design. Ann. Stat., 11, 1060-1068.

17. Rash D. (1988). Recent results in growth analysis. Statistics, 19, 585-604.

18. Robbins, H., Monro, S. (1951). A stochastic Approximation Method. Ann. Math. Stat., 22, 400-407.

19. Silvey, S.D. (1980). Optimal Design. Chapman and Hall.

20. Walter E., Pronzato L. (1990). Qualitative and quantitative experiment design for phenomenological models - a survey. Automatica, 26, 195-213.

21. White L.V. (1973). An extension of the general equivalence theorem to nonlinear models. Biometrika, 60, 345-348.

22. Wu, C.F.J, Wynn, H.P. (1978). The convergence of general step-length algorithms for regular optimum design criteria. Ann. Stat., 6, 1273-1285.

23. Wu, C.F.J. (1985a). Efficient sequential designs with binary data. JASA, 80, 974-984.

24. Wu, C.F.J. (1985b). Asymptotic inference from sequential design in a nonlinear situation. Biometrika, 72, 553-558.

25. Zapponi, A., Loizzo, A., Valente, P. (1989). Carcinogenic risk assessment: some comparisons between risk estimates derived from human and animal data. Exp. Pathol., 37, 1-4.

Acknowledgments

I would like to thank the unknown referee for his helpful comments.

A Dynamical-System Approach to Sequential Design

L. Pronzato, E. Walter and C. Kulcsár

"It has been suggested, with some irony, that the best time to design an experiment is after the experiment has been completed because one has then more knowledge on the process under study ... By designing an experiment sequentially, we can, in a sense, approximate this happy (but impossible) situation ..."

D. Steinberg and W. Hunter 38

1 Introduction

Sequential experimental design is a natural approach to face the dependence of the optimal experiment on the parameters to be estimated in the nonlinear case (see e.g. 17, 41). It aims at taking the information contained in previous observations into account when choosing new experimental conditions, i.e. new support points for the design. Classically, two phases are alternated: *estimation*, during which the data collected so far are used to obtain parameter estimates, and *design*, during which these estimates are used to select the best experimental conditions given the final purpose of the experiment (e.g. parameter estimation, hypothesis testing or model discrimination, response optimization, screening, search 32). Many intuitive schemes can be used, depending on the type of purpose considered (see the survey papers 13, 39). Here, we shall restrict our attention to parameter estimation (with some words on response optimization), but the methodologies and classification to be presented could be used in other contexts.

Section 2 is devoted to a tentative classification of sequential and nonsequential schemes. Connection with optimal stochastic control for dynamical systems is evidenced. *Closed-loop*, *open-loop feedback* and *open-loop* optimal strategies are considered. Section 3 deals with the most widely used batch sequential approach, based on a *heuristic certainty equivalence* principle. Convergence is studied, using the *Ordinary Differential Equation* method 26, in a situation where the experiments are sequentially applied to different physical systems (or individuals) in a population. The policy is shown to correspond to a Robbins-Monro stochastic approximation procedure. Fully sequential design is considered in Section 4. It corresponds to the case where a single support point is chosen after each new observation (see e.g. 17). Note the difference with the definition used in 22. An open-loop feedback policy is described in Section 5, also in the case where a population of physical systems is studied.

2 Classification of sequential-design policies

In this attempt to classify design policies we shall not distinguish batch-sequential from fully sequential approaches and exact from approximate designs. Most of this section is based on

3. The design to be chosen at step k (i.e. before the kth estimation stage) will be denoted by $\xi(k)$, with an admissible domain $\Xi(k)$. Let ξ_k^N denote the set of all designs chosen from step k to step N, with Ξ_k^N the corresponding admissible domain. The choice of the design criterion will not be discussed in this paper, and $J(\xi,\omega)$ will denote any criterion to be minimized with respect to ξ, with ω some characteristics of the process, unknown at this step (e.g. the value of the model parameters in a nonlinear regression problem). J may for instance be a scalar (convex or concave) function of the information matrix. Note that we shall call *information matrix* the matrix calculated as in a nonsequential procedure, although this is abusive due to the sequential nature of the design (see 19, 42). We consider the case where the total number N of design steps is fixed in advance. The *Closed-Loop Optimal* (CLO) solution is then obtained by solving the stochastic dynamical programming problem (see 4, 15)

$$\min_{\xi(1)\in\Xi(1)} \quad E_\omega\{\min_{\xi(2)\in\Xi(2)} E_\omega\{\dots E_\omega\{\min_{\xi(N)\in\Xi(N)} E_\omega\{J(\xi_1^N,\omega) \mid \mathcal{I}(N-1)\} \mid \mathcal{I}(N-2)\} \dots \mid \mathcal{I}(1)\} \mid \mathcal{I}(0)\}, \tag{1}$$

where $(\mathcal{I}(k))_k$ corresponds to an increasing sequence of σ-algebra and $\mathcal{I}(k)$ denotes all information concerning the system available at step k (i.e. prior knowledge $\mathcal{I}(0)$, observations $\mathbf{y}(1),\dots,\mathbf{y}(k)$ and experimental conditions $\xi(1),\dots,\xi(k)$). Once $\xi(1)$ has been applied, the problem is considered again, starting at $\xi(2)$ with prior information $\mathcal{I}(1)$.

Example 1: Consider the regression model

$$y(x) = a\exp(-\theta x) + \epsilon(x)\,,\ a>0\,, \tag{2}$$

where a is assumed to be known and θ is a scalar parameter to be estimated. The measurement errors $\epsilon(x)$ are independently normally distributed $\mathcal{N}(0,\sigma^2),\sigma = 0.1$. The information $\mathcal{I}(0)$ about θ corresponds to a discrete uniform prior $\pi^0(\theta)$ over $\Theta = \{1, 1.02, 1.04, \dots, 1.96, 1.98, 2\}$. We consider exact ELD-optimal design (see e.g. 41), i.e.

$$x_{ELD} = \arg\min_{x\in X} E_\theta\{-\ln\det M(\theta,x)\}\,,$$

with $M(\theta,x) = \frac{a^2x^2}{\sigma^2}\exp(-2\theta x)$ the information matrix for one measurement (here a scalar). Although a single measurement would suffice to estimate θ, we shall assume throughout this example that two measurements are allowed. With $N=2$ and the feasible domains for the two measurements given by $X(1)=[0,1], X(2)=[x_1,1]$ (let us say that x is time and that the second measurement cannot take place before the first one) we obtain for the CLO policy

$$x_{1\,ELD}^{CLO} = \arg\min_{x_1\in[0,1]} E_{y(1)}\{\min_{x_2\in[x_1,1]} E_\theta\{-\ln(\frac{a^2x_1^2}{\sigma^2}\exp(-2\theta x_1) + \frac{a^2x_2^2}{\sigma^2}\exp(-2\theta x_2)) \mid \mathcal{I}(1)\} \mid \mathcal{I}(0)\}\,,$$

where $E_\theta\{.\mid\mathcal{I}(1)\}$ is evaluated using the posterior distribution $\pi^1(\theta)$ (after one observation $y(1)$). As $y(1)$ is unknown at this stage, an expectation $E_{y(1)}\{.\mid\mathcal{I}(0)\}$ is performed, using the prior distribution π^0 and the distribution of measurement errors. Note that π^1 remains discrete, which is of special importance from a computational point of view (the weight α_i^0 of the ith support point θ_i for π^0 is simply updated into $\alpha_i^1 = \frac{\alpha_i^0 p(y(1)|\theta_i)}{\sum_i \alpha_i^0 p(y(1)|\theta_i)}$). A numerical

optimization (using two nested golden-search procedures and a numerical integration for the evaluation of $E_{y(1)}\{. \mid \mathcal{I}(0)\}$) yields

$$x_{1\,ELD}^{CLO} = 0.6443\,.$$

The value of $x_{2\,ELD}^{CLO}$ cannot be determined *a priori*, since it depends on the particular realization of $y(1)$.

◇

Even if the nonlinear programming problem associated with CLO design can theoretically be solved backward or forward in time 2, this is an extremely difficult task. To the best of our knowledge, within the experimental design context, the CLO approach has only been used in very simple situations. Zacks 45 considers a two-stage approach (i.e. $N = 2$). Bayard and Schumitzky 3 illustrate the feasibility of the forward-in-time approach by determining a classical D-optimal experiment for a nonlinear regression model (all expectations in (1) then disappear). A first suboptimal policy corresponds to *Open-Loop Feedback* (OLF) control 15, 40. The k-design step $\xi(k)$ is then chosen so as to minimize $E_\omega\{J(\xi_1^k,\omega) \mid \mathcal{I}(k-1)\}$, the designs in ξ_1^{k-1} being fixed. This policy is open-loop in the sense that the knowledge of the fact that the next design points will also be chosen sequentially is not taken into account (each design step is thus considered as the last one). It nevertheless contains feedback since the information $\mathcal{I}(k)$ is updated after new observations have been collected.

Example 1 (continued): Two support points are chosen sequentially according to an OLF policy. We obtain

$$x_{1\,ELD}^{OLF} = \arg \min_{x_1 \in [0,1]} E_\theta\{-\ln(\frac{a^2 x_1^2}{\sigma^2} \exp(-2\theta x_1))\}\,,$$

where the expectation E_θ is calculated with the prior distribution π^0. A numerical calculation gives

$$x_{1\,ELD}^{OLF} = 0.6667\,.$$

The value of $x_{2\,ELD}^{OLF}$ depends on the particular realization of $y(1)$, used to calculate π^1. Note that $x_{1\,ELD}^{OLF} > x_{1\,ELD}^{CLO}$, which could be expected since the OLF policy does not take advantage of the fact that a second measurement will be performed, with the constraint $x_2 \geq x_1$.

◇

An OLF policy will be considered in Section 5, in the context of population studies. Removing feedback in OLF control, one obtains an OL policy, which is nonsequential by nature,

$$\xi^{OL}{}_1^N = \arg \min_{\xi_1^N \in \Xi_1^N} E_\theta\{J(\xi_1^N,\omega) \mid \mathcal{I}(0)\}\,.$$

This has been widely considered in the literature under the name of optimal design *in the average sense* or *Bayesian design* (see e.g. 12 and the survey paper 41).

Example 1 (continued): The nonsequential exact ELD-optimal design with two measurements is given by

$$x_{1ELD}^{OL}, x_{2ELD}^{OL} = \arg \min_{x_1 \in [0,1], x_2 \in [x_1,1]} E_\theta\{-\ln(\frac{a^2 x_1^2}{\sigma^2} \exp(-2\theta x_1) + \frac{a^2 x_2^2}{\sigma^2} \exp(-2\theta x_2))\}.$$

A numerical calculation gives replicated measurements at

$$x_{1ELD}^{OL} = x_{2ELD}^{OL} = 0.6667.$$

◇

Finally, the most widely used sequential procedures correspond to a *Heuristic Certainty Equivalence* (HCE) control using feedback. At the kth design step, $\xi(k)$ is chosen so as to minimize $J(\xi_1^k, \hat{\omega}(k-1))$, where $\hat{\omega}(k-1)$ is a value of ω estimated from $\mathcal{I}(k-1)$ (e.g. $\hat{\omega}(k-1) = E\{\omega \mid \mathcal{I}(k-1)\}$), the designs in ξ_1^{k-1} being fixed. Applying this policy to Example 1, one obtains a sequential D-optimal design. This will be considered in Sections 3 and 4. Implementing it in open-loop without feedback, one gets classical D-optimal design.

In the context of parameter estimation for nonlinear regression models, OLF policies are almost as simple as HCE policies to implement. The main difficulty is to evaluate posterior densities for the unknown parameters (see 12). Simplifications can be achieved by considering either normal distributions and a linearization of the model response (as for instance in the Box and Hill approach to experimental design for model discrimination 9), or a discrete distribution for θ.

Contrary to the HCE or OLF policies, CLO design possesses the well known *dual effect* (see e.g. 16, 40): early support points may provide little gain in terms of precision on the parameter estimates, but may yield important knowledge about how to choose the next points. It is fully optimal, in the sense that all information about the past (from the prior distribution and previous measurements) and future (the fact that next support points will be chosen sequentially on the basis of the information to be collected now) is taken into account. The difficulty lies in the nested minimization and expectation steps. The procedure of *iteration in policy space*, suggested in 1, 2, 3, permits to reduce this complexity greatly by considering a CLO policy over a small number of steps (say $m << N$) and an open-loop policy for further steps. This is similar to ideas developed in the field of predictive control (see e.g. 6). For instance, the kth design $\xi(k)$ (with $m < N-k-1$) can be chosen so as to minimize

$$E_\omega \quad \{\min_{\xi(k+1) \in \Xi(k+1)} \dots E_\omega\{\min_{\xi(k+m) \in \Xi(k+m)} E_\omega\{\min_{\xi_{k+m+1}^N \in \Xi_{k+m+1}^N} E_\omega\{J(\xi_k^N, \omega) \mid \mathcal{I}(k+m)\} \mid \mathcal{I}(k+m-1)\} \dots \mid \mathcal{I}(k)\}.$$

The larger m is, the closer this policy gets to CLO design, which provides a trade-off between performances and computational complexity. It seems that the gain in performances progressively decreases while m increases. Choosing $m = 1$ or $m = 2$ should thus already yield the major part of what can be gained on the way from open-loop to closed-loop.

Remark 1 *In the context of bounded measurement errors with no information about the distribution of the errors between their bounds, mathematical expectations cannot be used. Average optimal design should then be replaced by worst-case optimal design ($E\{f(\omega) \mid \mathcal{I}\}$ being replaced by $\max_\omega[f(\omega) \mid \mathcal{I}]$). An open-loop policy is suggested in 24, 35, while the use of feedback is considered in 36.*

CLO strategies are fully optimal given the number N of design steps to be performed, and thus do not require any convergence study. On the opposite, convergence considerations are important for open-loop policies, for which N is not prespecified.

3 Convergence properties of classical sequential design

We consider the general situation where blocks of experiments may be sequentially performed on different physical systems (e.g. individuals in pharmacokinetics) that belong to a same population (in the sense that they can be described by models with the same structure but different parameter values). A classical procedure, corresponding to HCE design, consists for the next block in performing an optimal experiment for the empirical mean of the parameters estimated from each of the previous blocks.

Let θ^i denote the parameters of the ith individual to be considered. The observations $\mathbf{y}^i_X$, performed according to the exact design $X = \{x_1, x_2, \ldots, x_n\}$, are given by

$$\mathbf{y}^i_X = (y^i(x_1), \ldots, y^i(x_n))^T ,$$

with

$$y^i(x_j) = \eta(\theta^i, x_j) + \epsilon(\theta^i, x_j) ,$$

where $\eta(\theta, x)$ is the model response of an individual with parameters θ ($\theta \in \Theta$, open set of $\mathcal{R}^p$), for the design point x. We assume for simplicity that the same number n of observations is performed on each individual, $n \geq p = \dim \theta$. The measurement errors $\epsilon(\theta^i, x_j), j = 1, \ldots, n$, are assumed to be uncorrelated, with a (diagonal) covariance matrix $W_X(\theta^i)$. We assume that the underlying distribution is individual free, i.e. $\epsilon(\theta^i, x_j) = w(\theta^i, x_j, \omega^i_j)$, where w is a deterministic function and the ω^i_j's are i.i.d. random variables. For instance, the ω^i_j's can be normally distributed $\mathcal{N}(0,1)$, with $\epsilon(\theta^i, x_j) = (a \mid \eta(\theta^i, x_j) \mid^b + c)\omega^i_j$ and a, b, c given positive numbers. The design X is chosen on the basis of previous observations, i.e. using feedback.

In the context of pharmacokinetical experiments, where $\eta(\theta, x)$ is a nonlinear function of θ, D'Argenio 14 suggests to use

$$X^i = X_D(\hat{\theta}_m(i-1)) ,$$

where $X_D(\theta)$ is an exact D-optimal design for the parameters θ with n points of support, and where $\hat{\theta}_m(k)$ is given by

$$\hat{\theta}_m(k) = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}^i , \tag{3}$$

with $\hat{\theta}^i = \hat{\theta}(\mathbf{y}^i_{X^i})$ the value of θ estimated from the observations performed on the ith individual (e.g. using least-squares). X_D is obtained using the information matrix for normal errors,

$$M_X(\theta) = S^T_X(\theta) W^{-1}_X(\theta) S_X(\theta) ,$$

where

$$S^T_X(\theta) = \left(\frac{\partial \eta(\theta, x_1)}{\partial \theta}, \ldots, \frac{\partial \eta(\theta, x_n)}{\partial \theta} \right) ,$$

as

$$X^i = \arg\max_{X \in \mathcal{X}} \det M_X(\hat{\theta}_m(i-1)) ,$$

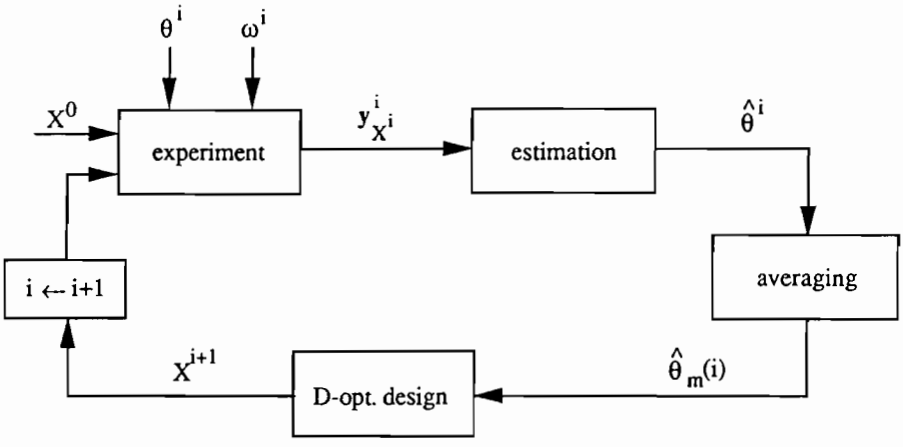


Figure 1: D-optimal design in a sequential population study.

with $\mathcal{X}$ the admissible experimental domain. The procedure is summarized by Figure 1.

Assume that the θ^{i}'s are i.i.d. random variables with distribution π. Define

$$\mu(i) = \hat{\theta}_m(i)\,,$$

and let $p_X(\hat{\theta} \mid \bar{\theta})$ denote the exact distribution of the estimates $\hat{\theta}(\mathbf{y}_X)$ when the model parameters take their true value $\bar{\theta}$. Using the *Ordinary Differential Equation* (ODE) method developed by Ljung 26, we can show that under some simple hypotheses (not detailed here for the sake of brevity, see 26) $\mu(i)$ asymptotically follows the trajectory of the deterministic differential equation defined by

$$\frac{d\mu(\tau)}{d\tau} = \mathbf{f}(\mu(\tau))\,, \tag{4}$$

where

$$\mathbf{f}(\bar{\mu}) = E\{\hat{\theta}(\mathbf{y}_{X_D(\bar{\mu})})\} - \bar{\mu} = \int_\theta \left(\int_{\hat{\theta}} \hat{\theta}\, p_{X_D(\bar{\mu})}(\hat{\theta} \mid \bar{\theta})\, d\hat{\theta} \right) \pi(\bar{\theta}) d\bar{\theta} - \bar{\mu}\,. \tag{5}$$

In most situations $\mathbf{f}(\bar{\mu})$ is finite for $\bar{\mu}$ in some set $\mathcal{D}_s$ (however, the possibility that $\mathbf{f}(\bar{\mu})$ does not exist will be illustrated by Example 2). The sequential policy described in Figure 1 thus corresponds to a Robbins-Monro stochastic approximation procedure for the determination of $\bar{\mu}$ satisfying $\mathbf{f}(\bar{\mu}) = 0$. It can converge only to stable stationary solutions of (4), and when these stable points are isolated the procedure cannot infinitely oscillate between them 26.

Remark 2 *An analytical expression for an approximation of $p_X(\hat{\theta} \mid \bar{\theta})$, more accurate than the classical normal approximation, is given e.g. in the survey paper 34. In some cases it even coincides with the exact distribution.*

The right-hand side of (5) can be written as

$$\mathbf{f}(\bar{\mu}) = E_\theta\{\mathbf{b}_{X_D(\bar{\mu})}(\theta)\} + E_\theta\{\theta\} - \bar{\mu}\,,$$

where the bias $\mathbf{b}_X(\bar{\theta}) = \int_{\hat{\theta}} (\hat{\theta} - \bar{\theta}) p_X(\hat{\theta} \mid \bar{\theta}) \, d\hat{\theta}$ *can be approximated by 10*

$$\mathbf{b}_X(\bar{\theta}) \simeq -\frac{1}{2} M_X^{-1}(\bar{\theta}) \frac{\partial \eta_X^T(\theta)}{\partial \theta}_{|\bar{\theta}} W_X^{-1} \mathbf{z}(\bar{\theta}) ,$$

with

$$z_i(\bar{\theta}) = trace\left(M_X^{-1}(\bar{\theta}) \frac{\partial^2 \eta(\theta, x_i)}{\partial \theta \partial \theta^T}_{|\bar{\theta}} \right) ,$$

and $\eta_X^T(\theta) = (\eta(\theta, x_1), \ldots, \eta(\theta, x_n))$.

An exhaustive study of the behaviour of (4) would be beyond the scope of this paper, and we shall simply present an illustrative example.

Example 2: Consider the one-dimensional regression model defined by (2) and assume first that the measurement errors $\epsilon(x)$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. The individual parameters θ^i are generated according to the normal distribution $\mathcal{N}(\theta^0, \sigma_\theta^2)$. Exact D-optimal design of size 1 is considered,

$$X_D^i = X_D(\mu(i-1)) = \frac{1}{\mu(i-1)} ,$$

(which is D-optimal e.g. for $\mathcal{X} = [0, \infty[$ and $\mu(i-1) > 0$). When $y_X^i < 0$, which occurs with a probability $P > 0 \, , \forall X \in \mathcal{X}$, the value of the least-squares estimate $\hat{\theta}(y_X^i)$ is infinite. This yields an infinite bias, so that (5) is not defined. In order to avoid this unrealistic situation, where negative observations can be obtained while the model response is always strictly positive, we shall assume now that the measurement errors are independently uniformly distributed in $[-\alpha\eta(\theta, x), \alpha\eta(\theta, x)]$, with $0 \leq \alpha < 1$. The least-squares estimator, given by $\hat{\theta}(y_X) = \frac{1}{x} \ln \frac{a}{y_X}$, is then always finite, provided that $x \neq 0$. (Note that it does not coincide with the maximum likelihood estimator, given by $\hat{\theta}_{ML}(y_X) = \frac{1}{x} \ln \frac{(1+\alpha)a}{y_X}$). Simple algebraic calculations give

$$f(\bar{\mu}) = \theta^0 + \frac{1}{x(\bar{\mu})} + \frac{\rho(\alpha)}{x(\bar{\mu})} - \bar{\mu} , \tag{6}$$

where

$$\rho(\alpha) = \frac{1}{2\alpha} \left((1-\alpha)\ln(1-\alpha) - (1+\alpha)\ln(1+\alpha) \right) ,$$

and where $x(\bar{\mu})$ defines the design policy used. For instance, $x(\bar{\mu}) = \frac{1}{\bar{\mu}}$ yields $f(\bar{\mu}) = \theta^0 + \rho(\alpha)\bar{\mu}$. The differential equation (4) is then globally stable since $\rho(\alpha)$ is negative, $\alpha \in [0, \infty[$. From 26, the procedure converges with probability 1 to the unique stable stationary solution given by $\mu^* = -\frac{\theta^0}{\rho(\alpha)}$. Note that it depends on the distribution of the θ^i's only through its mean θ^0. The evolution of $-\frac{1}{\rho(\alpha)}$ is given in Figure 2. One always has $\mu^* \geq \theta^0 \, , \alpha \in [0, \infty[$. The sequence $\hat{\theta}_m(k)$ converges to θ^0 only when α tends to 0 (i.e. when there is no measurement noise). From (6), when $\alpha \neq 0$ no design policy $x(\bar{\mu})$ permits to converge to θ^0.

◇

The sequential-design procedure described in Figure 1 generally does not converge to the optimal experiment for the mean value of the parameters in the population. The design should

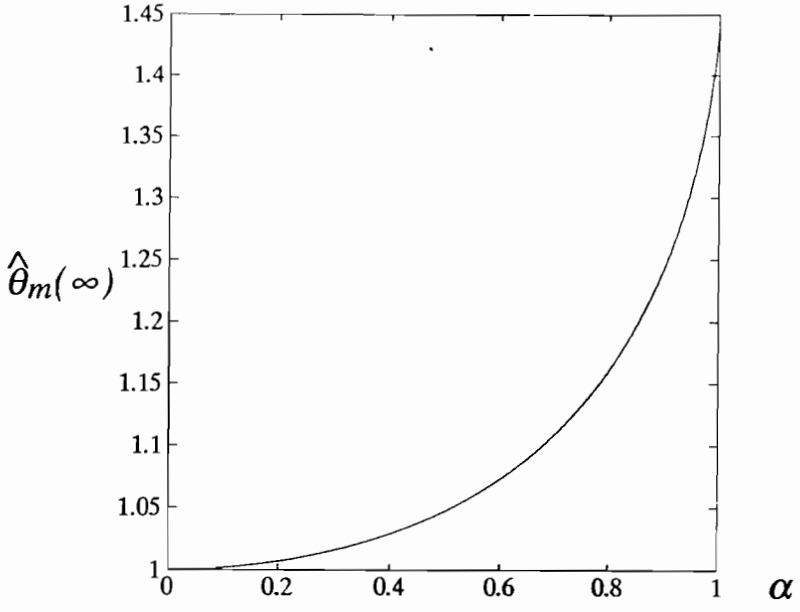


Figure 2: Convergence of $\hat{\theta}_m(k)$ in **Example 2, with** $\theta^0 = 1$.

thus not be based on the empirical mean of the parameters in the population (3). Another estimation of this mean (unbiased if possible) should be used, possibly together with some other characteristics of the distribution (see e.g. 37 for the estimation of such characteristics). A sequential approach based on an OLF control, with satisfying convergence properties, will be considered in Section 5.

Remark 3 *If all experiments were performed on the same individual with parameters* θ^0 *(which corresponds to classical batch sequential design), then convergence of the design could be studied within the same setting. The procedure would then correspond to a stochastic approximation scheme for estimating* $\bar{\mu}$ *satisfying* $\mathbf{f}(\bar{\mu}) = 0$, *with* $\pi(\bar{\theta}) = \delta(\bar{\theta} - \theta^0)$ *in (5). In the case of Example 2, the conclusions drawn from Figure 2 would remain the same.*

4 Fully sequential design

In this section, we restrict our attention to linear models, i.e.

$$\eta(\theta, \mathbf{x}) = \mathbf{x}^T\theta\,.$$

The experiments are performed on a single process with parameters θ, to be estimated using unweighted least-squares. The additive measurement are assumed i.i.d. Our aim is to give a summary of the convergence results available so far in this simple context and to point at some open problems. The information matrix at step $k+1$ is given by $M(k+1) = M(k) + \mathbf{x}(k+1)\mathbf{x}^T(k+1)$, and, denoting the average information matrix per sample by $R(k)$,

we have

$$\left\{ \begin{array}{rcl} R(k+1) & = & R(k) + \frac{1}{k+1}(\mathbf{x}(k+1)\mathbf{x}^T(k+1) - R(k)) \,, \\ \hat{\theta}(k+1) & = & \hat{\theta}(k) + \frac{1}{k+1} R^{-1}(k+1)\mathbf{x}(k+1)(y(k+1) - \mathbf{x}^T(k+1)\hat{\theta}(k)) \,, \end{array} \right. \tag{7}$$

with $\hat{\theta}(k)$ the parameter estimates at step k, and $y(k)$ the kth observation. Note that although $R(k)$ does not explicitly depend on θ, it may depend on the previous values of the estimates of θ through the regressors. The system (7) corresponds to the well known equations for recursive least-squares estimation. Differentw situations can be distinguished, leading to different approaches for the study of the convergence of the least-squares estimator and of the design.

First, $\mathbf{x}(k+1)$ may be a random variable independent of the past values of $\hat{\theta}$, R, $\mathbf{x}$ and y. This is classical in least-squares estimation, the convergence then depends on the distribution of the regressors (the convergence condition is one of persistency of excitation).

Second, $\mathbf{x}(k+1)$ may depend only on $y(k), y(k-1), \ldots$ and $\mathbf{x}(k), \mathbf{x}(k-1), \ldots$ as for auto-regressive models with exogeneous inputs. The convergence issue in this case is considered e.g. in 26, and also in 20 when $\mathbf{x}(k+1)$ is chosen so as to maximize $\det R(k+1)$.

Third, $\mathbf{x}(k+1)$ may depend only on $R(k)$. Consider for instance the situation

$$\mathbf{x}(k+1) = \arg\max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^T R^{-1}(k) \mathbf{x} \,,$$

which corresponds to the classical Wynn algorithm 44 for the construction of a D-optimal design measure. Convergence is proved several experimental design criteria e.g. in 43, 33.

Fourth, $\mathbf{x}(k+1)$ may depend only on $\hat{\theta}(k)$. Consider for instance the linear regression model $\eta(\theta, x) = \theta_0 + \theta_1 x + \theta_2 x^2$, where $x(k+1)$ is chosen so as to maximize $\eta(\hat{\theta}(k), x)$, i.e. $x(k+1) = -\frac{\hat{\theta}_1(k)}{2\hat{\theta}_2(k)}$ (with $\hat{\theta}_2(k)$ assumed to be negative). This corresponds to a self-tuning optimizer, whose convergence properties are studied in 11 using the ODE method. Convergence of $\hat{\theta}(k)$ towards θ is guaranteed only when a modified control policy is used, such as $x(k+1) = -\frac{\hat{\theta}_1(k)}{2\hat{\theta}_2(k)} + v(k+1)$, with the $v(k)$'s corresponding to a sequence of independent random variables, possibly with decreasing variance (this can again be interpreted as a condition of persistency of excitation).

Finally, a fifth situation is when $\mathbf{x}(k+1)$ depends both on $R(k)$ and $\hat{\theta}(k)$. For instance, one may wish to estimate $\mathbf{s}(\theta)$, a nonlinear vector function of θ. Let $\mathbf{s}'(\theta)$ denote $\frac{d\mathbf{s}^T(\theta)}{d\theta}$. The next design point $\mathbf{x}(k+1)$ can then be chosen in order to maximize $\Phi(M(k) + \mathbf{x}\mathbf{x}^T, \hat{\theta}(k)) = \phi[\mathbf{s}'^T(\hat{\theta}(k))(M(k) + \mathbf{x}\mathbf{x}^T)\mathbf{s}'(\hat{\theta}(k))]$, where $\phi[.]$ defines a scalar optimality criterion. Another choice (steepest-ascent approach), often leading to simpler calculations, is to maximize the Fréchet derivative of Φ at $M(k)$ and $\hat{\theta}(k)$ in the direction $\mathbf{x}\mathbf{x}^T$, i.e. to maximize $\lim_{\lambda \to 0+} \lambda^{-1} \Phi((1-\lambda)M(k) + \lambda \mathbf{x}\mathbf{x}^T, \hat{\theta}(k)) - \Phi(M(k), \hat{\theta}(k))$. The problem is considered in 42, but no general convergence result is obtained (except concerning the example treated in 18). The ODE method does not apply directly here (partly due to the fact that $\mathbf{x}(k+1)$ does not depend continuously on $R(k)$ and $\hat{\theta}(k)$). Further investigations are thus required, which could rely e.g. on the results in 23, 31, 5.

5 Open-loop feedback design in population studies

Following an OLF policy as suggested in Section 2, the design at step k will now be chosen on the basis of an estimate $\hat{\pi}(k-1)$ of the distribution of the individual parameters. The OLF

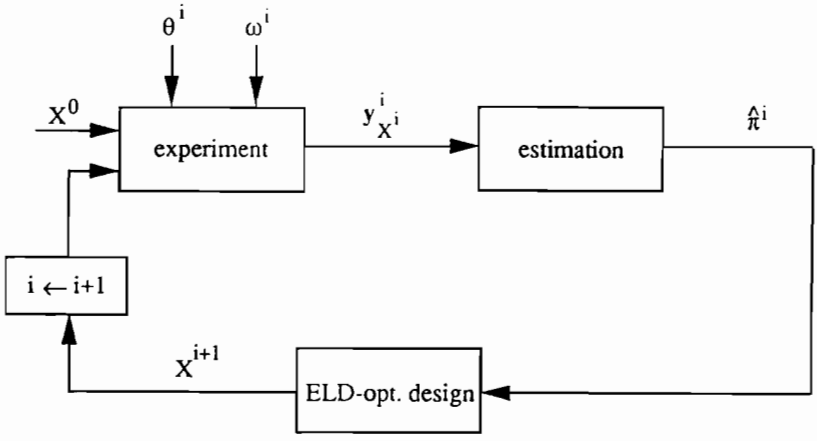


Figure 3: ELD-optimal design in a sequential population study.

procedure with ELD-optimal design is summarized in Figure 3. The design X^i is defined by

$$X^i = \arg\max_{x \in \mathcal{X}} E_\theta\{\ln \det M_X(\theta) \mid \hat{\pi}^{i-1}\}\,. \tag{8}$$

The estimation of $\hat{\pi}^i$ can be performed using e.g. the *maximum likelihood estimator for mixtures* 25. No details can be given here due to space limitation, and we can simply note the following points.

(i) The estimation is then not recursive: all previous observations performed on all previous individuals must be used at each step. A recursive determination of the distribution could be obtained through a parametrization, e.g. $\hat{\pi}^i$ could be searched within the class of normal distributions $\mathcal{N}(\hat{\theta}^i, \Omega^i)$, with a stochastic approximation method for updating the parameters $\hat{\theta}^i, \Omega^i$ of the distribution (see 30).

(ii) The maximum likelihood estimator corresponds to a discrete distribution, with a number of support points less than or equal to the number of individuals considered so far. As a consequence, the optimal design (8) can easily be determined (without requiring the use of numerical integration routines for the evaluation of the expectation).

(iii) The determination of the maximum likelihood distribution $\hat{\pi}^i$ can be performed with algorithms closely connected to those used in the design context (approximate theory) 7, 28, 8.

(iv) The problem of unicity of $\hat{\pi}^i$ is considered e.g. in 25, 28.

(v) Consistency of the maximum likelihood estimator is considered in 21.

Example 3: Consider again the regression model defined by (2), with $a = 10$, measurement errors i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 1$. A population of 100 individuals is considered, with the θ^i's i.i.d. $\mathcal{N}(\theta^0, \sigma_\theta^2)$, $\theta^0 = 1$, $\sigma_\theta = 0.1$. One measurement is performed on each individual. In this particular case, ELD-optimal design corresponds to D-optimal design for the mean

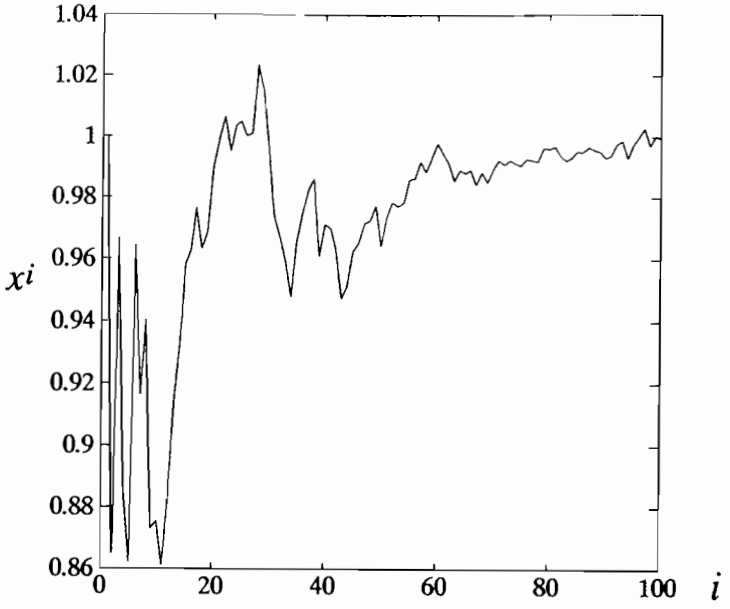


Figure 4: Evolution of x^i in Example 3.

value of θ, i.e.

$$x^i = \arg\max_{x \in \mathcal{X}} \frac{x^2}{\sigma^2} \exp(-2x E_\theta\{\theta \mid \hat{\pi}^{i-1}\}),$$

(the OLF policy thus coincides here with a HCE control). The optimal value x^* for the true distribution π is $x^* = 1$ (since $E_\theta\{\theta \mid \pi\} = \theta^0 = 1$). Figure 4 presents the evolution of x^i. The support point is seen to converge to the optimal design for the population.

◇

Remark 4 *A completely different situation would correspond to sequential design for estimating the distribution π itself. A characterization of the precision of this estimation would thus be required. First attempts in this direction (although in a nonsequential context) seem to be 27, 29.*

References

[1] D. Bayard. Proof of the quasi-adaptivity for the m-measurement feedback class of stochastic control policies. *IEEE Transactions on Automatic Control*, AC-32(5):447–451, 1987.

[2] D. Bayard. A forward method for optimal stochastic nonlinear and adaptive control. In *Proc. 27th Conf. on Decision and Control*, pages 280–285, Austin, Texas, December 1988.

[3] D. Bayard and A. Schumitzky. A stochastic control approach to optimal sampling design. Technical Report 90-1, Lab. of Applied Pharmacokinetics, School of Medicine, University of Southern California, Los Angeles, 1990.

[4] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, N.J., 1957.

[5] A. Benveniste, M. Metivier, and P. Priouret. *Algorithmes Adaptatifs et Approximations Stochastiques.* Masson, Paris, 1987.

[6] R. Bitmead, M. Gevers, and V. Wertz. *Adaptive Optimal Control, The Thinking Man's GPC.* Prentice Hall, New York, 1990.

[7] D. Böhning. Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference*, 11:57–69, 1985.

[8] D. Böhning. Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. *Biometrika*, 76(2):375–383, 1989.

[9] G. Box and W. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

[10] M. Box. Bias in nonlinear estimation. *Journal of Royal Statistical Society*, B33:171–201, 1971.

[11] A. Bozin and M. Zarrop. Self tuning optimizer — convergence and robustness properties. In *Proc. 1st European Control Conf.*, pages 672–677, Grenoble, July 1991.

[12] K. Chaloner. Optimal bayesian design for nonlinear estimation. Technical Report 468, School of Statistics, University of Minnesota, 1986.

[13] H. Chernoff. Approaches in sequential design of experiments. In J. Srivastava, editor, *A Survey of Statistical Design and Linear Models*, pages 67–90. North Holland, Amsterdam, 1975.

[14] D. D'Argenio. Optimal sampling times for pharmacokinetic experiments. *Journal of Pharmacokinetics and Biopharmaceutics*, 9(6):739–756, 1981.

[15] S. Dreyfus. *Dynamic Programming and the Calculus of Variations.* Academic Press, New York, 1965.

[16] A. Feldbaum. *Optimal Control Systems.* Academic Press, New York, 1965.

[17] I. Ford, C. Kitsos, and D. Titterington. Recent advances in nonlinear experimental design. *Technometrics*, 31(1):49–60, 1989.

[18] I. Ford and S. Silvey. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67(2):381–388, 1980.

[19] I. Ford, D. Titterington, and C. Wu. Inference and sequential design. *Biometrika*, 72(3):545–551, 1985.

[20] G. Goodwin and R. Payne. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, New York, 1977.

[21] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Math. Stat.*, 27:887–906, 1956.

[22] C. Kitsos. Fully sequential procedures in nonlinear design problems. *Computational Statistics and Data Analysis*, 8:13–19, 1989.

[23] H. Kushner and A. Shwartz. An invariant measure approach to the convergence of stochastic approximation with state dependent noise. *SIAM J. Control and Optimization*, 22(1):13–27, 1984.

[24] E. Landaw. Robust sampling designs for compartmental models under large prior eigenvalue uncertainties. In J. Eisenfeld and C. DeLisi, editors, *Mathematics and Computers in Biomedical Applications, IMACS*, pages 181–187. Elsevier, Amsterdam, 1985.

[25] B. Lindsay. The geometry of mixture likelihoods: a general theory. *Annals of Statistics*, 11(1):86–94, 1983.

[26] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, AC-22(4):551–575, 1977.

[27] A. Mallet. Méthodes d'estimation de lois à partir d'observations indirectes d'un échantillon : application aux caractéristiques de population de modèles biologiques. Thèse de Doctorat d'Etat, Université Pierre et Marie Curie, Paris 6, 1982.

[28] A. Mallet. A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, 73(3):645–656, 1986.

[29] A. Mallet and F. Mentré. An approach to the design of experiments for estimating the distribution of parameters in random models. In *Prep. 12th IMACS World Congress on Scientific Computation*, pages 134–137, Paris, July 1988.

[30] F. Mentré. Apprentissage de la loi de probabilité des paramètres d'un modèle par approximation stochastique. Thèse de Doctorat de 3ème cycle, Université Paris 7, 1984.

[31] M. Metivier and P. Priouret. Application of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, IT-30(2):140–151, 1984.

[32] J. O'Geran, H. Wynn, and A. Zhiglyavsky. Search. *Acta Applicandae Mathematicae*, 25:241–276, 1991.

[33] A. Pázman. *Foundations of Optimum Experimental Design*. VEDA, (co pub. Reidel, Dordrecht), Bratislava, 1986.

[34] A. Pázman. Small-sample distributional properties of nonlinear regression estimators (a geometric approach). *Statistics*, 21(3):323–367 (with discussion), 1990.

[35] L. Pronzato and E. Walter. Robust experiment design via maximin optimization. *Mathematical Biosciences*, 89:161–176, 1988.

[36] L. Pronzato and E. Walter. Sequential experimental design for parameter bounding. In *Proc. 1st European Control Conf.*, pages 1181–1186, Grenoble, July 1991.

[37] J. Steimer, A. Mallet, J. Golmard, and J. Boivieux. Alternative approaches to estimation of population pharmacokinetic parameters; comparison with NONMEM. *Drug. Metab. Review*, 15(14):265–292, 1984.

[38] D. Steinberg and W. Hunter. Experimental design: review and comment. *Technometrics*, 26(2):71–97, 1984.

[39] D. Titterington. Aspects of optimal design in dynamic systems. *Technometrics*, 22(3):287–299, 1980.

[40] E. Tse and Y. Bar-Shalom. An actively adaptive control for linear systems with random parameters via the dual control approach. *IEEE Transactions on Automatic Control*, AC-18(2):109–117, 1973.

[41] E. Walter and L. Pronzato. Qualitative and quantitative experiment design for phenomenological models — a survey. *Automatica*, 26(2):195–213, 1990.

[42] C. Wu. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72(3):553–558, 1985.

[43] C. Wu and H. Wynn. The convergence of general step-length algorithms for regular design criteria. *Annals of Statistics*, 6:1273–1285, 1978.

[44] H. Wynn. The sequential generation of D-optimum experimental designs. *Annals of Math. Stat.*, 41:1655–1664, 1970.

[45] S. Zacks. Problems and approaches in design of experiments for estimation and testing in nonlinear models. In P. Krishnaiah, editor, *Multivariate Analysis IV*, pages 209–223. North Holland, Amsterdam, 1977.

Designing Optimal Experiments for the Choice of Link Function for a Binary Data Model

Antonio C. Ponce de Leon and Antony C. Atkinson

The inclusion of an extra parameter to extend the link function in the modelling of binary data is investigated. The extended link includes the logistic and complementary log-log as special cases. We adress the problem of designing optimal experiments in this framework. We discuss the advantages of the approach which allows the choice among designs to estimate the link function parameter, the linear predictor parameters or both. Each design requires a different criterion function. Prior information is incorporated in the design criteria which depend on the parameters being estimated. Examples of locally optimal as well as optimal Bayesian designs are provided to illustrate the methods.

1 Introduction

Throughout this paper, the binary data models considered are a subclass of Generalized Linear Models. Thus, we adopt the terminology and notation of McCullagh & Nelder (1989) for modelling, but use optimal design theory notation elsewhere.

Suppose that a random sample $Y_1, ..., Y_n$ is to be observed, where Y_i follows a Binomial (m_i, π_i) distribution and that our main interest lies in the modelling of the relationship between the probability of success π_i and a set of covariates $\{x_{i1}, ..., x_{ip}\}$, $i = 1, ..., n$. According to generalized linear model assumptions, this relationship is described by a vector of linear predictors $\eta_\sim$, where $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ (p denotes the dimension of the vector of unknown parameters $\beta_\sim$) and a monotonic and differentiable function $g(.)$, called the link function, such that $\eta_i = g(\pi_i)$, $i = 1, ..., n$. Then the log likelihood function for the binomial distribution is

$$L(\pi_\sim, m_\sim; y_\sim) = \sum_{i=1}^{n} [y_i \log\left(\frac{\pi_i}{l - \pi_i}\right) + m_i \log(1 - \pi_i)] \tag{1.1}$$

where $\pi_i = g^{-1}(\eta_i)$. The term that does not depend on $\pi_\sim$ can be neglected.

In the modelling of binary data the choice of link function plays an important role, as shown in the examples of Section 2. However, only three functions are widely used in most applications. They are the logistic (logit), the inverse normal (probit) and the complementary log-log functions.

In this paper we suppose that for each experiment the set of covariates as well as the levels at which they are to be observed, may be chosen freely in a certain design region. Under these assumptions, the experimenter is capable of choosing a design that optimizes a given criterion emphasizing either the choice of link function or the estimation of the linear predictor. However, the link function specification should come first in any list of priorities,

for the estimation of the linear predictor parameters clearly depends on the link function. Having said that, there might be situations in which two or more link functions might fit the data rather well, even though the linear predictor parameter estimates differ significantly for each link, so as making the estimation of $\beta_{\sim}$ more crucial.

As far as optimal design is concerned, the problem of determining optimal experiments to estimate the parameters $\beta_{\sim}$ has been dealt with, for instance, by Chaloner & Larntz (1989). They presented examples of optimal Bayesian designs to estimate the linear predictor parameters $\beta_{\sim}$ as well as functions of them, such as the LD50 and LD95, in the case of logistic models. The problem of designing experiments for discriminating between two binary data models, more specifically models with different link functions, has been studied by Ponce de Leon & Atkinson (1992a) and (1992b), who give further references.

In fact, for binary data models both parameter estimation and model discrimination may be regarded as particular cases of a more generally formulated problem. In the next section we present a generalized link function for binary data models that brings a new insight to these two problems. The emphasis in this paper is on principles and numerical results. Full analytical details are given in Ponce de Leon (1992).

2 Generalized link function

Suppose that the link is the generalized link function (McCullagh & Nelder, *p.*378)

$$\eta_i = g(\pi_i, \lambda) = \log\left[\left\{\left(\frac{1}{1-\pi_i}\right)^{\lambda} - 1\right\} / \lambda\right], (\lambda \geq 0) \tag{2.1}$$

When $\lambda = 1$(2.1) reduces to the logistic link. Furthermore, it is straightforward to prove that $\lim_{\lambda \to 0} \eta_i = \log\{-\log(1-\pi_i)\}$, the complementary log – log link function.

So, the value of λ restricted to the interval $[0,1]$, may be interpreted as a measure of the distance between the logistic and the complementrary log – log models. Nevertheless, there is no reason whats ever for restricting attention to models in which λ belongs to this interval, for any non-negative value of λ is a potential candidate to provide a reasonable fit for binary data. Negative values of λ are not considered as numerical problems may arise in the computation of the link function inverse.

To illustrate how the maximum likelihood estimate for λ can be obtained we present two numerical examples using real data sets. It is interesting to notice that in both examples, $\hat{\lambda}$ lies outside the interval $[0,1]$.

Example 2.1 : Toxicity of Rotenone to *Macrosiphoniella sanborni*. Finney (1947, Ex.1, *p.*26, Table 2) gives the details. To investigate how the value of λ affects the goodness of fit for a binary data set, the criterion adopted was the deviance. The estimation of λ and $\beta_{\sim}$ was carried out in two steps. Firstly, we may suppose that the value of λ is known and proceed to estimate $\beta_{\sim}$. Next, we maximize the log likelihood *w.r.t.* λ, or equivalently minimize the deviance. Such a procedure is analogous to the so-called nested least squares. Analytical solutions for this problem appear to be very complicated. However, a numerical search for the minimal deviance over a grid for λ is effective and easy to implement. The model fitted to the rotenone data had the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, where $\{x_i\}$ are values of log concentration of rotenone. For each value of λ, the parameters β_0 and β_1 were estimated by iterative weighted least squares as described in McCullagh & Nelder (1989), pp. 40-43.

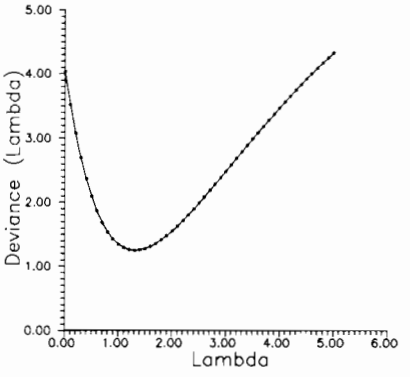


Figure 2.1 – Deviance of fit as a function of λ Example 2.1

Figure 2.1 shows the deviances as a function of λ over a grid in the interval $[0.0, 5.0]$. Although the change in deviance is not large, values of λ in the interval (1.25,1.5) yield smaller deviances, the optimum lying near $\lambda = 1.3$. Further analysis ought to be carried out before regarding any particular model as suitable.

The linear predictor parameter estimates vary with the value of λ. For example, when $\lambda = 0$ (complementary log-log), $\hat{\beta}_0 = -3.512$ and $\hat{\beta}_1 = 4.426$ whereas for $\lambda = 1$ (logistic), $\hat{\beta}_0 = -4.839$ and $\hat{\beta}_1 = 7.068$. These results are used later to justify the assumption of prior distributions for the linear predictor parameters being conditioned on the value of λ. In fact, there seems to be a trend in the behaviour of $\hat{\beta}_0$ and $\hat{\beta}_1$ that could be investigated more carefully.

Example 2.2: Milicer & Szczotka (1966) give the number of schoolgirls having menstruated as a function of age. The data are reproduced by Aranda-Ordaz (1981, Table 2). As in the previous example the linear predictor was assumed to be simply $\eta = \beta_0 + \beta_1 x$. The procedure described in Example 2.1, to estimate λ, β_0 and β_1 was again applied.

The deviances are shown in Figure 2.2. The search was carried out over a grid in the interval $[0.2, 6.0]$. As can be seen, the value of λ that provides the smallest deviance lies in the interval (1.25,1.7), more precisely close to 1.475.

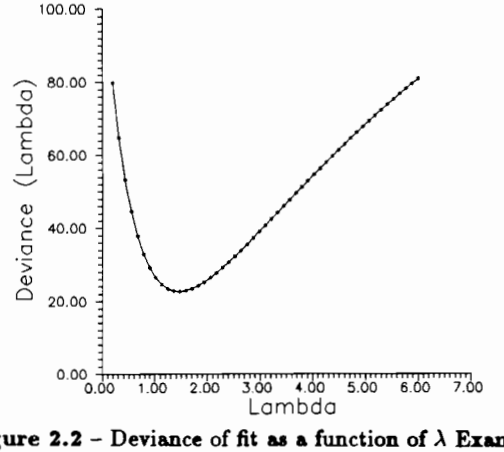


Figure 2.2 – Deviance of fit as a function of λ Example 2.2

In the next section we obtain Fisher's information matrix relative to the full problem, *i.e.* when λ is unknown. A brief discussion of optimal experimental design theory is also given.

3 Fisher's information matrix

Having analysed the results obtained in the examples presented in Section 2, we can now address the problem of designing optimal experiments to estimate λ and $\beta_{\sim}$. The main conclusion that can be drawn from these examples is that some importance must be given to the estimation of λ, whenever (2.1) is taken as the link function. Thus, from the optimal design theory point of view, it is sensible to focus the optimization on the estimation of λ rather than on the estimation of the linear predictor parameters $\beta_{\sim}$. Another possibility is to find the right balance between the two purposes.

To define the criterion functions, Fisher's information matrix for $(\lambda, \beta_{\sim})$ is required. After a series of differentiations of the log likelihood function (1.1) *w.r.t.* λ and $\beta_{\sim}$, we find that Fisher's information matrix is proportional to

$$M(\lambda, \eta_{\sim}(\beta_{\sim}), \xi_N) = \begin{bmatrix} \sum_1^n w_i \left(\frac{d\pi_i}{d\lambda}\right)^2 & \left(\sum_1^n w_i \frac{d\pi_i}{d\lambda}\frac{d\pi_i}{d\eta_i} x_{ir}\right)^t \\ \left(\sum_1^n w_i \frac{d\pi_i}{d\lambda}\frac{d\pi_i}{d\eta_i} x_{ir}\right) & \left(\sum_1^n w_i \left(\frac{d\pi_i}{d\eta_i}\right)^2 x_{ir} x_{is}\right) \end{bmatrix} \tag{3.1}$$

where

$$N = \sum_1^n m_i;\ w_i = \frac{p_i}{\pi_i(1-\pi_i)},\ p_i = \frac{m_i}{N}, i = 1, ..., n; \xi_N = \begin{Bmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{Bmatrix};$$

$$\frac{d\pi_i}{d\lambda} = \frac{-(\frac{1}{\lambda^2})\log(\lambda e^{\eta_i} + 1) + \frac{(1/\lambda)e^{\eta_i}}{\lambda e^{\eta_i}+1}}{(\lambda e^{\eta_i} + 1)^{1/\lambda}}, \quad i = 1, \dots, n;$$

$$\frac{d\pi_i}{d\eta_i} = \frac{\lambda e^{\eta_i}}{(\lambda e^{\eta_i} + 1)^{1/\lambda}}, \quad i = 1, \dots, n;\ r, s = 1, \dots, p.$$

The total number of observations N in the exact design ξ_N is assumed fixed. The weights $\{p_i\}$ may be interpreted as a measure of how informative the support points $\{x_{\sim_i}\}$ are for the estimation of the relevant parameter(*s*).

Exact design theory has been assumed so far. However, at this point we switch to the approximate theory in which the discrete design ξ_N is replaced by the design measure ξ over the design region $\mathcal{X}$. The details are given by Silvey (1980, *p.* 13,14). The advantage is in the properties of optimum approximate designs which, unlike discrete designs, satisfy an equivalence theorem (Kiefer & Wolfowitz, 1960). The information matrix (3.1) in terms of design measures becomes

$$M(\lambda, \beta_{\sim}, \xi) = \begin{bmatrix} M_{11} & M_{21}^t \\ M_{21} & M_{22} \end{bmatrix}$$

where

$M_{11} = \int w(\frac{d\pi}{d\lambda})^2 \xi(dx_{\sim}), M_{21} = [\int w \frac{d\pi}{d\eta}\frac{d\pi}{d\lambda} x_r \xi(dx_{\sim})]$ and $M_{22} = [\int w(\frac{d\pi}{d\eta})^2 x_r x_s \xi(dx_{\sim})]$.

The integrals are taken over the design region $\mathcal{X}$ and all variables have the same definition as before, but with no index. The exception is $w = 1/\{\pi(1-\pi)\}$.

In the next section we consider the choice of criterion function appropriate to the aim of the experiment. We also determine the derivative functions that are essential for checking optimality. Numerical examples are provided so as to illustrate the search for the optimum design.

4 The choice of criterion function

There are several purposes of interest in a binary data experiment, such as estimating a subset of parameters, estimating the LD50, or any other percentile, or estimating a function of the model parameters that might be meaningful. Obviously, the choice of criterion function needs to reflect the particular purpose. Here, we are concerned with the estimation of parameters and particular subsets of them. To be more precise, our main interest lies in three distinct but interrelated purposes, namely the estimation of

(i) the link function parameter λ ;

(ii) the vector of linear predictor parameters $\beta_{\sim}$;

(iii) both λ and $\beta_{\sim}$.

Due to the features of the first two problems the criterion to be adopted in (i) and (ii) is that of D_s-optimality, whereas D-optimality is suitable for the full problem (iii). For convenience, we split the formulation of the criteria and related derivative functions into these three cases. With matrices M, M_{11}, M_{12} and M_{22} defined in (3.2). The criterion for each case of interest follows

(i) Estimating λ

$$Maximize_{\xi\in\Xi}\Psi_1(M) = \log\{\frac{\det(M)}{\det(M_{22})}\} \tag{4.1}$$

or

$$Maximize_{\xi\in\Xi}\Psi_1(M) = \log\det(M_{11} - M_{12}M_{22}^{-1}M_{21})$$

(ii) Estimating $\beta_{\sim}$

$$Maximize_{\xi\in\Xi}\Psi_p(M) = \log\{\frac{\det(M)}{\det(M_{11})}\} \tag{4.2}$$

or

$$Maximize_{\xi\in\Xi}\Psi_p(M) = \log\det(M_{22} - M_{21}M_{11}^{-1}M_{12})$$

(iii) Estimating λ and $\beta_{\sim}$

$$Maximize_{\xi\in\Xi}\Psi(M) = \log\det(M) \tag{4.3}$$

The inclusion of a pair of weights to represent the relative importance of estimating λ and $\beta_{\sim}$ is an alternative formulation for (iii). In such a case the criterion function would be defined as a weighted conbination of criteria (i) and (ii). The weights could either be specified by the experimenter, based on some subjective criterion, or could be determined through the optimization procedure, by finding the maximum weighted combination, over a feasible set of weights. Criteria (i) and (ii) would then be particular cases of this generalized criterion when either purpose is allocated weight one.

As (4.1), (4.2) and (4.3) are concave functions on the set of design measures, a property required to apply optimal design theory, a theorem similar to the General Equivalence Theorem of Kiefer and Wolfovitz (1960) can be proven.

Optimum designs for all these criteria depend on the parameters λ and $\beta_{\sim}$. If λ were known, say $\lambda = 1$, the problem would reduce to designing for estimation of $\beta_{\sim}$. The resulting design would still require knowledge of $\beta_{\sim}$. The reverse problem of known $\beta_{\sim}$ with λ unknown does not make sense in practice, unless λ and $\beta_{\sim}$ are orthogonal. Example 2.1 shows that this is not necessarily the case. We are thus left with the dependence of the optimal designs on the parameter values. In order to check the optimality of any proposed design we require the derivative function of the design criterion.

Suppose that for fixed λ and $\beta_{\sim}$, $M(\xi^*)$ and $M(\xi_{x_{\sim}})$ denote the normalized information matrices at the optimal design ξ^* and at the design $\xi_{x_{\sim}}$, respectively, where $\xi_{x_{\sim}}$ is the measure assigning mass one to the point $x_{\sim} \in \mathcal{X}$.

It is standard in optimal design theory that the Fréchet derivative of the criterion function at $M(\xi^*)$ in the direction of $M(\xi_{x_{\sim}})$, often called the derivative function, provides a useful bound which is the key for checking the optimality of a given design. For the cases we regard here the bounds are

(i) Estimating λ - For all $x_{\sim} \in \mathcal{X}$,

$$\psi_1(x_{\sim}) = tr[M(\xi_{x_{\sim}})\{M(\xi^*)\}^{-1}] - tr[M_{22}(\xi_{x_{\sim}})\{M_{22}(\xi^*)\}^{-1}] \quad \leq 1 \tag{4.4}$$

(ii) Estimating $\beta_{\sim}$ - For all $x_{\sim} \in \mathcal{X}$,

$$\psi_p(x_{\sim}) = tr[M(\xi_{x_{\sim}})\{M(\xi^*)\}^{-1}] - tr[M_{11}(\xi_{x_{\sim}})\{M_{11}(\xi^*)\}^{-1}] \quad \leq p \tag{4.5}$$

(iii) Estimating λ and $\beta_{\sim}$ - For all $x_{\sim} \in \mathcal{X}$,

$$\psi(x_{\sim}) = tr[M(\xi_{x_{\sim}})\{M(\xi^*)\}^{-1}] \quad \leq p + 1 \tag{4.6}$$

In addition to providing a useful bound, the derivative function has the important feature that the bound is achieved at all the support points of the optimal design. Hence, it provides a straightforward checking procedure, which can be applied when searching for optimal designs.

The following are examples of locally optimal designs, that is designs which are calculated for a single assumed value of λ and $\beta_{\sim}$. In Example 4.1, the interest lies in the estimation of the link fun parameter λ, whereas in Example 4.2 the aim is to determine optimal designs for all three different purposes, using the same information on the parameters. In both examples we assume that the linear predictor structure is of the kind $\eta = \beta_0 + \beta_1 x$. Hence, the dimension of the problem is three.

Example 4.1: $\lambda = 1.4$, $\beta_0 = 0.5$ and $\beta_1 = 1.0$. The aim is to estimate λ. Therefore, criterion function (4.1) is applied, yielding the following optimal design.

$$\psi_1(\boldsymbol{x}) \leq 1, \forall \boldsymbol{x} \in \mathcal{X}; \Psi_1(\xi^*) = -5.766; \xi^* = \left\{ \begin{array}{ccc} -2.384 & 0.9266 & 3.335 \\ 0.6954 & 0.2691 & 0.0355 \end{array} \right\}.$$

To make sure that the above design is indeed optimal, Figure 4.1 shows the derivative function (4.4) in which the upper limit, $\psi_1(\boldsymbol{x}) = 1$ is achieved only at the support points for the optimal design.

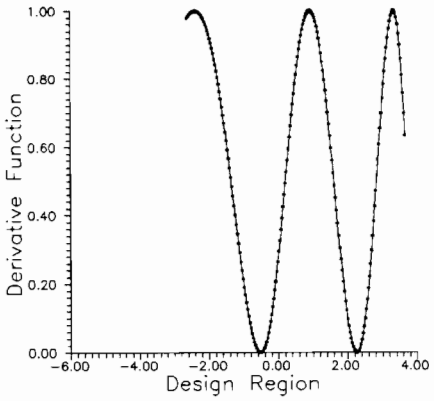


Figure 4.1 — Derivative function (4.4). **Example 4.1**

Example 4.2 : All criteria are used. The parameter values are $\lambda = 0.001, \beta_0 = 0.5$ and $\beta_1 = 1.0$. Table 4.1 shows the resulting optimal designs and respective optimal values of the criteria.

Table 4.1 — Optimal designs for three different criteria

Purpose	Upper limit	Value of the criterion at the optimal	Optimal design
Estimate λ	$\psi_1(\boldsymbol{x}) \leq 1$	-2.526	$\left\{ \begin{array}{ccc} -3.6330 & -0.1416 & 1.3760 \\ 0.2681 & 0.2276 & 0.5043 \end{array} \right\}$
Estimate $\beta_\sim$	$\psi_2(\boldsymbol{x}) \leq 2$	-31.62	$\left\{ \begin{array}{ccc} -2.8410 & -0.4758 & 1.2960 \\ 0.4997 & 0.3477 & 0.1526 \end{array} \right\}$
Estimate both	$\psi(\boldsymbol{x}) \leq 3$	-32.77	$\left\{ \begin{array}{ccc} -2.6450 & -0.1452 & 1.1170 \\ 0.3333 & 0.3333 & 0.3333 \end{array} \right\}$

In spite of using the same exact prior information about the parameters λ, β_0 and β_1, we obtain significantly distinct locally optimal designs. For instance, more than 50% of the weight is allocated to the only positive support point 1.376 in the optimal design to estimate λ, whereas to estimate β_0 and β_1, nearly 50% of the weight is allocated to the most negative support point of the design, -2.841. The unbalanced allocation of weights in both designs suggests that the greatest part of the information about λ is concentrated in positive values of the covariate $\boldsymbol{x}$, as opposed to negative values of $\boldsymbol{x}$, which appear to contain most part

of the information about β_0 and β_1. To reinforce this interpretation, the optimal design to estimate all the parameters allocates equal weights to the support points as though it were a combination between estimating λ and $\beta_\sim$, neither being emphasized.

Another feature of the designs in Table 4.1 is that they all are supported on three points, regardless the number of parameters they are meant to estimate. This can be explained by the fact that the number of support points in the optimal design depends strongly on the dimension of the problem, or the number of unknown parameters, rather than on the number of parameters to be estimated in the criterion adopted. Figures 4.2.a, 4.2.b and 4.2.c below show the derivative functions corresponding to the designs of Table 4.1. Note the upper limit is equal to 1, 2 and 3, respectively the number of parameters to be estimated in each criterion.

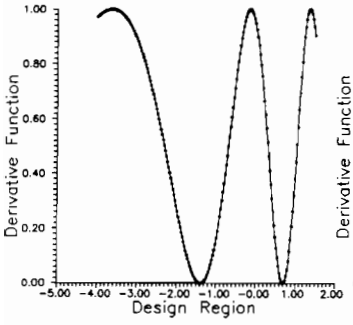


Fig. 4.2.a
$\Psi_1(x)$

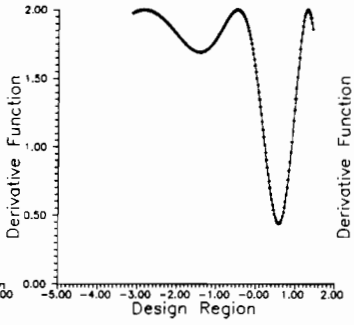


Fig.4.2.b
$\Psi_2(x)$

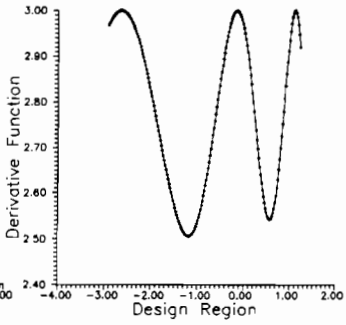


Fig 4.2.c
$\Psi(x)$

In the next section we introduce prior distributions for the parameters λ and $\beta_\sim$. The criterion and derivative functions for the same purposes (i), (ii) and (iii) are presented, together with examples.

5 Bayesian optimal designs to estimate λ and/or $\beta_\sim$

Because of the dependence of the information matrix on the unknown parameter values, the results obtained so far yield locally optimal designs for the parameters λ and/or $\beta_\sim$. However, by incorporating prior distributions into the model, Bayesian designs are also obtainable. These Bayesian designs are optimal experiments that result from the process of averaging the criterion function over the prior distributions of the parameters.

It is straightforward to prove that criterion functions defined as expectations of criteria (4.1), (4.2) and (4.3) are still concave functions on the set of design measures. Consequently all results from optimal design theory continue to hold. For more details about Bayesian designs, see Ponce de Leon & Atkinson (1991) and (1992a).

As suggested by the examples in Section 2 it is reasonable to take priors for the parameters $\{\beta_i\}$ conditional on the value of λ. This assumption, however is purely intuitive, with further investigation of this matter being advisable. Furthermore, the assumption of a conditional probability distribution, $\beta_\sim \mid \Lambda = \lambda$, is general in the sense that if values of $\{\beta_i\}$ are independent of λ all results obtained below will still hold. Taking the priors into consideration the criteria are defined as the expected values of criteria (4.1) to (4.3), expectations being taken in two steps, first over $\beta_\sim \mid \Lambda = \lambda$ and then over λ. Analogously, the derivative functions

are defined as the expected values of expressions (4.4) to (4.6). We present two examples of Bayesian optimal designs.

Example 5.1 : Suppose the interest in the experiment lies in estimating λ and that prior information about its value is available. We consider two cases. In the first, information about λ is relatively accurate, whereas in the second it is rather dispersed. Moreover, we suppose, in the first case, that the distribution of $\{\beta_i\}|\Lambda = \lambda$ is slightly inaccurate whilst independent of λ but, in the second, it is precise although dependent on λ. Prior distributions are shown below in Table 5.1. In both cases we assume that the linear predictor structure is simply $\beta_0 + \beta_1 \, x$.

Table 5.1 — Prior distributions for λ and $\{\beta_i\} \mid \Lambda = \lambda$ (Two cases)

Case	Lambda	Prob (λ)	$\{\beta_0, \beta_1\} \mid \lambda$	Prob $[\{\beta_0, \beta_1\} \mid \lambda]$
First	0.001	0.5	{0.5, 1.0}	0.5
			{0.5, 1.5}	0.5
	0.002	0.5	{0.5, 1.0}	0.5
			{0.5, 1.5}	0.5
Second	0.001	0.5	{0.5, 1.0}	1.0
	1.0	0.5	{0.0, 2.0}	1.0

The resulting Bayesian optimal design for the first case has only three support points as opposed to four in the second. This difference is not unexpected as the number of support points in the Bayesian optimal design is likely to increase as the accuracy of the prior information about λ decreases. Both the designs and the derivative functions are shown in Table 5.2 and Figures 5.1.a and 5.1.*b*, respectively.

Table 5.2 — Optimal Bayesian designs for Example 5.1

Case	Optimum value of criterion	Optimal design
First	-2.784	$\begin{Bmatrix} -2.9850 & -0.1544 & 1.0100 \\ 0.2781 & 0.2595 & 0.4624 \end{Bmatrix}$
Second	-4.153	$\begin{Bmatrix} -3.3440 & -0.5939 & 1.1270 & 2.1370 \\ 0.2418 & 0.4247 & 0.2902 & 0.0433 \end{Bmatrix}$

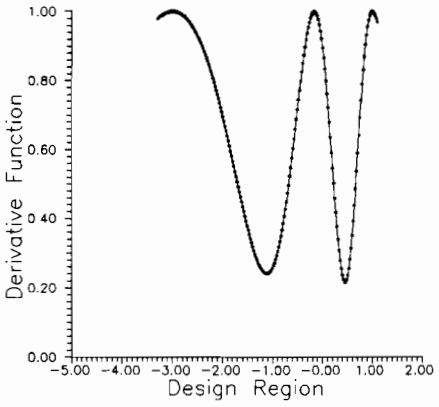


Figure 5.1.a – Derivative function (first case)

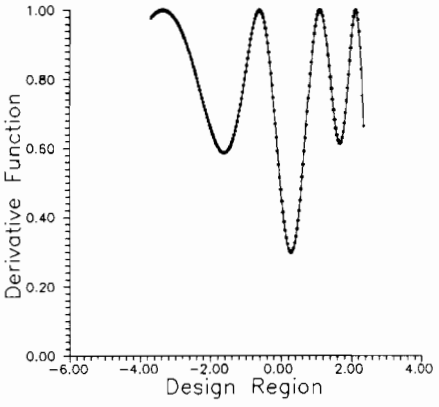


Figure 5.1.b – Derivative function (second case)

Example 5.2 : A rather more interesting situation arises when all three criteria are considered using the same prior information on λ and $\{\beta_i\} \mid \lambda$. Here, we consider prior distributions that are quite concentrated around specific values for both λ and $\{\beta_i\} \mid \Lambda = \lambda$. Table 5.3 shows the priors. Again, the linear predictor assumed was $\eta = \beta_0 + \beta_1 x$. The Bayesian optimal designs are displayed in Table 5.4.

Table 5.3 — Prior distributions for λ and $\{\beta_0, \beta_1\} \mid \Lambda = \lambda$

Lambda	Prob(λ)	$\{\beta_0, \beta_1\} \mid \lambda$	Prob $[\{\beta_0, \beta_1\} \mid \lambda]$
0.001	0.25	{0.5, 1.0}	0.5
		{0.5, 1.5}	0.5
0.002	0.25	{0.25, 1.0}	0.3
		{0.5, 1.5}	0.7
0.003	0.25	{0.4, 2.0}	0.2
		{0.5, 1.5}	0.8
0.004	0.25	{0.8, 0.9}	0.4
		{0.7, 0.95}	0.6

Table 5.4 — Optimal Bayesian designs for Example 5.2

Purpose	Upper limit	Optimum value of criterion function	Optimal design
Estimate λ	$\psi_1(x) \leq 1$	-2.849	$\begin{Bmatrix} -2.8920 & -0.1813 & 0.9594 \\ 0.2802 & 0.2613 & 0.4585 \end{Bmatrix}$
Estimate $\beta_{\sim}$	$\psi_2(x) \leq 2$	-29.14	$\begin{Bmatrix} -2.301 & -0.4248 & 0.9284 \\ 0.499 & 0.3496 & 0.1514 \end{Bmatrix}$
Estimate both	$\psi(x) \leq 3$	-30.34	$\begin{Bmatrix} -2.1440 & -0.1685 & 0.7978 \\ 0.3333 & 0.3333 & 0.3333 \end{Bmatrix}$

All Bayesian optimal designs have three support points. Similar results to those of Example 4.2 are obtained here. Again, the greater part of the information about λ is located at the most positive value of $x(45.85\%)$, whereas the most negative value of x contains almost 50% of the information about $\{\beta_i\}$. The optimal design for estimating λ and $\{\beta_i\}$ has equal weights, suggesting that there was a trade-off between the two purposes. To check optimality, we plot the derivative functions on the design region $\mathcal{X}$. The plots are shown in Figures 5.2.a, 5.2.*b* and 5.2.c.

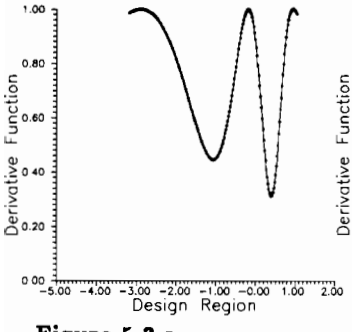


Figure 5.2.a
$\psi_1(x)$

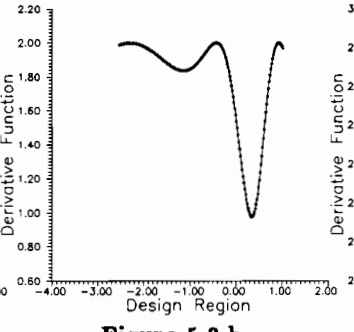


Figure 5.2.b
$\psi_2(x)$

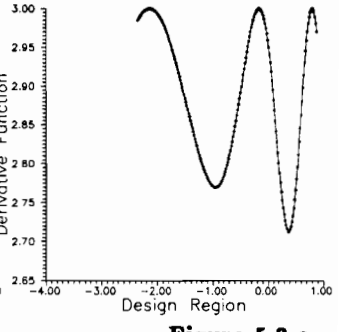


Figure 5.2.c
$\psi(x)$

A Fortran program was used to obtain all numerical results. NAG routine E04JAF was adopted as the optimization procedure. The computation was carried out on a PC 486 machine.

6 Conclusion

Locally and Bayesian optimal designs can be obtained for each of the three problems considered in this article. Optimal designs to estimate λ, the link function parameter, are

particularly useful, for other models than the logistic, probit and complementary log-log may also yield good fits. Bayesian designs depend on how reliable the priors assumed are, so a careful choice of priors is advisable in practice. Combining sequential and Bayesian methods to find optimal designs seems to be the best approach for this kind of problem. Comparison with designs to discriminate between models can be made for the case of designing to estimate λ. Results concerning bounds for the number of support points in the optimal design remain to be found, although it is intuitive that one of the factors determining the number is the accuracy of the priors. The lack of information about the number of support points complicates the search for the optimum. Simulation methods might be used to evaluate how informative the locally and Bayesian optimal designs are compared to other designs.

References

1. Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. Biometrika, 68, 357-63.

2. Chaloner, K. & Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. J. Statist. Plan. Inf. 21, 191-208.

3. Finney, D.J. (1947). *Probit Analysis.* Cambridge.

4. Kiefer, J. & Wolfowitz, J. (1960). The equivalence of two extremum problems. Can. J. Math. 12, 363-6.

5. McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models, 2nd Edition.* London & New York: Chapman and Hall.

6. Milicer, H. & Szczotka, F. (1966). Age at menarche in Warsaw schoolgirls in 1965. Human Biology, 38, 199-203.

7. Ponce de Leon, A.C. & Atkinson, A.C. (1991). Optimum experimental design for discriminating between two rival models in the presence of prior information. Biometrika 78, 3, pp. 601-8.

8. Ponce de Leon, A.C. & Atkinson, A.C. (1992a). Optimal designs for discriminating between two rival binary data models in the presence of prior information. Proceedings of Probastat' 91, Bratislava, Czechoslovakia.

9. Ponce de Leon, A.C. & Atkinson, A.C. (1992b). The design of experiments to discriminate between two rival generalized linear models. Lecture Notes in Statistics, Springer-Verlag, Vol. 78, 159-164.

10. Ponce de Leon, A.C. (1992). Optimum experimental design for model discrimination and generalized linear models. PhD thesis. The London School of Economics and Political Sciences, University of London.

11. Silvey, S.D. (1980). *Optimal Design. An Introduction to the Theory for Parameter Estimation.* London: Chapman and Hall.

On the Construction of Optimal Designs with Applications to Binary Response and to Weighted Regression Models

B. Torsney and A.K. Musrati

1 Introduction

There are a variety of problems in the statistical arena which demand the calculation of one or several optimising probability distributions or measures and hence are examples of the general problems we consider. These include optimal weighted regression design problems and binary response problems. Our interest in this contribution is to show that, for a certain class of two parameter generalised linear and weighted linear regression models, the problem can be reduced to a canonical form. This simplifies the underlying problem and designs are constructed for a number of contexts with a single variable using geometric and other arguments.

2 Weighted linear regression

The ingredients of a weighted linear regression design problem are:

(1) Model:

$$E(y) = \alpha + \beta z,\ z \in Z = [a, b]$$
$$\mathrm{Var}(y) = \sigma^2 / w(z),$$

for some weight function $w(z)$ (see below).

(2) Design:

Design points $z_1, z_2, \ldots, z_r$, $z_i \in Z$ with weights $p_1, p_2, \ldots, p_r$ where the variables p_i can take any value between and including 0 and 1. i.e.

$$\sum_{i=1}^{r} p_i = 1,\ 0 \leq p_i \leq 1.$$

(3) Matrix:

The information matrix M is of the form

$$M = M(p) = \sum_{i=1}^{r} p_i w(z_i) \begin{pmatrix} 1 \\ z_i \end{pmatrix} (1\ z_i),$$

$$= E_p\{w(z)\begin{pmatrix}1\\ z\end{pmatrix}(1\ z)\},$$

$$= E_p\{\underline{g}\underline{g}^t\}, \text{ where } \underline{g} = \begin{pmatrix}g_1\\ g_2\end{pmatrix} \text{ and } g_1 = \sqrt{w(z)}, g_2 = zg_1.$$

(4) Criteria:

Assume $\mathrm{Cov}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} \propto M^{-1}$. There are various criteria. We consider D-optimality which chooses to maximise $\det(M)$.

(5) Weight Functions:

We consider three weight functions from the literature (Fedorov (1972), Karlin and Studden (1966)). These and the corresponding widest possible design space Z_w are now listed.

weight function $w_i(z)$	**design space** Z
(i) $w_1(z) = (1-z)^{\alpha+1}(1+z)^{\beta+1}$	$Z \subseteq Z_w = (-1,1), \alpha, \beta > -1$
(ii) $w_2(z) = z^{\alpha+1}e^{-z}$	$Z \subseteq Z_w = (0,\infty), \alpha > -1$
(iii) $w_3(z) = e^{-z^2}$	$Z \subseteq Z_w = (-\infty,\infty)$

3 Binary regression

The ingredients of a binary regression design problem are:

(1) Model:

$$y/x \approx \text{ Bi } (1,\eta)$$

where

$$\begin{aligned}\eta &= F(\gamma+\delta x), a^* \le x \le b^* \\ &= F(\underline{\theta}^t\ \underline{s}, \text{with } \underline{s} = (1,x)^t, \underline{\theta} = (\gamma,\delta)^t.\end{aligned} \tag{3.1}$$

(2) Design: Let p_x denote a design on $\mathcal{X} = [a^*, b^*]$.

(3) Matrix: The Fisher information matrix is given by

$$M(p_x,\underline{\theta}) = \int_{\mathcal{X}} w(\gamma+\delta x)\begin{pmatrix}1\\ x\end{pmatrix}(1\ x)p(dx), \text{where } w(.) = \frac{f^2}{F[1-F]}.$$

(4) Criteria: Local D-optimality: choose p_x for fixed $\underline{\theta}$ to maximise $\det(M)$.

(5) Canonical Transformation:

Our objective is to find optimal designs for all $\underline{\theta}$. We can simplify this task by transforming to a canonical problem.

$$\begin{aligned} \text{Let } z &= \gamma + \delta x = \underline{\theta}^t \underline{x}. \\ \text{Let } \underline{t} &= \begin{pmatrix} 1 \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} \end{aligned} \tag{3.2}$$

Then (i) $x \in \mathcal{X} = [a^*, b^*] \Leftrightarrow z \in Z = [a, b]$ for some a, b.

$$\begin{aligned} \text{(ii)} \quad M &= \int_Z w(z) \begin{pmatrix} 1 \\ z \end{pmatrix} (1 \; z) p(dz) \\ &= E_p\{\underline{g}\underline{g}^t\}, \text{ where } \underline{g} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \text{ and } g_1 = \sqrt{w(z)}, g_2 = z g_1. \end{aligned}$$

(iii) a design p_x on $\mathcal{X}$ induces a design p on Z.

(iv) $M(p, \underline{\theta}) = B M_x B^t = B E\{w(\underline{\theta}^t \underline{s}) \underline{s}\underline{s}^t\} B^t$

(v) $\det\,(M(p, \theta)) = \det(M_x) \det(B^2)$.

Thus choosing p_x to maximise $\det(M(p, \underline{\theta}))$ is equivalent to choosing p to maximise $\det(M)$ i.e. it is equivalent to a weighted linear regression design problem with weight function $w_f(z) = \frac{f^2(z)}{F(z)[1-F(z)]}$. We consider nine choices of $f(z)$ (see Table 1). In these cases $w_f(z)$ and $w_i(z)(i = 1, 2, 3)$ share a key feature which has strong implications for D-optimal designs. This is that the set G definded as

$$G = G(Z) = \{\underline{g}(z) = (g_1, g_2)^t : g_1 = \{w(z)\}^{1/2}, g_2 = z g_1, z \in Z\},$$

is a closed convex curve beginning and ending at the origin as z ranges from its lower limit to the upper limit. For example see figures 1 to 6.

4 Determining optimal designs

An ideal approach has two steps;

(i) First identify or characterise the support points of an optimal design, that is z values with positive weight;

(ii) Then determine these optimal weights

In the two parameter case there exists

(a) A D-optimal design with a support of two or three points (well established);

(b) Explicit solutions for the optimal weights in either case.

5 Determining support points

Geometric characterisation:

There is a geometrical rule concerning the set G above, which (potentially) identifies the support points of a D-optimal design. Namely these are the points of contact between G and the smallest ellipsoid centred on the origin containing G.

The geometry of G will clearly be crucial. We have noted that in all our examples it is a closed convex curve anchored at the origin.

6 Explicit D-optimal weights

(1) Two design points:

Suppose that a design p assigns weights p_1, p_2 to two points z_1, z_2 such that $\underline{g}(z_1), \underline{g}(z_2) \in \mathbf{R}^2$ are linearly independent. Then the information matrix of this design M, is given by

$$M = \{p_1 \underline{g}(z_1)\underline{g}(z_1)^t + p_2 \underline{g}(z_2)\underline{g}(z_2)^t\}.$$

The determinant of M is equal to

$$\det(M) = p_1(1 - p_1) = w(z_1)w(z_2)(z_1 - z_2)^2,$$

since $p_1 + p_2 = 1$, which implies that $(p_2 = 1 - p_1)$.

The determinant above is proportional to the simple function $f(p_1) = p_1(1 - p_1)$ of p_1. Thus an elementary one variable optimisation technique shows that the determinant is maximised at $\hat{p}_1 = \hat{p}_2 = 1/2$ (which verifies standard result).

(2) Three design points:

Suppose that a design p assigns weights p_1, p_2, p_3 to three points z_1, z_2, z_3 such that any two of $\underline{g}(z_1), \underline{g}(z_2), g(z_3) \in \mathbf{R}^2$ are linearly independent of each other. Then the information matrix of this design is given by

$$M = \{p_1 \underline{g}(z_1)\underline{g}(z_1)^t + p_2 \underline{g}(z_2)\underline{g}(z_2)^t + p_3 \underline{g}(z_3)\underline{g}(z_3)^t\}.$$

$$M = \sum_{i=1}^{3} p_i \underline{g_i}\underline{g_i^t}.$$

Let $G_{ij} = (\underline{g_i} : \underline{g_j})$, so that G_{ij} is a 2×2 matrix, and denote its determinant by $D_{ij} = \det(G_{ij})$. Then $\det(M)$ is given by

$$\varphi = \det(M) = p_1 p_2 D_{12}^2 + p_1 p_3 D_{13}^2 + p_2 p_3 D_{23}^2,$$

where

$$D_{12}^2 = w(z_1)w(z_2)(z_1 - z_2)^2,$$

$$D_{13}^2 = w(z_1)w(z_3)(z_1 - z_3)^2,$$

$$D_{23}^2 = w(z_2)w(z_3)(z_2 - z_3)^2.$$

To find the optimal weights we must maximise the determinant above with resspect to the variables p_1, p_2, p_3 subject to $\sum p_j = 1$. First order conditions are

$$\therefore \frac{\partial \varphi}{\partial p_j} = \sum p_j \frac{\partial \varphi}{\partial p_i} = 2\varphi,$$

which yield three linear equations in p_1, p_2, p_3. Solving this linear system by the well established elimination method gives the oiptimal value of p_i as

$$\hat{p}_i = \frac{D_i}{\sum_{j=1}^{3} D_j}.$$

where

$$D_1 = D_{23}^2(D_{12}^2 + D_{13}^2 - D_{23}^2),$$

$$D_2 = D_{13}^2(D_{12}^2 + D_{23}^2 - D_{13}^2),$$

$$D_3 = D_{12}^2(D_{13}^2 + D_{23}^2 - D_{12}^2).$$

7 Results for all models (except DEXP & DREC.)

(a) Widest possible choice

The D-optimal designs on Z_w has only two support points, say z_1 and $z_2(z_1 < z_2)$ for seven choices of the distribution function F in Table 1 (excluding the double exponential and the double reciprocal ones which will be considered in detail), and for the three weight functions $w_i(z)(i = 1, 2, 3)$. See Tables 1 and 2.

(b) General $Z = [a, b]$.

Consider the problem of finding D-optimal designs for general $Z = [a, b]$. For the seven choices of the distribution function F in Table (1) excluding the double exponential and the double reciprocal distributions, and for the three weight functions $w_i(z)(i = 1, 2, 3)$. The D-optimal designs (appear) to have two support points which are categorised by a common form of solution. In fact the conclusions of Ford, Torsney, and Wu (1992) extend to $w_i(z)(i = 1, 2, 3,)$. Denote the support points of the best two point design on the widest possible choice of Z, i.e. Z_w, by a^* and b^* and on Z by z_1, z_2, where $a^* \leq b^*, z_1 < z_2$.

(1) Case $b \geq b^*$ and $a \leq a^*$:

If the two-point design on a^* and b^* is D-optimal for Z_w then it is D-optimal for $Z = [a, b]$. Otherwise, it is only guaranteed to be D-optimal among two-point designs. We conclude that

this design is globally D-optimal for seven of the nine choices of F in Table (1), and the three weight functions $w_i(z)(i = 1, 2, 3)$.

(2) Case $Z = [-b, b], b \leq b^*$:

For a symmetric distribution function F, if the function $w(z)z$ is non decreasing over $Z = [0, b]$, then $z_1 = -b$ and $z_2 = b$. The first two symmetric distributions in Table (1), and in addition the symmetric weight functions $w_1(z)[\text{for}\alpha = \beta], w_3(z)$ satisfy this condition on $w(z)$.

(3) Case $b \leq b^*$:

If the function $w(z)(z - \ell_1)^2$ is non-decreasing in z over $Z = [\ell_1, b]$ for any $\ell_1 \geq a$, then $z_2 = b$ and $z_1 = z_b(a)$, where $z_b(a) = \max\{a, \ell(b)\}, \ell(b)$ being the value which maximises $w(z)(z - b)^2$ over $Z = [a, b]$. For any F such that $w(z)$ is log concave and differentiable over $Z = [a^*, b^*], w(z)(z - \ell_1)^2$ is non-decreasing over $Z = [\ell_1, b^*]$ for any $\ell_1 \geq a^*$. Wu (1988) shows such log concavity is respect of the logistic and skewed logistic distributions and the complimentary log-log distribution. The property is also enjoyed by the three weight functions $w_i(z)(i = 1, 2, 3)$.

(4) Case $a \geq a^*$:

If the function $w(z)(z - u_2)^2$ is non-increasing in z over $Z = [a, u_2]$ for any $u_2 \leq b$, then $z_1 = a$ and $z_2 = z_a(b)$, where $z_a(b) = \min\{b, u(a)\}, u(a)$ being the value which maximises $w(z)(z-a)^2$ over $Z = [a, b]$. It can be shown that for the examples cited in case 3, $w(z)(z-u_2)^2$ is non-increasing over $Z = [a^*, u_2]$ for any $u_2 \leq b^+$.

(5) Case $a \geq a^*, b \leq b^*$:

If $w(z)$ is log-concave and differentiable over $Z = [a^*, b^*]$, then $z_1 = a$ and $z_2 = b$. This follows from combining the results in cases (3) and (4) above.

We summarise the above statements in Table (3) The values z_1, z_2 are only guaranteed to be D-optimal among two-point designs on the appropriate Z.

8 Results for (DEXP & DREC) models

(a) Widest possible choice.

The D-optimal designs $Z_w = (-\infty, \infty)$ for both these symmetric models prove to have three support points, say $\{-z^*, 0, z^*\}$ with optimal weights $\hat{p}, 1 - 2\hat{p}, \hat{p})$ where z^* and $\hat{p}$ must maximise the determinant of the information matrix under the symmetric design $\{-z, 0, z\}$ with weights $p, 1 - 2p, p)$. In fact the optimal weight $\hat{p}$ can be determined explicitly for given z^* from the following equation

$$\hat{p} = \frac{1}{4[1 - w(z^*)]},$$

and z^* is the optimal value of z which maximises the resultant determinant under the above design. i.e.

$$\det(M) = \frac{z^2 w(z)}{4[1 - w(z)]}.$$

We list the support points and the optimal weights for both models in Table (4). Also see figures 7 to 10.

(b) General $Z = [a, b] \not\ni 0$.

Consider the case when the interval $Z = [a, b]$ does not contain zero. In this case we believe that the D-Optimal designs have two support points at least one of which is an end point Z(a if $a > 0, b$ if $b < 0$). That is

$$\text{Supp}(p^*) = \left\{ \begin{array}{ll} \{a, \min\}, & a > 0 \\ \{\max, b\}, & b < 0 \end{array} \right\},$$

where $\min = \min\{b, b^*(a)\}, b^*(a)$ being the value which maximises $w(z)(z-a)^2$ over $Z = [a, \infty)$ for any $a > 0$, and $\max = \max\{a, a^*(b)\}, a^*(b)$ being the value which maximises $w(z)(z-b)^2$ over $Z = (-\infty, b]$ for any $b < 0$.

In particular, if $Z = [0, \infty)$ then the D-optimal design is a two-point design supported on $\{0, u_1\}$ with optimal equal weights $(\frac{1}{2}, \frac{1}{2})$ where u_1 must maximise the resultant determinant $z^2 w(z)$ over $Z = [0, \infty)$. By symmetry $\{-u_1, 0\}$ are the support points of the D-optimal design on $Z = (-\infty, 0]$.

(c) General $Z = [a, b] \in 0$.

We now consider the case when the interval $Z = [a, b]$ contains zero, i.e. $a < 0$ and $b > 0$. In this case the D-optimal designs are supported on either two points or three-points. These designs are categorised by a general form of solution. Define the following terms

(i) Let $-u_1$ denote the negative support of the global D-optimal design on $Z = (-\infty, 0]$. Thus $-u_1 = -1.841$ for the double exponential distribution and $-u_1 = -1.618$ for the double reciprocal distribution.

(ii) Let $-u_2$ denote the negative support point of the global D-optimal design on the widest possible choice of Z, i.e. $Z_w = (-\infty, \infty)$. Thus $-u_2 = -1.5936$ for the double exponential distribution and $-u_2 = -\sqrt{2}$ for the double reciprocal distribution.

(iii) Let $-u_3$ be the smallest value of a^* such that the D-optimal design on the set $\{a^*, 0\}$ is optimal on $Z = [a^*, 0]$ but it is not optimal on $Z = [a^*, y]$ for any positiv y. (We note that $-u_3$ is obtained by solving the equation $F'(0) = 0$) and the function $F = \underline{g}^t M^{-1} \underline{g}$ denotes the variance function of the estimated responce surface. Thus $-u_3 = -1$ for the double exponentisl distribution and $-u_3 = -0.5$ for the double reciprocal distribution.

(iv) Let $-u_4$ denote the critical value of $-k$ at which the D-optimal design on $Z_k = [-k, k] \forall k$ changes from a 3-point to a 2-point design. Thus $-u_4 = -0.4055$ for the double exponential distribution and $-u_4 = -0.1974$ for the double reciprocal distribution.

(v) Consider the D-optimal design on $Z = (-\infty, 0]$ with support points $\{-u_1, 0\}$ (see(i)). Let $z^\sim(u_1)$ be the smallest positive value of z such that $F(z) = \underline{g}^t M^{-1} \underline{g} = 2$. Thus $z^\sim(u_1) = 0.3528$ for the double exponential distribution and $z^\sim(u_1) = 0.5062$ for the double reciprocal distribution.

(1) Case $a < -u_2$ and $b > u_2$:

The D-optimal design models in this case is that for Z_w. So it is a threepoint design supported on Supp $(p^*) = \{-u_2, 0, u_2\}$. We now assume $b < u_2$ and $b \leq |a|$, then

(2) Case $a < -u_1$:

Here the support points of the D-optimal design are classified as follows

$$\text{Supp}(p^*) = \left\{ \begin{array}{ll} \{-u_1, 0\}, & b < z^{\sim}(u_1) \\ \{a^*(b), 0, b\}, & z^{\sim}(u_1) < b < u_2 \end{array} \right\}$$

where $a^*(b)$ is the value of a^* which maximises the determinant of M, where M is the design matrix under the design $\{a^*, 0, b\}$. We note that always $a^*(b) > u_1$ as empirical results suggest.

(3) Case $-u_1 < a < -u_2$:

The support points of the D-optimal design in this case are either two points or three points classified as follows

$$\text{Supp}(p^*) = \left\{ \begin{array}{ll} \{a, 0\}, & b < z^{\sim}(a) \\ \{a, 0, b\}, & z^{\sim}(a) < b < z^{\sim}(u_1) \\ \{\max, 0, b\}, & z^{\sim}(u_1) < b < u_2 \end{array} \right\},$$

where

(i) max = $\max\{a, a^*(b)\}$, $a^*(b)$ being the value of a^* which maximises the determinant of M, where M is as in case 2.

(ii) and $z^{\sim}(a)$ is the value of z such that $F(z) = \underline{g}^t M^{-1} \underline{g} = 2$ under the design on $\{a, 0\}$.

(4) Case $-u_2 < a < -u_3$:

The support points of the D-optimal design in this case are classified as follows

$$\text{Supp}(p^*) = \left\{ \begin{array}{ll} \{a, 0\}, & b < z^{\sim}(a) \\ \{a, 0, b\}, & z^{\sim}(a) < b < |a| \end{array} \right\},$$

where $z^{\sim}(a)$ is as in case (3) above.

(5) Case $-u_3 < a < -u_4$:

The support points of the D-optimal designs in this case are classified as follows

$$\text{Supp}(p^*) = \left\{ \begin{array}{ll} \{a, b\}, & b < z^{+}(a) \\ \{a, 0, b\}, & z^{+}(a) < b < |a| \end{array} \right\},$$

where $z^{\sim}(a)$ is the (unique) value of b such that $F(0) = \underline{g}^t M^{-1} \underline{g} = 2$ under the D-optimal design on the set $\{a, b\}$.

(6) Case $a > -u_4$: Finally, the D-optimal designs in this case are supported on two points. Namely $\text{Supp}(p^*) = \{a, b\}, b < |a|$.

Table 1; Supports z_1, z_2 of two-point D-optimal designs on $Z_w = (-\infty, \infty)$. (Note $s = \text{sign}(z)$).

Name	$f_i(z)$	$F_i(z)$	z_1	z_2
1) Logit	$e^{-z}(1+e^{-z})^{-2}$	$(1+e^{-z})^{-1}$	-1.543	1.543
2) Probit	$\frac{1}{\sqrt{2\pi}}e^{(-z^2/2)}$	$\Phi(z)$	-1.138	1.138
3) Double Exponential	$\frac{1}{2}e^{-\lvert z\rvert}$	$\frac{1+2}{2} - \frac{s}{2}e^{-\lvert z\rvert}$	-0.768	0.768
4) Double Reciprocal	$\frac{1}{2}(1+\lvert z\rvert)^{-2}$	$\frac{(1+s)}{2} - \frac{s}{2}(1+\lvert z\rvert)^{-1}$	-0.390	0.390
5) Complementary Log-Log	$\exp(z-e^2)$	$1-\exp(-e^2)$	-1.338	0.980
6-9) Skewed Logit	$m[F_1(z)]^{m-1}f_1(z)$	$(1+e^{-z})^{-m}$	—	—
6) $m = 1/3$	...	...	-4.409	0.552
7) $m = 2/3$	...	...	-2.284	1.191
8) $m = 3/2$	...	...	-0.939	1.898
9) $m = 3$	...	...	-0.060	2.525

Table 2; Supports z_1, z_2 of two-point D-optimal designs on Z_w for the three weight functions $w_i(z)(i = 1, 2, 3)$.

Weight function	**Support points**
1) $w_1(z) = (1-z)^{\alpha+1}(1+z)^{\beta+1}$	$z_i = \frac{(\beta-\alpha)(\alpha+\beta+3)\pm 2\sqrt{(\alpha+2)(\beta+2)(\alpha+\beta+3)}}{(\alpha+\beta+3)(\alpha+\beta+4)}$
2) $w_2(z) = z^{\alpha+1}e^{-z}$	$z_i = (\alpha+2) \pm \sqrt{(\alpha+2)}, i = 1, 2.$
3) $w_3(z) = e^{-z^2}$	$z_1 = \frac{-1}{\sqrt{2}}, z_2 = \frac{1}{\sqrt{2}}$

Table 3: Supports of two-point D-optimal designs on a general $Z = [a, b]$.

$Z = [a, b]$	z_1	z_2
1) $a \leq a^*, b \geq b^*$	a^*	b^*
2) $a = -b, b \leq b^*$	$-b$	b
3) $a > -\infty, b \leq b^*$	$z_b(a) = \max\{a, \ell(b)\}$	b
4) $a \geq a^*, b < \infty$	a	$z_a(b) = \min\{b, u(a)\}$
5) $a \geq a^*, b \leq b^*$	a	b

Table 4: Support points and optimal weights of D-optimal designs on $Z_w = (-\infty, \infty)$.

Model	**Support points**	**Optimal Weights**
1) Double Exponential.	(-1.5936,0,1.5936)	(0.2819,0.4362,0.2819)
2) Double Reciprocal.	$(-\sqrt{2}, 0, \sqrt{2})$	(0.2617,0.4766,0.2617)

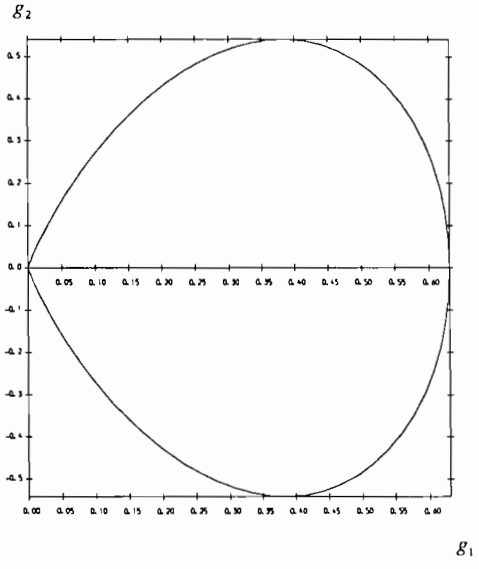


Figure (1)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in Z_w = (-\infty, \infty)\}$, for the symmetric logistic distribution.

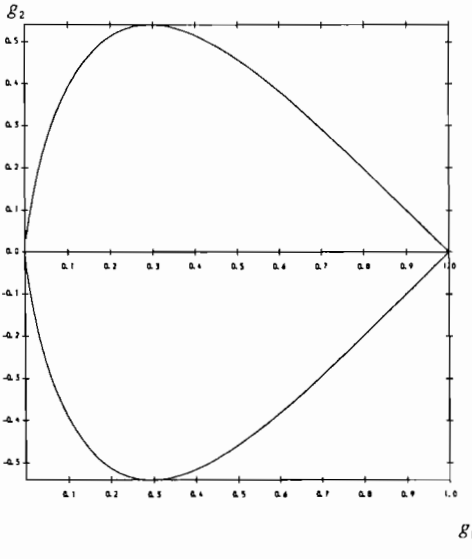


Figure (2)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in Z_w = (-\infty, \infty)\}$, for the symmetric double exponential distribution.

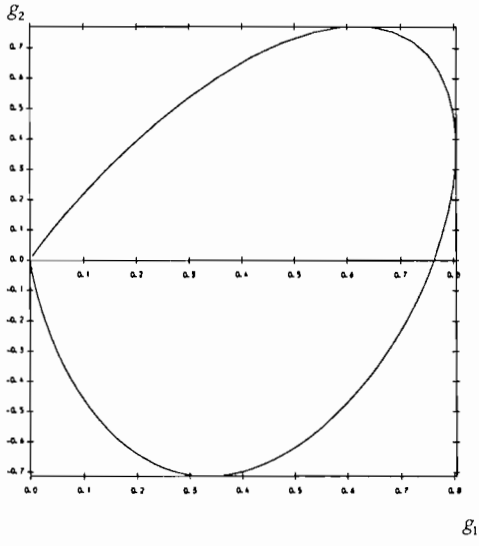


Figure (3)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in Z_w = (-\infty, \infty)\}$, for the asymmetric complementary log-log distribution.

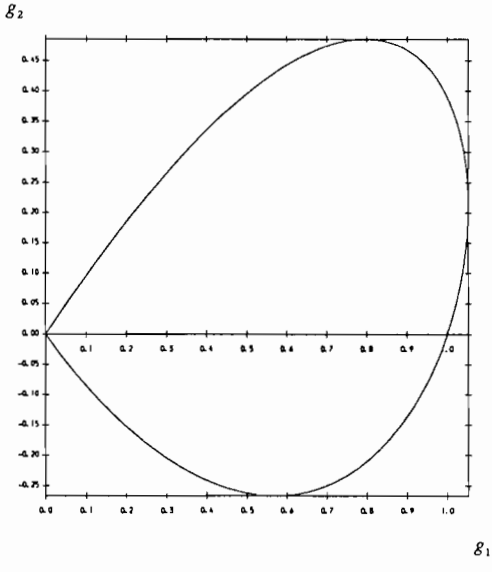


Figure (4)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \subseteq Z_w = (-1, 1)\}$, for the asymmetric weight function $w_1(z) = (1-z)^{\alpha+1}(1+z)^{\beta+1}$ with $(\alpha = 1, \beta = 2)$.

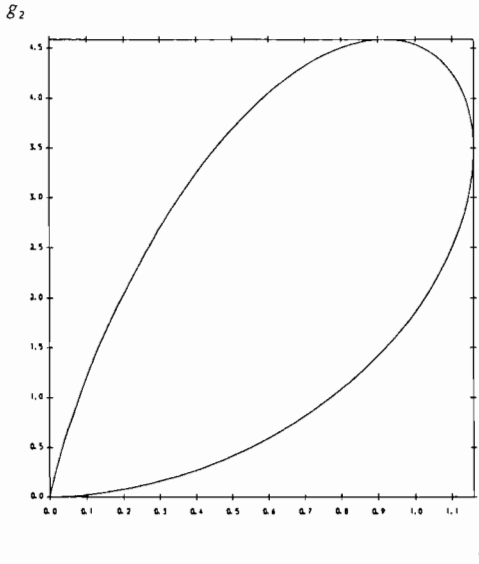


Figure (5)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \subseteq Z_w = (0, \infty)\}$, for the asymmetric weight function $w_2(z) = z^{\alpha+1}e^{-1}$ with $(\alpha = 2)$.

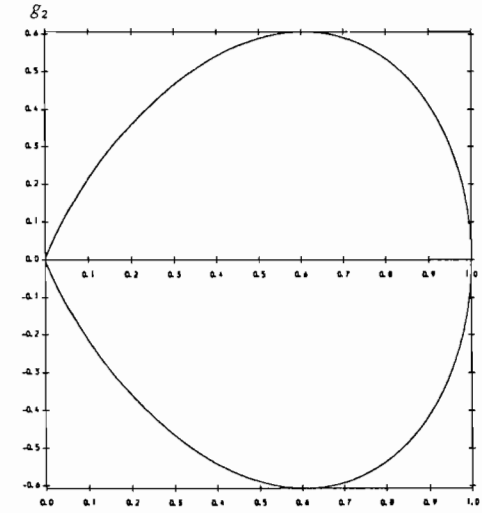


Figure (6)

Plot of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in Z = (-\infty, \infty)\}$, for the symmetric weight function $w_3(z) = e^{-z^2}$.

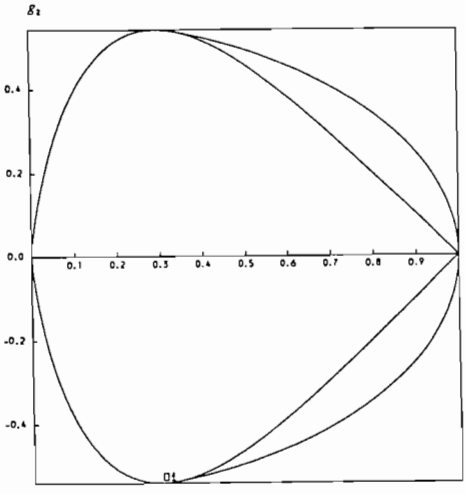


Figure (7)

Combined plots of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in \mathbf{R}\}$ and the ellipsoid $Q = \{(g_1, g_2)^t : (g_1, g_2)^t M^{*-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = 2\}$ for the double exponential distribution and the case $a < -u_2$ and $b > u_2$, where M^* is the global D-optimal design matrix on $Z_w = (-\infty, \infty)$, whose support points are $\{-u_2, 0, u_2\}(u_2 = 1.5936)$ and weights $\{0.2819, 0.4362, 0.2819\}$.

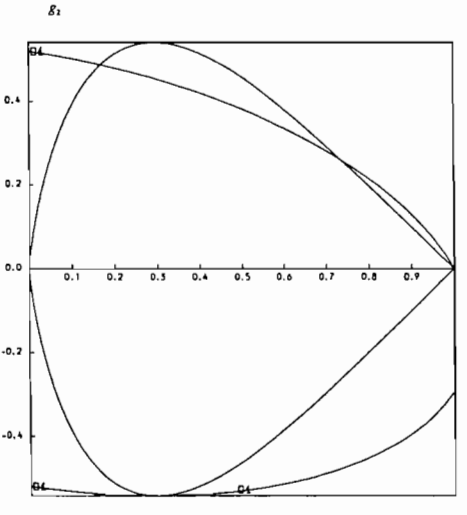


Figure (8)

Combined plots of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in \mathbf{R}\}$ and the ellipsoid $Q = \{(g_1, g_2)^t : (g_1, g_2)^t M^{*-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = 2\}$ for the double exponential distribution and the case $a < -u_1$, where M^* is the global D-optimal design matrix on $Z = (-\infty, 0]$, whose support points are $\{-u_1, 0\}, b < z^{\sim}(u_1)(u_1 = 1.841)$.

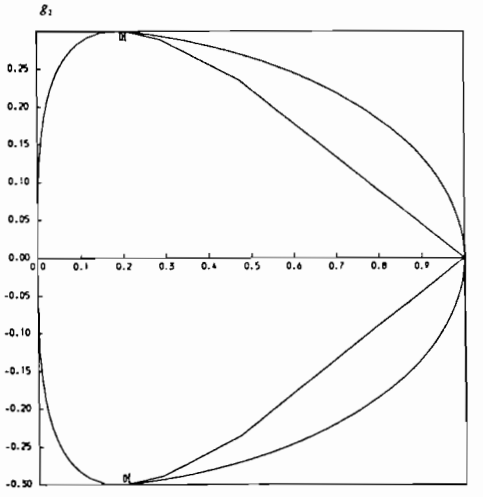


Figure (9)

Combined plots of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in \mathbf{R}\}$ and the ellipsoid $Q = \{(g_1, g_2)^t : (g_1, g_2)^t M^{*-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = 2\}$ for the double exponential distribution and the case $a < -u_2$ and $b > u_2$, where M^* is the global D-optimal design matrix on $Z_w = (-\infty, \infty)$, whose support points are $\{-u_2, 0, u_2\}(u_2 = \sqrt{2})$ and weights $\{0.2617, 0.4766, 0.2617\}$.

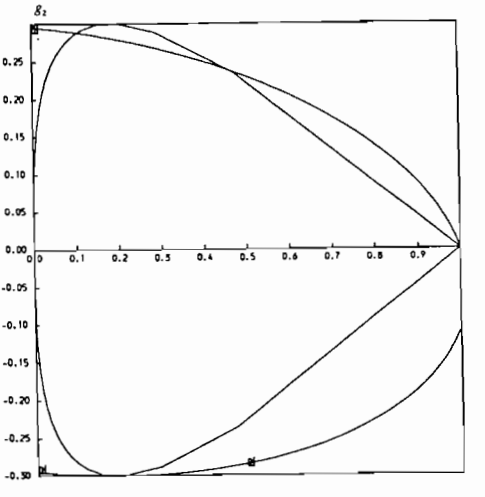


Figure (10)

Combined plots of the set $G = \{(g_1, g_2)^t : g_1 = \sqrt{w(z)}, g_2 = zg_1, z \in \mathbf{R}\}$ and the ellipsoid $Q = \{(g_1, g_2)^t : (g_1, g_2)^t M^{*-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = 2\}$ for the double exponential distribution and the case $a < -u_1$, where M^* is the global D-optimal design matrix on $Z = (-\infty, 0]$, whose support points are $\{-u_1, 0\}, b < z^{\sim}(u_1)(u_1 = 1.618)$.

References

1. Fedorov, V.V. (1972). Theory of optimal experiments. Academic Press:New York and London

2. Ford, I. (1976). Optimal static and sequential design: A critical review. Unpublished Ph.D.thesis. University of Glasgow.

3. Ford, I., Torsney, B. and Wu, C.F.J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. J.Roy.Soc. Vol.54, No.2,pp.569-583.

4. Karlin, S and Studden, W.J. (1966). Optimal experimental designs. Ann. Math. Stat., Vol.37, 783-815.

5. Silvey, S.D. (1980) Optimal design. London: Chapman and Hall.

6. Torsney, B (1981). Algorithms for a constrained optimisation problem with applications in statistics and optimum design. Unpublished Ph.D. thesis. University of Glasgow.

7. Torsney, B. (1983) "A moment inequality and monotonicity of an algorithm", proceedings of the international symposium on semi-infinite programming and applications at the University of Texas at Austin (K.O. Kortanek and A.V. Fiacco, eds.), lecture notes in economics and mathematical systems 215, 249-260.

8. Torsney, B. (1988). "Computing optimising distributions with applications in design, estimation and image processing". In optimal design and analyses of experiments (Dodge, Y.,Fedorov, V.V & Wynn, H.P. (Eds.)). Elsevier science publishers B.V. (North Holland), pp.316-370.

9. Torsney, B and Alahmadi, Abdulhadi, M (1992). Further developments of Algorithms for constructing optimising distributions. In model oriented data analysis. A survey of recent methods. Proceedings of the 2nd IIASA workshop in StKyrik, Bulgaria, May 28-June 1, 1990, pp121-129.

10. Wu, C.F.J.(1988) "Optimal design for percentile estimation of a quantal response curve". In optimum design and analysis of experiments (Dodge, Y., Fedorov, V.V. Wynn, H.P. (Eds.)). Elsevier science publishers B.V. (North Holland), pp. 213-223.

Behaviour of Asymptotically Optimal Designs for Robust Estimation at Finite Sample Sizes

Christine Müller

Estimating θ or a linear aspect of θ of a conditionally contaminated linear model $y(x) = f(x)^T\theta + \epsilon(x)$, where the conditional distributions given $x \in X$ of the errors $\epsilon(x)$ may be different contaminated normal distributions, an asymptotic bias will appear. Bounding the bias by some bound b estimators can be characterized which minimize the trace of the asymptotic covariance matrix under all estimators with bias bounded by b. In Müller (1987) it was shown that A-optimal design measures are also optimal design measures for these optimal robust estimators. But these results hold only asymptotically. Here by a Monte-Carlo study for linear and quadratic regression it is shown that the behaviour of optimal robust estimators at A-optimal and non A-optimal designs with finite sample size does not much differ from the asymptotic behaviour.

1 Introduction

A general linear model

$$y_{iN} = f(x_{iN})'\theta + \epsilon_{iN}, \qquad i = 1, ..., N$$

is considered where y_{iN} are the observations, $x_{iN} \in X$ are experimental conditions, $f : X \to \mathbb{R}^m$ is a known "regression" function, $\theta \in \mathbb{R}^m$ is an unknown parameter vector, ϵ_{iN} are error variables. To enable asymptotic considerations the corresponding design measures ξ_N of the designs $x_N := (x_{1N}, ..., x_{NN})$, $N \in \mathbb{N}$, should satisfy

$$\xi_N := \sum_{i=1}^{N} e_{x_{iN}} \overset{N\to\infty}{\longrightarrow} \xi \text{ weakly } (e_x \text{ denoting the Dirac measure on } x).$$

In classical linear models it is assumed that the error variables $\epsilon_{1N}, ..., \epsilon_{NN}$ are independent and identically distributed and often it is assumed that they are normally distributed with mean 0 and known variance σ^2. I.e. without loss of generality it is assumed

$$\epsilon_{iN} \sim n_{(0,1)}$$

($n_{(\mu,\sigma^2)}$ denoting the normal distribution with mean μ and variance σ^2).

But if some outlying observations (gross errors) may appear the normal distribution is not correct. Then a conditionally contaminated linear model is adequate (see Bickel (1981, 1984), Rieder (1985, 1987), Müller (1992)). In such a model it is assumed that the $\epsilon_{1N}, ..., \epsilon_{NN}$ are independent and distributed according to a contaminated normal distribution where the contamination may be different for different experimental conditions. I.e.

$$\epsilon_{iN} \sim Q_{iN}(dz) := (1 - N^{-1/2}\gamma(x_{iN}))\, n_{(0,1)}(dz) + N^{-1/2}\gamma(x_{iN})\, g(z, x_{iN})\, n_{(0,1)}(dz)$$

with $\int \gamma(x)\,\xi(dx) \le 1$, $\int g(z,x) n_{(0,1)}(dz) = 1$, $g(z,x) \ge 0$ for all $z \in I\!R$, $x \in X$. Thereby the markov kernel $g(\cdot,x)n_{(0,1)}$ models the form and $\gamma(x)$ the proportion of contamination. The set $\mathcal{U}$ of all sequences $(Q^N := \bigotimes_{i=1}^N Q_{iN})_{N\in I\!N}$ defines a conditionally contamination neighbourhood around the classical model $(n_{(0,1)}^N)_{N\in I\!N}$.

For estimating a linear aspect $\varphi(\theta) = L\theta$, $L \in I\!R^{l\times m}$, of the unknown parameter vector θ a big class of estimators are the asymptotically linear (AL-) estimators. This class of estimators includes all estimators $(\widehat{\varphi}_N)_{N\in I\!N}$ for φ which satisfy

$$\sqrt{N}[\widehat{\varphi}_N - \varphi(\theta) - \frac{1}{N}\sum_{i=1}^N \psi(y_{iN} - f(x_{iN})'\theta, x_{iN})] \overset{N\to\infty}{\longrightarrow} 0$$

in probability for the classical model $(n_{(0,1)}^N)_{N\in I\!N}$ where ψ is called the influence function of the estimator and should satisfy

$$\psi \in \Psi(\xi) := \{\psi : I\!R \times X \to I\!R^l; \int |\psi(z,x)|^2\, n_{(0,1)}(dz)\xi(dx) < \infty,$$
$$\int \psi(z,\cdot) n_{(0,1)}(dz) = 0, \int \psi(z,x) f(x)'z\, n_{(0,1)}(dz)\xi(dx) = L\}.$$

Under some regularity conditions M-estimators $\widehat{\theta}_N$ for θ which are defined as solutions of

$$\sum_{i=1}^N \psi(y_{iN} - f(x_{iN})'\widehat{\theta}_N, x_{iN}) = 0$$

are AL-estimators with influence function ψ. Also one-step-M-estimators $\widehat{\varphi}_N$ (briefly called OM estimators) for $\varphi(\theta) = L\theta$ which are given by

$$\widehat{\varphi}_N = L\widehat{\theta}_N + \frac{1}{N}\sum_{i=1}^N \psi(y_{iN} - f(x_{iN})'\widehat{\theta}_N, x_{iN})$$

where $\widehat{\theta}_N$ is some initial estimator for θ are AL-estimators with influence function ψ. See Bickel (1975), Müller (1992). In particular all Gauss-Markov-estimators for $\varphi(\theta) = L\theta$ (briefly called GM-estimators) are AL-estimators with influence function $\psi(z,x) = L\,M(\xi)^- f(x)\,z$ where $M(\xi) := \int f(x)f(x)'\xi(dx)$.

At conditionally contaminated linear models AL-estimators $\widehat{\varphi}_N$ are asymptotically normal distributed. I.e. for all $(Q^N)_{N\in I\!N} \in \mathcal{U}$ the distribution of $\sqrt{N}(\widehat{\varphi}_N - \varphi(\theta))$ converges for $N \to \infty$ to a normal distribution with mean (asymptotic bias)

$$b(\psi, (Q^N)_{N\in I\!N}) := \int \psi(z,x)\,\gamma(x)\,g(z,x)\,n_{(0,1)}(dz)\,\xi(dx)$$

and covariance matrix

$$V(\psi) := \int \psi(z,x)\psi(z,x)'\, n_{(0,1)}(dz)\,\xi(dx).$$

Thereby the maximum asymptotic bias satisfies

$$\sup\{b(\psi,(Q^N)_{N\in I\!N});\ (Q^N)_{N\in I\!N} \in \mathcal{U}\} = \|\psi\|_\infty \qquad (\|\psi\|_\infty = \operatorname{ess\,sup}_{n_{(0,1)}\otimes\xi} |\psi|).$$

See Bickel (1984), Rieder (1985, 1987), Müller (1987, 1992), Kurotschka and Müller (1992).

For robust AL-estimators the maximum asymptotic bias should be bounded, i.e. $\|\psi\|_\infty \le b < \infty$ and asymptotically optimal robust AL-estimators with bias bound b are those AL-estimators which have an influence function $\psi_{b,\xi}$ satisfying

$$\psi_{b,\xi} = \arg\min\{\operatorname{tr} \int \psi\psi' d(n_{(0,1)}\otimes\xi);\ \psi \in \Psi(\xi) \text{ with } \|\psi\|_\infty \le b\}.$$

Characterizations of $\psi_{b,\xi}$ are given in Hampel (1978), Krasker (1980), Rieder (1985) for $\varphi(\theta) = \theta$ and in Müller (1987), Kurotschka and Müller (1992) for arbitrary $\varphi(\theta) = L\theta$. A special case appears for $b = \infty$ where the Gauss-Markov-estimator is asymptotically optimal, i.e. $\psi_{\infty,\xi}(z,x) = L\,M(\xi)^{-} f(x) z$.

Asymptotically optimal design measures for estimation with bias bound b are those design measures ξ_b satisfying (see Müller (1987, 1991))

$$\xi_b \;=\; \arg\min\{\mathrm{tr} \int \psi_{b,\xi}\, \psi'_{b,\xi}\, d(n_{(0,1)} \otimes \xi);\; \xi \in \Xi\}.$$

A special case of this optimality criterium for design measures ξ is the classical A-optimality criterium because for $b = \infty$

$$\int \psi_{\infty,\xi}\, \psi'_{\infty,\xi}\, d(n_{(0,1)} \otimes \xi) \;=\; L\,M(\xi)^{-}\, L'.$$

But the following theorem shows that A-optimal design measures are also optimal for robust estimation with bias bound $b < \infty$.

Theorem 1.1 *(Müller (1987, 1991))*
Let $\Xi = \{\xi;\ \varphi(\theta)$ *is identifiable at* $\xi\}$, $f : X \to \mathbb{R}^m$ *is continuous,* X *compact and* $b \geq \min\{\| \psi \|_\infty;\ \psi \in \bigcup_{\xi \in \Xi} \Psi(\xi)\}$. *Then*

ξ^* *is A-optimal in* Ξ

iff

ξ^* *is asymptotically optimal in* Ξ *for robust estimation with bias bound* b.

Asymptotically A-optimal designs are also optimal for robust estimation. But what happens for finite sample sizes? Therefore a Monte-Carlo study as described in Section 2 of this paper was done for linear and quadratic regression. The results of the study are given in Section 3 of this paper.

2 Description of the Monte-Carlo study

For estimating $\theta = (\theta_0, \theta_1)'$ of a linear regression model

$$y_{iN} = \theta_0 + \theta_1 x_{iN} + \epsilon_{iN}$$

the behaviour of optimal AL-estimators for θ with 3 different bias bounds was explored at 3 different designs with sample size $N = 10, 20, \ldots, 2000$ and this was done for 3 different contaminated error distributions and 3 different true parameter vectors. The designs were

$$x_N^A := \begin{pmatrix} -1 & 1 \\ \frac{N}{2} & \frac{N}{2} \end{pmatrix} \overset{N\to\infty}{\longrightarrow} \xi_A = \frac{1}{2}(e_{-1} + e_1)$$

where ξ_A is the A-optimal design measure, i.e. the asymptotically optimal design measure for estimation with some bias bound,

$$x_N^2 := \begin{pmatrix} -1 & 1 \\ \frac{N}{4} & \frac{3N}{4} \end{pmatrix} \overset{N\to\infty}{\longrightarrow} \xi_2 = \frac{1}{4} e_{-1} + \frac{3}{4} e_1$$

and

$$x_N^3 := \begin{pmatrix} -1 & 1 \\ \frac{N}{8} & \frac{7N}{8} \end{pmatrix} \overset{N\to\infty}{\longrightarrow} \xi_3 = \frac{1}{8}e_{-1} + \frac{7}{8}e_1.$$

The following AL-estimators for θ were regarded: The least squares estimator (LSE) $\widehat{\theta}_N^*$ and one-step-M-estimators (OME)

$$\widehat{\theta}_N = \widehat{\theta}_N^* + \frac{1}{N}\sum_{i=1}^N \psi_{b,\xi}(y_{iN} - f(x_{iN})'\widehat{\theta}_N^*, x_{iN})$$

with $b = 4$ and $b = 8$ where the initial estimator $\widehat{\theta}_N^*$ was the LS estimator and the score functions $\psi_{b,\xi}$ were given by

$$\psi_{4,\xi_A}(z,x) = \operatorname{sgn}(z)\frac{\min\{|z|, 4\cdot 0.704\}}{\sqrt{2}\cdot 0.704}\begin{cases}(-1,1)' & \text{for } x=-1,\\ (1,1)' & \text{for } x=1,\end{cases}$$

$$\psi_{8,\xi_A}(z,x) = \operatorname{sgn}(z)\frac{\min\{|z|, 8\cdot 0.707\}}{\sqrt{2}\cdot 0.707}\begin{cases}(-1,1)' & \text{for } x=-1,\\ (1,1)' & \text{for } x=1,\end{cases}$$

$$\psi_{4,\xi_2}(z,x) = \operatorname{sgn}(z)\begin{cases}\frac{\min\{|z|,4\cdot 0.218\}}{\sqrt{2}\cdot 0.218}(-1,1)' & \text{for } x=-1,\\ \frac{\min\{|z|,4\cdot 1.061\}}{\sqrt{2}\cdot 1.061}(1,1)' & \text{for } x=1,\end{cases}$$

$$\psi_{8,\xi_3}(z,x) = \operatorname{sgn}(z)\begin{cases}\frac{\min\{|z|,8\cdot 0.109\}}{\sqrt{2}\cdot 0.109}(-1,1)' & \text{for } x=-1,\\ \frac{\min\{|z|,8\cdot 1.237\}}{\sqrt{2}\cdot 1.237}(1,1)' & \text{for } x=1.\end{cases}$$

These OM estimators were used because they are easy to compute. While the LS estimator is asymptotically optimal for estimation with bias bound $b = \infty$, i.e. without bias bound, the OM estimators are asymptotically optimal for robust estimation with bias bound $b = 4$ and $b = 8$, respectively, (see for the special form of $\psi_{b,\xi}$ Müller (1987) or Kurotschka and Müller (1992)). The regarded contaminated distributions of ϵ_{iN} were

$$\epsilon_{iN} \sim (1 - N^{-\frac{1}{2}})n_{(0,1)} + N^{-\frac{1}{2}}n_{(\mu(x_{iN}),1)}$$

where $\mu = (\mu(-1), \mu(1)) = (5,5)$, $\mu = (\mu(-1), \mu(1)) = (5,10)$, $\mu = (\mu(-1), \mu(1)) = (10,5)$ and the regarded true parameter vectors $\theta = (\theta_0, \theta_1)'$ were $\theta = (1,1)'$, $\theta = (0,5)'$, $\theta = (5,0)'$.

For estimating the linear aspect $\varphi(\theta) = (\theta_1, \theta_2)'$ of a quadratic regression model

$$y_{iN} = \theta_0 + \theta_1 x_{iN} + \theta_2 x_{iN}^2 + \epsilon_{iN}$$

the behaviour of optimal AL-estimators for $\varphi(\theta)$ with 2 different bias bounds was explored at 2 different designs with sample size $N = 10, 20, \ldots, 2000$ and this was done for 2 different contaminated error distributions and 2 different true parameter vectors. The designs were

$$x_N^A := \begin{pmatrix} -1 & 0 & 1 \\ \frac{N}{2+\sqrt{2}} & \frac{\sqrt{2}N}{2+\sqrt{2}} & \frac{N}{2+\sqrt{2}} \end{pmatrix} \overset{N\to\infty}{\longrightarrow} \xi_A = \frac{1}{2+\sqrt{2}}(e_{-1} + \sqrt{2}e_0 + e_1)$$

where ξ_A is the A-optimal design measure, i.e. the asymptotically optimal design measure for estimation with some bias bound, and

$$x_N^2 := \begin{pmatrix} -1 & 0 & 1 \\ \frac{N}{3} & \frac{N}{3} & \frac{N}{3} \end{pmatrix} \overset{N\to\infty}{\longrightarrow} \xi_2 = \frac{1}{3}(e_{-1} + e_0 + e_1).$$

The following AL-estimators for θ were regarded: The Gauss-Markov estimator (GME) $\widehat{\varphi}_N^*$ for $\varphi(\theta)$ and the one-step-M-estimator (OME)

$$\widehat{\varphi}_N = \begin{pmatrix} 0\,1\,0 \\ 0\,0\,1 \end{pmatrix} \widehat{\theta}_N^* + \frac{1}{N} \sum_{i=1}^{N} \psi_{b,\xi}(y_{iN} - f(x_{iN})'\widehat{\theta}_N^*, x_{iN})$$

with $b = 4$ where the initial estimator $\widehat{\theta}_N^*$ was the LS estimator for θ and the score functions $\psi_{b,\xi}$ were given by

$$\psi_{4,\xi_A}(z,x) = \operatorname{sgn}(z) \frac{\min\{|z|, 4 \cdot 0.345\}}{0.345} \begin{cases} \frac{1}{\sqrt{2}}(-1,1)' & \text{for } x = -1, \\ (0,-1)' & \text{for } x = 0, \\ \frac{1}{\sqrt{2}}(1,1)' & \text{for } x = 1, \end{cases}$$

$$\psi_{4,\xi_2}(z,x) = \operatorname{sgn}(z) \begin{cases} \frac{\min\{|z|, 4 \cdot 0.432\}}{\sqrt{2} \cdot 0.432}(-1,1)' & \text{for } x = -1, \\ \frac{\min\{|z|, 4 \cdot 0.154\}}{0.154}(0,-1)' & \text{for } x = 0, \\ \frac{\min\{|z|, 4 \cdot 0.432\}}{\sqrt{2} \cdot 0.432}(1,1)' & \text{for } x = 1. \end{cases}$$

While the GM estimator is asymptotically optimal for estimation with bias bound $b = \infty$, i.e. without bias bound, the OM estimator is asymptotically optimal for robust estimation with bias bound $b = 4$ (see for the special form of $\psi_{b,\xi}$ Müller (1987) or Kurotschka and Müller (1992)). The regarded contaminated distributions of ϵ_{iN} were

$$\epsilon_{iN} \sim (1 - N^{-\frac{1}{2}}) n_{(0,1)} + N^{-\frac{1}{2}} n_{(\mu(x_{iN}),1)}$$

where $\mu = (\mu(-1), \mu(0), \mu(1)) = (5,5,5)$, $\mu = (\mu(-1), \mu(0), \mu(1)) = (0,5,0)$ and the regarded true parameter vectors $\theta = (\theta_0, \theta_1, \theta_2)'$ were $\theta = (1,1,1)'$, $\theta = (0,0,5)'$.

For linear regression as well as for quadratic regression the errors ϵ_{iN}, $i = 1, \ldots, N$, for a sample size N were generated by the random number generator of the programming language GAUSS. For different designs different errors were generated because the contaminated error distributions depend on the experimental conditions. But from the same observations which were calculated from the errors and the true parameter vector the different estimators were calculated. For a given sample size N this was repeated M times where $M = 500$ were used for $N = 10, 20, 50, 100, 200, 500, 1000, 2000$ and $M = 1000$ for $N = 10, 20, 30, 50, 70, 100, 130, 160$. To compare the distribution of the estimator $\widehat{\varphi}_N$ at finite sample sizes N with the asymptotic distribution of $\sqrt{N}(\widehat{\varphi}_N - \varphi(\theta))$ the following values were calculated from the estimated values $\widehat{\varphi}_{Nj}$ in the M repetitions. Thereby $\widehat{\varphi}_{Nj} = (\widehat{\varphi}_{Nj}^1, \ldots, \widehat{\varphi}_{Nj}^l)'$ denotes the estimation for $\varphi(\theta) = (\varphi^1(\theta), \ldots, \varphi^l(\theta))'$ in the j'th repetition, $j = 1, \ldots, M$.

Simulated bias (Bias) of $\sqrt{N}\widehat{\varphi}_N$:

$$\widehat{B}_N := \left| \frac{1}{M} \sum_{j=1}^{M} \sqrt{N}(\widehat{\varphi}_{Nj} - \varphi(\theta)) \right|.$$

Trace of the simulated covariance matrix (Tr Cov) of $\sqrt{N}\widehat{\varphi}_N$:

$$\widehat{V}_N := \sum_{\alpha=1}^{l} \frac{1}{M-1} \sum_{j=1}^{M} N \left(\widehat{\varphi}_{Nj}^{\alpha} - \frac{1}{M} \sum_{j=1}^{M} \widehat{\varphi}_{Nj}^{\alpha} \right)^2.$$

Trace of the simulated mean squared error matrix (MSE) of $\sqrt{N}\widehat{\varphi}_N$:

$$\widehat{MSE}_N := \sum_{\alpha=1}^{l} \frac{1}{M} \sum_{j=1}^{M} N \left(\widehat{\varphi}_{Nj}^{\alpha} - \varphi^{\alpha}(\theta) \right)^2.$$

These values were plotted against the sample size N. All calculations were done with the programming language GAUSS.

3 Results

Figure 1 shows for the linear regression the trace of the simulated covariance matrix $\widehat{V}_N$ of the LS estimator and the OMS estimator with bias bound $b = 4$ for the designs x_N^A and x_N^2 with sample sizes N from 10 up to 2000 (the realized sample sizes are marked at the horizontal axis with thick tick marks). At the right hand side of the figure the asymptotic values are marked which are

2	for the LS estimator at x_N^A,
2.667	for the LS estimator at x_N^2,
2.002	for the OM estimator at x_N^A,
2.940	for the OM estimator at x_N^2.

This shows that the convergence is very slow. In a repetition of the study with other random numbers the same behaviour appeared. The same behaviour appeared also for other contaminated error distributions, namely for contaminated error distributions with $\mu = (\mu(-1), \mu(1)) = (5, 10)$ and $\mu = (\mu(-1), \mu(1)) = (10, 5)$, for other parameter vectors, namely for $\theta = (0, 5)'$ and $\theta = (5, 0)'$ and for the design x_N^3. The same held also for the quadratic regression, see Figure 2. Therefore these results shows that it is very necessary to investigate the behaviour of the regarded estimators at small sample sizes.

In particular Figure 1 shows that for the LS estimator as well as for the OM estimator the trace of the covariance matrix at the A-optimal design x_N^A was smaller than at the non A-optimal design x_N^2. Therefore the asymptotic relation between the designs held also for finite sample sizes, in particular for small sample sizes of $N = 10$ and $N = 20$. But in opposite to the asymptotic behaviour, for x_N^A as well as for x_N^2 for $N \geq 30$ the trace of the covariance matrix of the OME estimator was smaller than of the LS estimator. This is due to the fact that the variances of the error variables ϵ_{iN} increases very much when outliers appears. Because the proportion of outliers decreases with $\sqrt{N}$, in particular for the LS estimator the convergence to the asymptotic variance is slow.

In contrast to the trace of the simulated covariance the convergence of the simulated bias to the asymptotic bias was quick, in particular for the LS estimator for which for the designs x_N^A, x_N^2 and x_N^3 for $\mu = (\mu(-1), \mu(1)) = (5, 5)$ the asymptotic bias is equal to $|\int \psi_{\infty,\xi}\, n_{(5,1)}(dz)\xi(dx)| = 5$. For the OM estimator for $N \geq 30$ the simulated bias was less than the bias bound $b = 4$. See Figure 3. The over-shoot over the bias bound $b = 4$ for $N < 30$ will be a consequence of the special choice of the initial estimator of the OM estimator which provides that for small sample sizes the OM estimator behaves more like the LS estimator which is the initial estimator. The special choice of the initial estimator may also provide the well behaviour of the OM estimator concerning the trace of the covariance matrix.

From the behaviour of the simulated bias and of the trace of the simulated covariance matrix it is clear that concerning the trace of the simulated mean squared error matrix a difference between the designs x_N^A and x_N^2 appeared and that a more important difference between the LS estimator and the OM estimator appeared (see Fig. 4).

Similar results concerning the trace of the covariance matrix, the bias and the trace of the mean squared error matrix were obtained for other random numbers and for other parameter vectors. For other contaminated error distributions the behaviour of the trace of the covariance matrix and the bias and therefore also of the trace of mean squared error matrix changed in the following way: For contaminated error distributions with $\mu = (\mu(-1), \mu(1)) =$

$(5,10)$ and $\mu = (10,5)$ the bias of the LS estimator converged quickly to the asymptotic values which are in both cases equal to $\frac{5\sqrt{10}}{2} \approx 7.906$ while the bias of the OM estimator fell under the bias bound $b = 4$ only for $N \geq 100$ for the A-optimal designs ξ_N^A and for $N > 130$ for the non A-optimal designs ξ_N^2. For $\mu = (5,10)$ the difference of x_N^A and x_N^2 concerning the trace of the covariance matrix vanished. This is due to the fact that the contaminations with mean 5 and 10 at the experimental conditions $x = -1$ and $x = 1$ provide a greater variance at $x = 1$ than at $x = -1$. In such situations x_N^A is not any more A-optimal and instead of x_N^2 a design performs better which puts more observations on $x = 1$. Therefore x_N^2 performed better. The opposite is true for $\mu = (10,5)$. Then the trace of the covariance matrix at x_N^2 is very large so that the differences between x_N^A and x_N^2 are greater than for $\mu = (5,5)$.

To compare the A-optimal design x_N^A with other designs one should take into account that the minimum bias bound which is possible depends on the design. For x_N^A and x_N^2 the minimum bias bound is $\sqrt{\pi} \approx 1.77 < 4$ and $2\sqrt{\pi} \approx 3.54 < 4$, respectively. But for x_N^3 the minimum bias bound is $4\sqrt{\pi} \approx 7.09$. See Müller (1987), Kurotschka and Müller (1992). Therefore in the comparison of x_N^A and x_N^3 a bias bound of $b = 8$ was chosen. For x_N^A this bias bound was so large that for contaminations with $\mu = (5,5)$ the OM estimator behaved like the LS estimator. But at x_N^3 the trace of the covariance matrix of the OM estimator with $N \geq 50$ was significantly smaller than the variance of the LS estimator. Comparing the designs for all regarded N even the trace of the covariance of the OM estimator at x_N^3 was significantly greater than the variance of the OM and LS estimator at x_N^A.

For estimating $\varphi(\theta) = (\theta_1, \theta_2)'$ of the quadratic regression model at the A-optimal design x_N^A and the non A-optimal designs x_N^2 similar results were obtained as for estimating θ in the linar regression model if contamination distributions with $\mu = (\mu(-1), \mu(0), \mu(1)) = (5,5,5)$ were used (see Fig. 2). In particular the trace of the covariance matrix of the OM estimator as well as of the GM estimator was at x_N^A smaller than at x_N^2 and at both designs for $N \geq 30$ the variance of the OM estimator was smaller than the variance of the GM estimator. Only the difference between the designs was not so great as for linear regression which is due to the special choice of x_N^2. Also the asymptotic values do not differ very much for x_N^A and x_N^2 (see Fig. 2). For contamination distributions with $\mu = (0,5,0)$ the difference between the two designs was greater.

Furthermore the bias showed a different behaviour. Namely for the contamination distributions with $\mu = (5,5,5)$ the bias of the GM estimator as well as of the OM estimator was approximately equal to 0 which is due to the special aspect and the equality of the contamination distributions. Using contamination distributions with $\mu = (0,5,0)$ for the GM estimator for all regarded N a bias of approximately 5 appeared where for the OM estimator the bias was less than the bias bound $b = 4$ for $N \geq 20$.

4 Conclusion

Although in this study only two regression models were regarded the results may hold for other linear models, in particular for linear models with qualitative factors because the regarded designs had finite support. Therefore one can make the following conclusions.

In general concerning the trace of the covariance matrix also for small sample sizes an A-optimal design behaves better than a non A-optimal design. In particular this effect always appears for equal contamination distributions at the different experimental conditions. Only

for unfavourably posed contaminations a non A-optimal design may be better than the A-optimal design. But this effect will be obtained for the OM estimator as well as for the LS or GM estimator. If the different proportions and forms of the contamination at the different experimetal conditions are known special other designs should be used. But in general the proportions and forms of the contamination is unknown so that an A-optimal design is the best choice.

It was surprising that in general for $N \geq 30$ the trace of the covariance matrix of the OM estimators was smaller than the variance of the LS or GM estimators although for very small sample sizes and asymptotically the opposite is true. Therefore in a contaminated linear model an OM estimator with a LS estimator as initial estimator is a very good choice.

Because OM estimators with other initial estimators may have a different behaviour in a seperate study the behaviour of OM estimators with other initial estimators should be investigated.

References

[1] BICKEL, P.J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428-434.

[2] BICKEL, P.J. (1981). Quelque aspects de la statistique robuste. In École d'Été de Probabilités de St. Flour. *Springer Lecture Notes in Math.* **876**, 1-72.

[3] BICKEL, P.J. (1984). Robust regression based on infinitesimal neighbourhoods. *Ann. Statist.* **12**, 1349-1368.

[4] HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the ASA Statistical Computing Section*, ASA, Washington, D.C., 59-64.

[5] KRASKER, W.S. (1980). Estimation in linear regression models with disparate data points. *Econometrica* **48**, 1333-1346.

[6] KUROTSCHKA, V. and MÜLLER, Ch.H. (1992). Optimum robust estimation of linear aspects in conditionally contaminated linear models. *To appear in Ann. Statist.*

[7] MÜLLER, Ch.H. (1987). Optimale Versuchspläne für robuste Schätzfunktionen in linearen Modellen. *Ph. D. thesis.* Freie Universität Berlin.

[8] MÜLLER, Ch.H. (1991). Optimal designs for robust estimation in conditionally contaminated linear models. *Submitted to J. Statist. Plann. Inference.*

[9] MÜLLER, Ch.H. (1992). One-step-M-estimators in conditionally contaminated linear models. *Preprint No. A-92-11*, Freie Universität Berlin, Fachbereich Mathematik.

[10] RIEDER, H. (1985). Robust estimation of functionals. *Technical Report.* Universität Bayreuth.

[11] RIEDER, H. (1987). Robust regression estimators and their least favorable contamination curves. *Stat. Decis.* **5**, 307-336.

Fig.1. Linear regression: Variance for estimating $\varphi(\vartheta)=\vartheta$

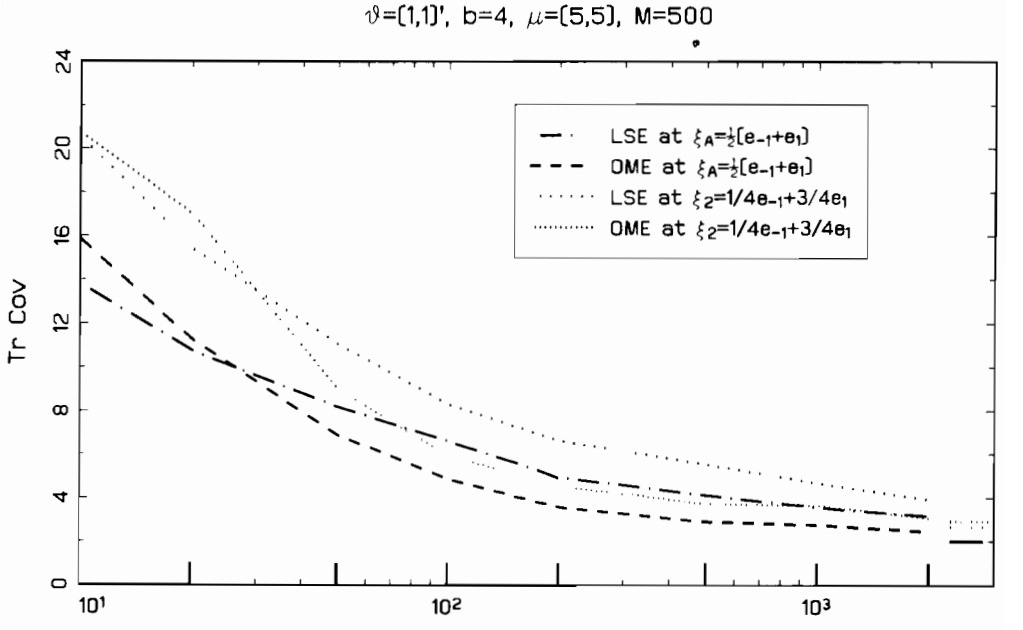


Fig.2. Quadratic regression: Variance for estimating $\varphi(\vartheta)=(\vartheta_1,\vartheta_2)'$

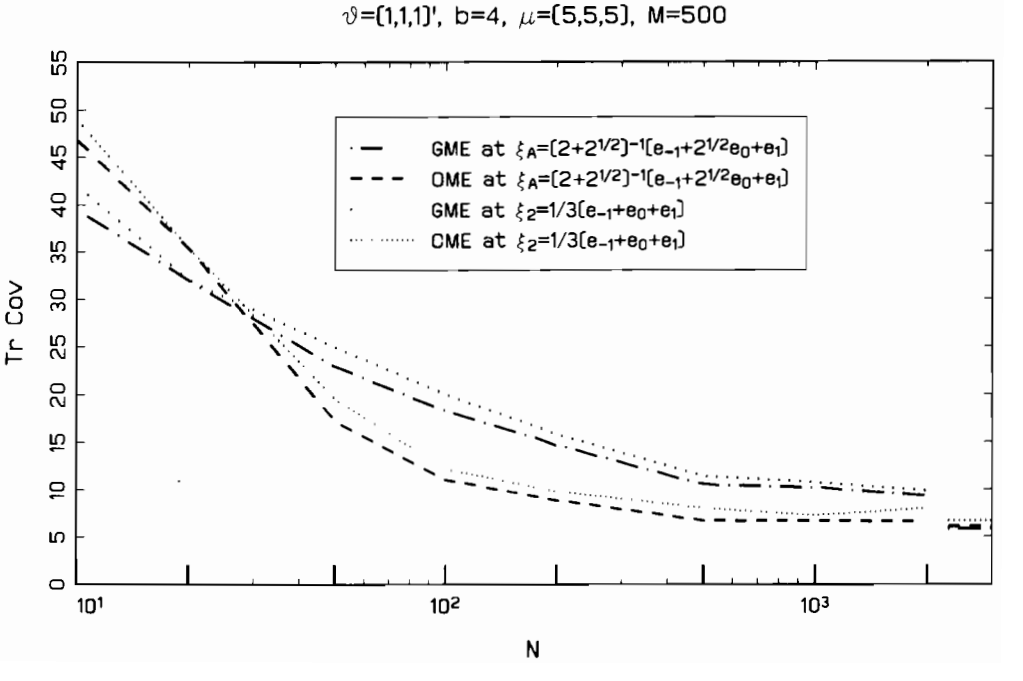


Fig.3. Linear regression: Bias for estimating $\varphi(\vartheta)=\vartheta$

$\vartheta=(1,1)'$, b=4, $\mu=(5,5)$, M=1000

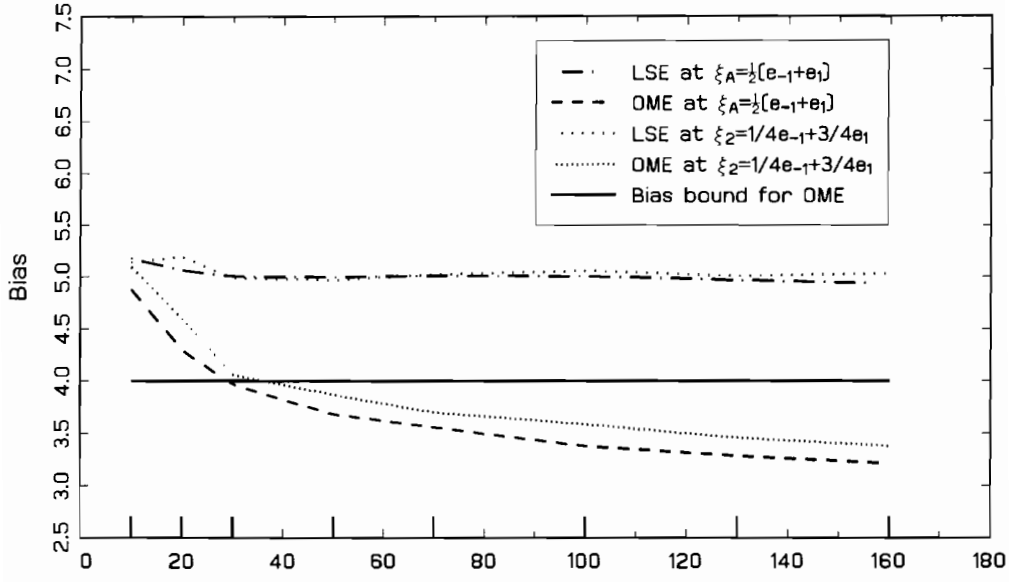
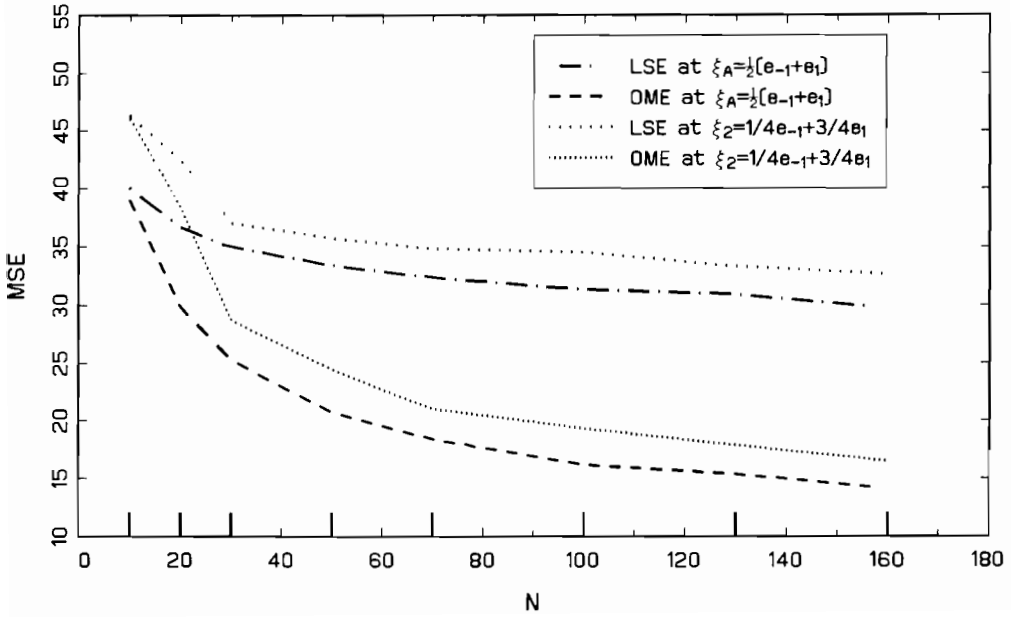


Fig.4. Linear regression: Mean squared error for estimating $\varphi(\vartheta)=\vartheta$

$\vartheta=(1,1)'$, b=4, $\mu=(5,5)$, M=1000



D-optimal Design for Polynomial Abel Inversion

Victor P. Kozlov

The optimum set of observation positions in the problem of restoration of a radial structure of axially symmetric object is determined explicitly under assumption that unknown radial distribution may be described by a polynomial of even degree. The new result which specifies the D-optimal design for polynomial regression with weight function that completes the related result of Karlin and Studden (1966) is proved. The numerical values of D-optimal design points are given for half-degrees of polynomial up to 15.

1 Introduction

We consider one special form of the Abel integral equation which often appears in different applications, such as plasma diagnostics, X-ray or emission tomography. This form of Abel equation coincides with partial case of Radon transform for circularly symmetric functions:

$$F(p) = 2\int_p^R \frac{r, f(r), dr}{\sqrt{r^2 - p^2}} \qquad \text{[Abel]} \tag{1}$$

$$= \int_{-\sqrt{R^2-p^2}}^{\sqrt{R^2-p^2}} f(\sqrt{p^2 + s^2})\, ds \qquad \text{[Radon]}$$

Here function f usually represents a radial distribution of some physical quantity (emission or mass density) over a circle of radius R while $F(p)$ is the value of integral along straight line (the ray) situated at the distance p from the center of circle (the "ray-sum value").

It is well known that the equation (1) may be solved at least formally by explicit inverse transform

$$f(r) = -\frac{1}{\pi r}\frac{d}{dr}\int_r^R \frac{p, F(p)\, dp}{\sqrt{p^2 - r^2}} \tag{2}$$

yet it is equally well known that the solution (2) is unstable due to the need of differentiation , so for stable numerical solution one needs some kind of a priori information. A simple way to introduce such information is to fix some parameter model for unknown function f. Really a kind of polynomial expansion is mostly used [Minerbo and Levy (1969), Kosarev (1973)].

Here we use a slightly different approach to produce the polynomial model. Let radius $R = 1$. Suppose that unknown distribution over unit circle may be represented by the bivariate polynomial of full degree n :

$$f_n(x,y) = \sum_{k+l \le n} a_{kl} x^k y^l \quad , \quad x^2 + y^2 \le 1 \tag{3}$$

Then under condition of rotational invariance

$$f_n(r\cos\theta, r\sin\theta) = f(r) \tag{4}$$

which is necessary to apply the Abel equation we obtain the polynomial which contains only even powers of r and may be written as

$$f_m(r) = \sum_{\nu=0}^{m} (\nu + \frac{1}{2})\, a_\nu R^0_{2\nu}(r) \tag{5}$$

where $m = [\frac{n}{2}]$, $R^0_{2\nu}(r)$ are the radial Zernike polynomials associated with zero-order harmonic, and a_ν 's are the unknown coefficients.

Due to Cormack (1964) it is well known that the Abel or Radon transform of the model (4) is given by

$$F_m(p) = \sqrt{1-p^2} \sum_{\nu=0}^{m} a_\nu U_{2\nu}(p) \tag{6}$$

where $U_{2\nu}(p)$ are the Chebyshev polynomials of the second kind:

$$U_{2\nu}(p)\sqrt{1-p^2} = \sin((l+1)\arccos p) \quad .$$

2 Model of experiment

We suppose that the unknown parameters of the model (4) are to be estimated on the base of a set of measurements of values (5) at different p 's:

$$y_j = F_m(p_j) + e_j \quad , \quad j = 1, ..., N \quad . \tag{7}$$

By using the standard notations and assumptions about errors e_j 's :

$$e = (e_1, ..., e_N)' \quad , \quad E(e) = 0, \ \mathrm{cov}(e) = \sigma^2 W^{-1} \quad , \quad W = \mathrm{diag}(w_1, ..., w_N) \quad , \quad \mathrm{tr}(W) = 1$$

$$u(p) = (u_0(p), ..., u_m(p))' \quad , \quad u_\nu(p) = \sqrt{1-p^2} U_{2\nu}(p) \quad , \quad \nu = 0, 1, ..., m$$

we can write the information matrix of the experiment (6) as the moment matrix of design measure ζ :

$$M(\zeta) = \int u(p)'u(p)\zeta(dp) \quad , \quad \mathrm{supp}\zeta \subset [0,1] \quad .$$

Here ζ is the discrete probability measure on interval $[0,1]$, which assigns some weights w_j 's to the points p_j 's. Since the covariance matrix of the best linear unbiased estimates (BLUE) of unknown parameters a_ν 's is proportional to M^{-1} , all the characterizations of the experiment accuracy are based on some functional's of information matrix $M(\zeta)$.

3 Optimal design

According to traditional definition, the D-optimal design ζ^* is the solution of maximizing problem

$$\mathrm{Det}\, M(\zeta^*) = \sup \mathrm{Det}\, M(\zeta) \tag{8}$$

where *supremum* is taken over all probability measures on interval $[0,1]$. This design minimizes the generalized variance of parameter estimates.

It is well known that the D-optimal design is invariant relative to all non-degenerate linear transforms of unknown parameters or basis functions. Taking into account that the model (5) contains only the even powers of p , it may be transform by substituting

$$t = 1 - 2p^2$$

into the model

$$F_m(t) = \sqrt{1+t} \sum_{k=0}^{m} c_k t^k \quad , \quad t \in [-1,+1] \quad . \tag{9}$$

This model is similar but not the same as the well known polynomial model with weight function which was studied by Karlin and Studden (1966). One of their result deals with the model

$$F_m(t) = w(t)^{\frac{1}{2}} \sum_{k=0}^{m} c_k t^k \quad , \quad t \in [-1,+1] \quad , \tag{10}$$

where weight function $w(t)$ is of the form

$$w(t) = (1-t)^{(\alpha+1)}(1+t)^{(\beta+1)} \quad , \quad \alpha > -1, \beta > -1 \quad .$$

The D-optimal design for model (5*b*) is defined in terms of zeros of Jacobi polynomial $P_{m+1}^{(\alpha,\beta)}(t)$. Formally our case corresponds to $\alpha = -1, \beta = 0$, but it should be treated anew , because Jacobi polynomial with index $\alpha = -1$ is not defined strictly.

The main result of this paper may be formulated as follows :

THEOREM The D-optimal design for model (5) is unique. It assigns the equal weights $(m+1)^{-1}$ to the points

$$p_j = \sqrt{(1-t_j)/2} \quad , \quad j = 0, 1, ..., m,$$

where t_j 's are the points of unique D-optimal design for model (5a) and are the roots of polynomial equation

$$(1-t)P_m^{(1,0)}(t) = 0$$

where $P_m^{(1,0)}(t)$ is the Jacobi polynomial of degree m and index's $\alpha = 1, \beta = 0$ for interval [-1,+1].

PROOF differs from that of Karlin and Studden (1966) only by calculation details. First from oscillation properties of the polynomials one can state that number of observation points is to be equal to the number of unknown parameters $m+1$ and that one of these points should be the bound point $t_0 = 1\,(p_0 = 0)$. All the rest m points t_j 's of optimal design are to be inner for interval $[-1, +1]$, and the polynomial

$$y(t) = (t - t_1)...(t - t_m)$$

which have the zeros in this points satisfies the differential equation

$$(1-t^2)y'' - (1+3t)y' + m(m+2)y = 0$$

which determines the Jacobi polynomial $P_m^{(1,0)}$ up to a constant factor.

Remark 3.1. By inverting the sign of t it is not difficult to see that D-optimal design for model

$$F_m(t) = \sqrt{1-t}\sum_{k=0}^{m} c_k t^k \quad , \quad t \in [-1, +1] \quad . \tag{11}$$

is defined by zeros of polynomial $(1+t)P_m^{(0,1)}(t)$ (see Bateman and Erdelyi (1953)).

4 Numerical results

For calculation of Jacobi polynomials we use the standard recurrence relations [Bateman and Erdelyi (1953)] :

$$P_{n+1}(t) = (a_n t + b_n)P_n(t) - c_n P_{n-1}(t)$$

$$a_n = \frac{2n+3}{n+2} \quad , \quad b_n = \frac{1}{(n+2)(2n+1)} \quad , \quad c_n = \frac{n(2n+3)}{(n+2)(2n+1)}.$$

The coefficients in the representation

$$P_n(t) = \sum_{k=0}^{n} A_k^n t^k$$

were obtained recursively from the induced relations:

$$A_k^{n+1} = a_n A_{k-1}^n + b_n A_k^n - c_n A_k^{n-1} \quad , \quad A_k^n = 0 \quad \text{if } k > n$$

$$A_0^0 = 1 \quad , \quad A_0^1 = \frac{1}{2} \quad , \quad A_1^1 = \frac{3}{2} \quad .$$

The zeros of polynomials were calculated by using a standard software.

A set of D-optimal designs is calculated for half-degrees $m = 1, .., 40$. But these calculations may be of little use if the full computation scheme remains unstable due to, for instance, matrix inversion which is needed for BLUE calculation.

To check the stability of full computation scheme we use the statement of the Kiefer - Wolfowitz equivalence theorem:

$$\max_{p \in [0,1]} d(\zeta^*, p) = m + 1$$

where

$$d(\zeta, p) = u(p)' M^{-1}(\zeta) u(p) \quad .$$

Since the condition above is a need and sufficient one for D-optimality it should be true up to calculation errors. Hence we may compute the left hand side of this equality and compare it with $m + 1$. It was found that,in spite of independence of D-optimal design on partial choice of polynomial basis, the computation stability varies significantly with this choice.For example, the simple theoretical model like (5a) with single-term basis shows the evident instability as half-degree m exceeds 6 even while using double precision arithmetic . On the contrary, the model (5) with the Chebyshev's basis does not show any sign of instability up to $m = 40$.

Part of D-optimal designs is presented here in the Table 1 (up to m=15).

The values D_{max} ,left hand side of D-optimality condition, are presented in the bottom lines of each part of the table.As one can see for all designs the statement of the Kiefer-Wolfowitz theorem is fulfilled. The values of Det(M) also presented in the table correspond to Chebyshev's basis (5).

5 Conclusion remarks

5.1. The D-optimal design according to Kiefer-Wolfowitz equivalence theorem is also G-optimal ,i.e. it minimizes the maximum of variance of BLUE of regression function - in our case of ray-sum model (5).But the main interest is concentrated on "density" model (4) for which D-optimal design is non- G- optimal.This is typical for inverse problems where a

model of regressions of observations does not coincide with a model of unknown function to be estimated on the base of experiment.For such experiments the search of "true" G-optimal design is more complicated problem.

5.2. As one can see from Table 1, for high degrees of the polynomial model the D-optimal design requires the very fine spatial resolution of the measurements, which leads in limit to bad accuracy due to low intensity of very narrow sounding beam. So we come to the problem of measurements of intensity distribution on the whole rather then of local values of ray-sums (5).A certain approach based on Banach space formulation of design problem for regression experiment developed early by the author (see Ermakov ed. (1983)) may be useful for designing of experiment of this type.

References

1. Bateman H. and Erdelyi A. 1953, Higher Transcendental Functions,vol.II, New York, McGraw-Hill.

2. Cormack A.M. 1964, Representation of a Function by Its Line Integrals, with Some Radiological Applications, II, J.of Appl. Phys., vol.35, no.10, pp.2908-2913.

3. Ermakov S.M. ed. 1983, Mathematical Theory of Design of Experiment, Moscow, Nauka.

4. Karlin S.and Studden W.J. 1966, Optimal experimental designs, Annals of Math. Statist., vol.37, no.4, pp.783-815.

5. Kosarev E.L. 1973, On numerical solution of Abel equation, J.vychislit. matematiki i matematich. fisiki, t.13, no.6, ss. 1591-1595.

6. Minerbo G.N. and Levy M.E. 1969, Inversion of Abel's integral equation by means of orthogonal polynomials, SIAM J. Numer. Anal., vol.6, no.4, pp.598-616.

Table 1. POINTS OF D-OPTIMAL DESIGN FOR ABEL INVERSION

j	m= 1	m= 2	m= 3	m= 4	m= 5
0	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.816497	0.595862	0.460804	0.373845	0.313903
2		0.919211	0.768462	0.645298	0.551848
3			0.954679	0.850386	0.749683
4				0.971028	0.895537
5					0.979893
D_{max}	2.000000	3.000000	4.000000	5.000000	6.000000
$Det(M)$	$5.9259\ 10^{-1}$	$3.2768\ 10^{-1}$	$1.7601\ 10^{-1}$	$9.3045\ 10^{-2}$	$4.8689\ 10^{-2}$
j	m= 6	m= 7	m= 8	m= 9	m=10
0	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.270286	0.237197	0.211267	0.190415	0.173289
2	0.480381	0.424548	0.379956	0.343626	0.313511
3	0.664326	0.593822	0.535560	0.487008	0.446118
4	0.814258	0.739698	0.674398	0.617967	0.569259
5	0.922996	0.856861	0.792507	0.733505	0.680633
6	0.985233	0.940915	0.886392	0.830862	0.778049
7		0.988696	0.953245	0.907680	0.859558
8			0.991070	0.962088	0.923520
9				0.992768	0.968642
10					0.994023
D_{max}	7.000000	8.000000	9.000000	10.000000	11.000000
Det(M)	$2.5300\ 10^{-2}$	$1.3079\ 10^{-2}$	$6.7349\ 10^{-3}$	$3.4573\ 10^{-3}$	$1.7703\ 10^{-3}$
j	m=11	m=12	m=13	m=14	m=15
0	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.158977	0.146840	0.136420	0.127376	0.119455
2	0.288169	0.266567	0.247943	0.231727	0.217485
3	0.411309	0.381376	0.355394	0.332652	0.312591
4	0.527064	0.490307	0.458086	0.429663	0.404437
5	0.633642	0.591945	0.554886	0.521846	0.492276
6	0.729289	0.684872	0.644636	0.608246	0.575313
7	0.812399	0.767760	0.726234	0.687941	0.652772
8	0.881566	0.839416	0.798667	0.760069	0.723918
9	0.935618	0.898804	0.861032	0.823850	0.788074
10	0.973635	0.945064	0.912551	0.878587	0.844626
11	0.994978	0.977524	0.952578	0.923686	0.893033
12		0.995721	0.980613	0.958652	0.932828
13			0.996311	0.983107	0.963632
14				0.996786	0.985149
15					0.997176
D_{max}	12.000000	13.000000	14.000000	15.000000	16.000000
Det(M)	$9.0461\ 10^{-4}$	$4.6145\ 10^{-4}$	$2.3504\ 10^{-4}$	$1.1956\ 10^{-4}$	$6.0753\ 10^{-5}$

Minimizing the Largest of the Parameter Variances. V(β)-optimality

Jesus López-Fidalgo

We introduce a new criterion function that directly minimizes the variances of the estimators of the parameters. Covariances are ignored. Although the calculations are more laborious than in other more manageable criteria, such as that of D-optimality, we gain in fidelity to our objective.

1 Introduction

In this section we establish notation. For an exhaustive introduction to the optimal experimental design see FEDOROV (1972), SILVEY (1980) or PAZMAN (1986).

Let X be the **design space** or **experimental domain**, which is assumed to be compact. Let $y(x)$ be the observation carried out at point x, which will be a random variable with known variance and unknown mean. By state, a function of the type:

$$\theta \ : \ X \rightarrow \mathbf{R} / \theta(x) = E\{y(x)\},$$

will be understood. The set of all states will be a linear space and it will be denoted by Θ. It is convenient to study the **linear regression model with uncorrelated observations:**

$$\theta(x) = E(y|x) = \alpha^t f(x) = \alpha_1 f_1(x) + \ldots + \alpha_m f_m(x),$$

where $f^t(x) = (f_1(x), \ldots, f_m(x))$ is a continuous known vector function.

Let a design be a discrete probability measurement, ξ. The support of the design will be given by $X_\xi = \{x \in X \ : \ \xi(x) > 0\}$.

The information matrix associated with the design ξ is defined as :

$$M(\xi) = \sum_{x \in X} f(x) f^t(x) \sigma^{-2}(x) \xi(x).$$

The image set of this matrix will be denoted:

$$\mathcal{M}[(\xi)] = \{M(\xi)u \ : \ u \in R^m\}.$$

Let g be a functional on the space Θ and ξ a design. We define:

$$\mathrm{var}_\xi g = \begin{cases} g^t M^-(\xi) g & \text{if } g \in \mathcal{M}[M(\xi)] \\ \infty & \text{if } g \notin \mathcal{M}[M(\xi)], \end{cases}$$

where g is the vector corresponding to g on the basis $f_1, \ldots, f_m$ of the linear space Θ.

The set of all the designs in the model will be denoted by Ξ, whereas the set of all the information matrices will be:

$$\mathcal{M} = \{M(\xi) \ : \ \xi \in \Xi\} \text{ and } \mathcal{M}_+ = \{M(\xi) \in \mathcal{M} \ : \ \det M(\xi) \neq 0\}.$$

A criterion function (see PAZMAN, 1986) will be a $\Phi : \mathcal{M} \to \mathbf{R} \cup \{+\infty\}$ function bounded from below, such that:

$$\mathrm{var}_\xi g \leq \ \mathrm{var}_\eta g \ \ \forall \text{ functional g of } \Theta \Rightarrow \Phi[M(\xi)] \leq \Phi[M(\eta)]$$

that satisfies the properties:

a) U_Φ, an open set of the $\mathcal{L}(\mathcal{M})$, the linear space of symmetric matrices spanned by the set $\mathcal{M}$, exists such that $\mathcal{M}_+ \subset U_\Phi$, and Φ is defined, finite and convex in U_Φ

b) If $M_n \in \mathcal{M}_+, n = 1, 2, \ldots$ and $\lim_{n\to\infty} M_n = M \in \mathcal{M} - \mathcal{M}_+$ then:

$$\lim_{n\to\infty} \Phi(M_n) = \infty.$$

A design that minimizes $\Phi[M(\xi)]$ will be called Φ-optimum.

2 Definition of the criterion and properties

Definition 1.- We define the following criterion function:

$$\Phi_{V(\beta)}\{M(\xi)\} = \max_i \ \mathrm{var}_\xi(\alpha_i) = \begin{cases} \max_i \{M^{-1}(\xi)\}_{ii} & \text{if } \det M(\xi) \neq 0 \\ \infty & \text{if } \det M(\xi) = 0. \end{cases}$$

In the following propositions we demonstrate that it is truly a criterion function.

Proposition 1. (PAZMAN, 1986, page 63).- If g is a functional, the function $\Psi\{M(\xi)\} = \mathrm{var}_\xi(g)$ is lower semicontinuous. It is continuous in $M(\xi)$ if $\det M(\xi) \neq 0$ or if $\mathrm{var}_\xi(g) = \infty$.

Proposition 2. The criterion function defined above is continuous.

Proof: In the open set of nonsingular information matrices this function is the maximum of a finite family of continuous functions, and is thus continuous. If $M(\xi_0)$ is singular; then i exists such that α_i is not estimable, and thus $\mathrm{var}_\xi(\alpha_i) = \infty$, from which it is deduced that $\Phi_{V(\beta)}[M(\xi)] = \infty$, and we know that the function:

$$\varphi_i : \mathcal{M} \to \mathbf{R} / \varphi_i\{M(\xi)\} = \ \mathrm{var}_\xi(\alpha_i)$$

is continuous in $M(\xi_0)$, so that for each K that we establish there will be a neighborhood of $M(\xi_0)$, U, such that if $M(\xi) \in U$, then $\varphi_i\{M(\xi)\} \geq K$ and therefore, also:

$$\Phi_{V(\beta)}\{M(\xi)\} = \max_i \text{var}_\xi(\alpha_i) \geq K,$$

and we have thus proved the continuity of our function.

Proposition 3. The $\Phi_{V(\beta)}$ function is convex in the set of information matrices, and strictly convex in the set of nonsingular matrices.

Proof: This function is the maximum of a finite family of convex functions and therefore convex (see PAZMAN, 1986, pp 62). Moreover, the convexity will be strict in the set of nonsingular matrices.

3 Differentiability of the V(β)-optimality criterion function

Proposition 4. The function ϕ_i is differentiable in the set of nonsingular matrices.

Proof: We denote by E_{ij} the matrix whose elements are all null except the one situated in row i, column j, which equals one. Thus, we have:

$$\frac{\partial}{\partial M_{ij}}\{M^{-1}(\xi)\}_{kk} = \lim_{\epsilon \to 0} \frac{[\{M(\xi) + \epsilon E_{ij}\}^{-1}]_{kk} - \{M^{-1}(\xi)\}_{kk}}{\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\{M(\xi)+\epsilon E_{ij}\}_{kk}}{\det\{M(\xi)+\epsilon E_{ij}\}} - \frac{M_{kk}(\xi)}{\det M(\xi)}}{\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{M_{kk}(\xi)+\epsilon M^{ij}_{kk}(\xi)}{\det M(\xi)+\epsilon M_{ij}(\xi)} - \frac{M_{kk}(\xi)}{\det M(\xi)}}{\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{M_{kk}(\xi)\det M(\xi) + \epsilon M^{ij}_{kk}(\xi)\det M(\xi) - M_{kk}(\xi)\det M(\xi) - \epsilon M_{ij}(\xi)M_{kk}(\xi)}{\epsilon \det M(\xi)\{\det M(\xi) + \epsilon M_{ij}(\xi)\}}$$

$$= \lim_{\epsilon \to 0} \frac{\epsilon\{M^{ij}_{kk}(\xi)\det M(\xi) - M_{ij}(\xi)M_{kk}(\xi)\}}{\epsilon \det M(\xi)\{\det M(\xi) + \epsilon M_{ij}(\xi)\}} = \frac{M^{ij}_{kk}(\xi)\det M(\xi) - M_{ij}(\xi)M_{kk}(\xi)}{\det M(\xi)^2}$$

where $M_{ij}(\xi)$ is the cofactor in $M(\xi)$ of component i,j. Also $M^{ij}_{kk}(\xi)$ is the determinant of the matrix which results from the elimination of rows i and k, and columns j and k, multiplied by the factor $(-1)^{i+j}$ if $i,j < k$ or $i,j > k$, or multiplied by the factor $(-1)^{i+j+1}$ otherwise. Therefore the gradient takes the form:

$$\left\{\frac{M^{ij}_{kk}(\xi)\det M(\xi) - M_{ij}(\xi)M_{kk}(\xi)}{\det M(\xi)^2}\right\}_{ij.}$$

However, this does not assure the differentiability of our criterion function for nonsingular matrices. We shall, then, use the following definition:

Definition 2: For each $k = 1, 2, \ldots, m$ the sets:

$$H_k = \{M(\xi) \in \mathcal{M} / \max_i \text{var}_\xi(\alpha_i) = \text{var}_\xi(\alpha_k)\}$$

$$H_k^+ = H_k \cap \mathcal{M}_+$$

$$\jmath_{s,\tau} = \overline{(\bigcap_{k=1}^{s} H^+_{\tau(k)} - \bigcap_{k=s+1}^{m} H^+_{\tau(k)})}, \quad \tau \in S_m, s = 1, 2, \ldots, m$$

are defined, where we denote the symmetric group of order m by S_m and the sets:

$$\bigcap_{k=1}^{s} H^+_{\tau(k)} - \bigcap_{k=s+1}^{m} H^+_{\tau(k)}, \tau \in S_m, \quad s = 1, 2, \ldots, m$$

form a partition of $\mathcal{M}_+$ and the $\jmath_{s,\tau}$ for $\tau \in S_m, s = 1, 2, \ldots, m$ are the topological closure of these sets in $\mathcal{M}$. In accordance with the notation used up to now we denote:

$$\jmath^+_{s,\tau} = \{M \in \jmath_{s,j} : M \text{ is nonsingular}\}.$$

We can then establish the following result:

Proposition 5. The sets $\jmath_{s,\tau}$ for $\tau \in S_m, s = 1, 2, \ldots, m$ are compact.

Proof: Since they are closed sets contained in a compact $\mathcal{M}$, it is necessary to demonstrate that the interior of the sets:

$$\bigcap_{k=1}^{s} H^+_{\tau(k)} - \bigcap_{k=s+1}^{m} H^+_{\tau(k)}, \tau \in S_m, s = 1, 2, \ldots, m$$

coincides with the interior of sets $\jmath_{s,\tau}$ for $\tau \in S - m, s = 1, 2, \ldots, m$, which is demonstrated by the following lemma of basic topology:

Lemma 1: Let A be a convex subset of a convex set $\mathcal{M}$ contained in the metric linear space E. Then:

a) A° and $\overline{A}$ are also convex in $\mathcal{M}$.

b) $A^\circ = \overline{A^\circ}$

Proposition 6. If the sets $\jmath_{s,\tau}$ are locally convex and connected, then the function of definition 1 will be differentiable in the interior of each of the sets:

$$\bigcap_{k=1}^{s} H^+_{\sigma(k)} - \bigcap_{k=s+1}^{m} H^+_{\sigma(k)}, \sigma \in S_m, k = 1, 2, \ldots, m$$

and its gradient is given by:

$$\nabla\Phi_V[M(\xi)] = \{\frac{M_{tt}^{ij}(\xi)\det M(\xi) - M_{ij}(\xi)M_{tt}(\xi)}{\det M(\xi)^2}\}_{ij}.$$

where

$$M_{tt}(\xi) = \max_i M_{ii}(\xi).$$

Proof: In order to calculate the gradient, we will work in the neighborhood opened in $\mathcal{L}(\jmath_{s,r})$ of each one of the $\jmath_{s,r}^{+}$ defined in the form:

$$U_{s,r} = \{M \in \mathcal{L}(\jmath_{s,r}) : M \text{ is nonsingular}\},$$

and we shall refer to the gradient in that sense. The proof will be carried out in the interior of the set:

$$\bigcap_{k=1}^{s} H_k^{+} - \bigcap_{k=s+1}^{m} H_k^{+},$$

which does not entail a loss of generality. We shall then calculate:

$$\frac{\partial\Phi_{V(\beta)}[M(\xi)]}{\partial M_{ij}} = \lim_{\epsilon\to 0}\frac{\Phi_{V(\beta)}[M(\xi)+\epsilon E_{ij}] - \{M^{-1}(\xi)\}_{11}}{\epsilon},$$

and now if we assume that:

$$\Phi_{V(\beta)}[M(\xi)+\epsilon E_{ij}] = \max_l\{[M(\xi)+\epsilon E_{ij}]^{-1}\}_{ll} = \{[M(\xi)+\epsilon E_{ij}]^{-1}\}_{tt},$$

then :

$$\frac{\partial\Phi_{V(\beta)}[M(\xi)]}{\partial M_{ij}} = \lim_{\epsilon\to 0}\frac{\{[M(\xi)+\epsilon E_{ij}]^{-1}\}_{tt} - \{M^{-1}(\xi)\}_{11}}{\epsilon} =$$

$$\lim_{\epsilon\to 0}\frac{\frac{[M(\xi)+\epsilon E_{ij}]_{tt}}{\det[M(\xi)+\epsilon E_{ij}]} - \frac{M_{11}(\xi)}{\det M(\xi)}}{\epsilon} = \lim_{\epsilon\to 0}\frac{\frac{M_{tt}(\xi)+\epsilon M_{tt}^{ij}(\xi)}{\det M(\xi)+\epsilon M_{ij}(\xi)} - \frac{M_{11}(\xi)}{\det M(\xi)}}{\epsilon}.$$

However, the previous maximum will be reached at $k = t$ when:

$$\frac{M_{tt}(\xi)+\epsilon M_{tt}^{ij}(\xi)}{\det M(\xi)+\epsilon M_{ij}(\xi)} \geq \frac{M_{ll}(\xi)+\epsilon M_{ll}^{ij}(\xi)}{\det M(\xi)+\epsilon M_{ij}(\xi)}, \quad l = 1,2,\ldots,m$$

is satisfied, that is, when:

$$M_{tt}(\xi)+\epsilon M_{tt}^{ij}(\xi) \geq M_{ll}(\xi)+\epsilon M_{ll}^{ij}(\xi), \quad l = 1,2,\ldots,m,$$

and when ϵ is made sufficiently small, this maximum will be reached when $M_{tt}(\xi) = \max M_{ii}(\xi)$, so that t does not depend on i,j.

Once this is defined, we can continue with the calculation of the gradient:

$$\frac{\partial \Phi_{V(\beta)}\{M(\xi)\}}{\partial M_{ij}} =$$

$$= \lim_{\epsilon \to 0} \frac{M_{tt}(\xi) \det M(\xi) + \epsilon M_{tt}^{ij}(\xi) \det M(\xi) - M_{tt}(\xi) \det M(\xi) - \epsilon M_{ij}(\xi) M_{tt}(\xi)}{\epsilon \det M(\xi)\{\det M(\xi) + \epsilon M_{ij}(\xi)\}}$$

$$= \lim_{\epsilon \to 0} \frac{\epsilon\{M_{tt}^{ij}(\xi) \det M(\xi) - M_{ij}(\xi) M_{tt}(\xi)\}}{\epsilon \det M(\xi)\{\det M(\to \xi) + \epsilon M_{ij}(\xi)\}} = \frac{M_{tt}^{ij}(\xi) \det M(\xi) - M_{ij}(\xi) M_{tt}(\xi)}{\det M(\xi)^2}$$

4 Computation of error

We adapt two results of the general theory of experimental optimal design for the calculation of the error committed on taking a design as optimal using $V(\beta)$-optimality.

Proposition 7. Let us assume that $\Phi_{V(\beta)}[M(\mu)] < \infty$ and let $\delta > 0$, thus satisfying $\partial \Phi_{V(\beta)}[M(\mu), M] \geq -\delta, \quad M \in M$ then the bound:

$$\Phi_{V(\beta)}[M(\mu)] - \inf\{\Phi_{V(\beta)}[M(\xi)] | \xi \in \Xi\} \leq \delta$$

is satisfied

Proof: It is automatically transferred from proposition IV.28, PAZMAN (1986).

Proposition 8. -If $M(\mu) \in \jmath_{*,\tau}^{+}$ and $\delta > 0$ so that:

$$f^t(x) \nabla \Phi_{V(\beta)}[M(\mu)] f(x) \geq \text{ tr } \nabla \Phi_{V(\beta)}[M(\mu)] M(\mu) - \delta, \quad x \in X,$$

then the bound:

$$\Phi_V[M(\mu)] - \inf\{\Phi_V[M(\xi)] | M(\xi) \in \jmath_{*,\tau}\} \leq \delta$$

is verified.

Proof: It is automatically transferred by changing $\mathcal{M}$ for $\jmath_{*,\tau}$ in the proposition V.2, PAZMAN (1986).

Observation.- The disadvantage of the latter proposition with respect to the former is that it calculates the error in each of the $\jmath_{*,\tau}$ and not globally. However, it is undoubtedly of interest since it works with the gradient and not with the directional derivative.

5 Calculation of the gradient of the V(β)-optimality criterion function in the biparametric case

The problem which now arises is as follows:

$$\theta(x) = \alpha f(x) + \beta g(x), \quad x \in X,$$

where we can assume that $\sigma(x) = 1$ just by redefining the problem taking $f^{\sim}(x) = \sigma^{-1}(x)f(x)$. So that the information matrix will have the form:

$$M(\xi) = \begin{pmatrix} \sum_{x \in X} \xi(x) f(x)^2 & \sum_{x \in X} \xi(x) f(x) g(x) \\ \sum_{x \in X} \xi(x) f(x) g(x) & \sum_{x \in X} \xi(x) g(x)^2 \end{pmatrix}$$

For simplicity of notation we shall call:

$$a = \sum_{x \in X} \xi(x) f(x)^2, \quad b = \sum_{x \in X} \xi(x) f(x) g(x), \quad c = \sum_{x \in X} \xi(x) g(x)^2,$$

so that the inverse matrix of the information matrix, whenever it exists, will be:

$$M^{-1}(\xi) = \begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix} =$$

$$= \frac{1}{\det M(\xi)} \begin{pmatrix} \sum_{x \in X} \xi(x) g(x)^2 & -\sum_{x \in X} \xi(x) f(x) g(x) \\ -\sum_{x \in X} \xi(x) f(x) g(x) & \sum_{x \in X} \xi(x) f(x)^2 \end{pmatrix},$$

and the criterion function is then:

$$\Phi_{V(\beta)}[M(\xi)] = \max\{\frac{c}{ac - b^2}, \frac{a}{ac - b^2}\},$$

and since the information matrix is always positive semidefinite, and we are assuming this one to be nonsingular, then $ac - b^2 > 0$. Therefore:

$$\begin{aligned} H_1 &= \{M(\xi) \text{ nonsingular} : c \geq a\} \cup \{\alpha_1 \text{ is not estimable for } \xi\} \\ &= \{M(\xi) \text{ non singular} : \sum_{x \in X} \xi(x)\{g(x)^2 - f(x)^2\} \geq 0\} \cup \{\alpha_1 \text{ is not estimable for } \xi\} \end{aligned}$$

$$\begin{aligned} H_2 &= \{M(\xi) \text{ nonsingular} : c \leq a\} \cup \{\alpha_2 \text{ is not estimable for } \xi\} \\ &= \{M(\xi) \text{ nonsingular} : \sum_{x \in X} \xi(x)\{g(x)^2 - f(x)^2\} \leq 0\} \cup \{\alpha_2 \text{ is not estimable for } \xi\}. \end{aligned}$$

Let us now calculate the gradient in:

$$\mathcal{J}_1^{\circ} = \overline{H_1 - H_2}^{\circ} = \{M(\xi) \text{ nonsingular} : c > a\}.$$

We will have:

$$\Phi_{V(\beta)}\{M(\xi)\} = \frac{\sum_{x \in X} \xi(x) g(x)^2}{\det M(\xi)} = \frac{c}{ac - b^2},$$

so that the gradient will be:

$$[\nabla \Phi_{V(\beta)}\{M(\xi)\}]_{ij} = \lim_{\epsilon \to 0^+} \frac{\Phi_{V(\beta)}\{M(\xi) + \epsilon E_{ij}\} - \Phi_{V(\beta)}\{M(\xi)\}}{\epsilon},$$

but:

$$\{M(\xi) + \epsilon E_{11}\}^{-1} = \frac{1}{(a+\epsilon)c - b^2} \begin{pmatrix} c & -b \\ -b & a+\epsilon \end{pmatrix},$$

and taking a sufficiently small ϵ, the determinant continues to be positive and:

$$\Phi_{V(\beta)}\{M(\xi) + \epsilon E_{11}\} = \frac{c}{(a+\epsilon)c - b^2},$$

and the gradient is:

$$[\nabla \Phi_{V(\beta)}\{M(\xi)\}]_{11} = \lim_{\epsilon \to 0^+} \frac{\frac{c}{(a+\epsilon)c - b^2} - \frac{c}{ac - b^2}}{\epsilon} = \frac{-c^2}{(ac - b^2)^2}.$$

Analogously:

$$[\nabla \Phi_{V(\beta)}\{(\xi)\}]_{12} = \lim_{\epsilon \to 0^+} \frac{\frac{c}{ac - (b+\epsilon)b} - \frac{c}{ac - b^2}}{\epsilon} = \frac{cb}{(ac - b^2)^2}$$

$$[\nabla \Phi_{V(\beta)}\{M(\xi)\}]_{21} = \frac{cb}{(ac - b^2)^2}$$

$$[\nabla \Phi_{V(\beta)}\{M(\xi)\}]_{22} = \lim_{\epsilon \to 0^+} \frac{\frac{c+\epsilon}{a(c+\epsilon) - b^2} - \frac{c}{ac - b^2}}{\epsilon} = \frac{-b^2}{(ac - b^2)^2}.$$

Therefore in $\overline{H_1 - H_2}^{\circ}$ the gradient of the criterion function is:

$$\nabla \Phi_{V(\beta)}\{M(\xi)\} = \frac{1}{(ac - b^2)^2} \begin{pmatrix} -c^2 & cb \\ cb & -b^2 \end{pmatrix}.$$

In the same way the gradient in:

$$\mathcal{J}_2^{\circ} = \overline{H_2 - H_1}^{\circ} = \{M(\xi) \text{ non singular}: c < a\}$$

will be:

$$\nabla\Phi_{V(\beta)}\{M(\xi)\} = \frac{1}{(ac-b^2)^2}\begin{pmatrix} -b^2 & ab \\ ab & -a^2 \end{pmatrix}.$$

On the other hand:

$$j_3^\circ = \overline{H_1 \bigcap H_2}^\circ = \emptyset.$$

6 Discussion

The interest in this criterion is found in the fact that it only concerns the minimization of the largest parameter variances, which is its main objective. Thus, it does not care about covariances. It is useful to remark that while E-optimality criterion function is the maximum of the variances on the unitary functionals, $V(\beta)$-optimality is constrained to the functionals of the canonical basis. In other words, while E- optimality minimizes the inverse eigenvalues of the information matrix, $V(\beta)$-optimality minimizes the largest diagonal element of the inverse of the information matrix.

We find two disadvantages connected with this criterion:

1. It depends upon the scaling of the points of the space X. In fact, with some suitable scaling the criterion is equivalent to other known criteria.

2. The sequential algorithm for constructing $V(\beta)$-optimal designs is complicated. For example, when using algorithms involving the gradient it is necessary to seek on the different sets $j_{s,\tau}$ with the corresponding gradient on each one. This procedure is done in a similar way as that of ATKINSON and FEDOROV (1975) for discriminating between three or more models.

References

1. ATKINSON A. C. and FEDOROV V.V. (1975). Optimal design: Experiments for discriminating between several models. Biometrika, Vol. 62, No. 2, pp 289-303.

2. FEDOROV V.V. (1972). Theory of optimal experiments. Academic Press. New York.

3. FEDOROV V.V. (1980). Convex design theory. Math. Operationsforsch. Statist., Ser.statistics., Vol. 11, No. 3, pp 403-413.

4. PAZMAN A. (1986). Foundations of optimum experimental design. D. Reidel publishing company. Dordrecht.

5. SILVEY S.D. (1980). Optimal design. Chapman and Hall. London.

Some Two-Stage Procedures for Treating the Behrens-Fisher Problem

Rainer Schwabe

For designing an experiment some knowledge about the underlying model is crucial. In particular, usually a known variance-covariance structure is assumed. If however, the variances might vary from level to level of some controlled factores the situation becomes more complicated. One attempt to attack this problem is to use a two-stage procedure: On the first stage a fixed number of experiments is made and the variances are estimated; on the second stage the number of additional experiments at each level combination is determined according to these estimates. We will illustrate this procedure by treating the situation of comparing the means of two groups with possibly different variances which is known as the Behrens-Fisher problem.

1 Introduction

One of the oldest still not completely solved problems in statistics is the comparison of the means of two populations with possibly different variances. This problem dates back at least to some work of Behrens in 1929 and has become more popular by the work of Fisher in 1935 who exhibited his fiducial arguments in this setting. Due to these roots the following situation has been coined the Behrens-Fisher problem (for further references see e.g. Stuart and Ord(1991)):

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent observations from two populations normally distributed with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. We are interested in the magnitude of the difference $\mu_1 - \mu_2$ of the means or, in particular, wether the means are equal (i.e. $\mu_1 - \mu_2 = 0$) or not.

An obvious statistic for the difference $\mu_1 - \mu_2$ is the difference $\overline{X^{(n_1)}} - \overline{Y^{(n_2)}}$ of the arithmetic means $\overline{X^{(n_1)}} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $\overline{Y^{(n_2)}} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i$ respectively. $\overline{X^{(n_1)}} - \overline{Y^{(n_2)}}$ is known to be normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2$ or equivalently

$$\frac{\overline{X^{(n_1)}} - \overline{Y^{(n_2)}} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

is standard normal.

We note that the variance of this difference is minimized subject to a fixed total amount $n = n_1 + n_2$ of observations if the ratio n_1/n_2 is approximately equal to σ_1/σ_2 . In this case $\frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2 \approx \frac{1}{n}(\sigma_1 + \sigma_2)^2$ which for unequal underlying variances $\sigma_1^2 \neq \sigma_2^2$ is reasonably smaller than the value $\frac{2}{n}(\sigma_1^2 + \sigma_2^2)$ for both: equal sample sizes $n_1 = n_2$ and sample sizes proportional to the variances $n_1/n_2 \approx \sigma_1^2/\sigma_2^2$ (for further readings in the theory of optimum

design (optimum allocation), we refer for example to the monographs of Fedorov(1972), Bandemer e. a.(1977), Silvey(1980), and Pázman(1986)).

If, however, as in the present case the variances σ_1^2 and σ_2^2 are unknown neither the distribution of $\overline{X^{(n_1)}} - \overline{Y^{(n_2)}}$ can be totally specified nor a good choice of n_1 and n_2 can be given in general. Only for the very special and in most situations unrealistic case that the ratio of the variances $\varrho = \sigma_1^2/\sigma_2^2$ is known the difference $\overline{X^{(n_1)}} - \overline{Y^{(n_2)}}$ can be studentized appropriately such that

$$\frac{\overline{X^{(n_1)}} - \overline{Y^{(n_2)}} - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{\varrho n_2})(\frac{n_1 - 1}{n_1 + n_2 - 2}\widehat{\sigma}_1^2 + \varrho\frac{n_2 - 1}{n_1 + n_2 - 2}\widehat{\sigma}_2^2)}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, where $\widehat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1 - \overline{X^{(n_1)}})^2$ and $\widehat{\sigma}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y^{(n_2)}})^2$ are the usual estimates for the variances σ_1^2 and σ_2^2 respectively. Also n_1/n_2 approximately equal to $\sqrt{\varrho}$ will be optimum as can easily be checked.

In general we are only able to insert the estimates $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ for the unknown variances in (1) and arrive at the Behrens-Fisher statistic

$$\frac{\overline{X^{(n_1)}} - \overline{Y^{(n_2)}} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1}\widehat{\sigma}_1^2 + \frac{1}{n_2}\widehat{\sigma}_2^2}}. \tag{2}$$

As the distribution of the Behrens-Fisher statistic cannot be derived explicitly we will give bounds in the next section (see Mickey and Brown(1966) and Scheffé(1970)).

In the following sections we will mainly be concerned with the design problem of choosing sample sizes such that either $\mu_1 - \mu_2$ can be estimated with a prescribed variance or a confidence interval for $\mu_1 - \mu_2$ of bounded length can be obtained. The latter procedure also yields a test for equality of the means with prescribed power by the usual interplay between tests and confidence regions. Because even in the one-sample situation estimates of prescribed accuracy cannot be obtained by a fixed sample size if the variance is unknown we introduce Stein's(1945) two-stage procedure for that case in section 3 (for more details we refer to Chatterjee(1991) and Stuart and Ord(1991)).

Section 4 is dedicated to direct applications of Stein's two-stage procedures by Chapman(1950) and Ghosh(1975) to the present situation of the Behrens-Fisher problem and a modified, more reasonable procedure is introduced. In section 5 the two-stage procedure is combined with a pretended optimum allocation rule (Banerjee(1967)) which will be compared to the procedures of section 4. In section 6 we end up with some remarks on a fixed total sample size and further applications.

2 The special d-solution

Let T_ν denote a t-distributed random variable with ν degrees of freedom and $t_\nu;\alpha$ the α-quantile of that distribution. According to the results of Mickey and Brown(1966) we notice that the Behrens-Fisher statistic (2) is less dispersed than a t-distribution with $\nu = \min(n_1 - 1, n_2 - 1)$ degrees of freedom:

Theorem 1

For all t:

$$P\left(\left|\frac{\overline{X^{(n_1)}} - \overline{Y^{(n_2)}} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1}\widehat{\sigma}_1^2 + \frac{1}{n_2}\widehat{\sigma}_2^2}}\right| \geq t\right) \leq P(|T_\nu| \geq t)\,. \qquad \square$$

For a proof of Theorem 1 see Mickey and Brown(1966). Scheffé's(1970) special d-solution is a direct consequence of this result:

Corollary 1

The interval given by the bounds $\overline{X^{(n_1)}} - \overline{Y^{(n_2)}} \pm t_{\nu;1-\alpha/2}\sqrt{\frac{1}{n_1}\widehat{\sigma}_1^2 + \frac{1}{n_2}\widehat{\sigma}_2^2}$ is a confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$.

3 Stein's two-stage procedures

In his pioneering paper Stein(1945) introduced two two-stage procedures for estimating the mean μ of a normal distribution with given accuracy:

Let $(X_i)_{i\in\mathbb{N}}$ be *iid* normally distributed with mean μ and variance σ^2. On the first stage we observe a preliminary sample $(X_1, \ldots, X_{n_0})$ of size n_0 and estimate the unknown variance σ^2 by $\widehat{\sigma}_0^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(X_i - \overline{X^{(n_0)}})^2$. After that we determine the total sample size $N \geq n_0$ depending on the estimated variance $\widehat{\sigma}_0^2$ and take $N - n_0$ additional observations at the second stage if $N > n_0$.

For any given $z > 0$ (the actual value of which will be determined later) we can describe Stein's(1945) procedures as follows:

Stein's first procedure:

Choose $N = \max([\widehat{\sigma}_0^2/z] + 1, n_0 + 1)$, where $[a]$ denotes the largest integer less or equal to a. Then there exist random variables A_0 and A_1 such that

$$\begin{aligned} n_0 A_0 + (N - n_0)A_1 &= 1 \\ n_0 A_0^2 + (N - n_0)A_1^2 &= \frac{z}{\widehat{\sigma}_0^2}\frac{1}{N}\,. \end{aligned}$$

Define $\widetilde{X^{(N)}} := A_0\sum_{i=1}^{n_0} X_i + A_1\sum_{i=n_0+1}^{N} X_i$. Then $\frac{\widetilde{X^{(N)}}-\mu}{\sqrt{z}}$ has a t-distribution with $n_0 - 1$ degrees of freedom. Hence $E(\widetilde{X^{(N)}}) = \mu_1$ for $n_0 \geq 3$, $\mathrm{Var}(\widetilde{X^{(N)}}) = z\frac{n_0-1}{n_0-3}$ for $n_0 \geq 4$, and $P(|\widetilde{X^{(N)}} - \mu| \leq t) = P(|T_{n_0-1}| \leq \frac{d}{\sqrt{z}})$.

Stein's second procedure:

Choose $N = \max([\widehat{\sigma}_0^2/z] + 1, n_0)$. Let $\overline{X^{(N)}} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then $\sqrt{N}\frac{\overline{X^{(N)}}-\mu}{\sqrt{\widehat{\sigma}_0^2}}$ has a t-distribution with $n_0 - 1$ degress of freedom. Hence $E(\overline{X^{(N)}}) = \mu$ for $n_0 \geq 3$,

$$\mathrm{Var}(\overline{X^{(N)}}) = E\left(N\frac{(\overline{X^{(N)}}-\mu)^2}{\widehat{\sigma}_0^2}\frac{\widehat{\sigma}_0^2}{N}\right) \leq z \cdot \mathrm{Var}\left(\sqrt{N}\frac{\overline{X^{(N)}}-\mu}{\sqrt{\widehat{\sigma}_0^2}}\right) = z\frac{n_0-1}{n_0-3}\,,$$

for $n_0 \geq 4$, and

$$\begin{aligned} P(|\overline{X^{(N)}} - \mu| \leq t) &= P\left(\left|\sqrt{N}\frac{\overline{X^{(N)}}-\mu}{\sqrt{\widehat{\sigma}_0^2}}\right| \leq t\sqrt{\frac{N}{\widehat{\sigma}_0^2}}\right) \\ &\geq P\left(\left|\sqrt{N}\frac{\overline{X^{(N)}}-\mu}{\sqrt{\widehat{\sigma}_0^2}}\right| \leq \frac{t}{\sqrt{z}}\right) \\ &= P(|T_{n_0-1}| \leq \tfrac{t}{\sqrt{z}}). \end{aligned}$$

Now for both procedures an estimate of μ of prescribed accuracy $\mathrm{Var}(\widetilde{X^{(N)}}) = c$ resp. $\mathrm{Var}(\overline{X^{(N)}}) \leq c$ is given if the constant z is chosen according to $z = \frac{n_0-3}{n_0-1}c$. Similarly $z = d^2/t^2_{n_0-1;1-\alpha/2}$ yields a cofindence interval for μ with bounds $\widetilde{X^{(N)}} \pm d$ resp. $\overline{X^{(N)}} \pm d\sqrt{\widehat{\sigma}_0^2/(Nz)}$ of bounded length $2d$ with confidence level $1-\alpha$.

Although for Stein's procedures the distribution is given exactly only for the first one the second exhibits a lot of advantages: first the sample size is smaller in case $\widehat{\sigma}_0^2/z < n_0$, the estimate $\overline{X^{(N)}}$ is more accurate, the confidence interval will be shorter, and - what is more important - the statistic $\overline{X^{(N)}} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is more appealing and easier to calculate.

Furthermore $\mathrm{Var}(\overline{X^{(N)}}) \leq \mathrm{Var}(\overline{X^{(n_0)}})$ and hence $\mathrm{Var}(\overline{X^{(N)}})$ will tend to zero for n_0 tending to infinity, whereas $\mathrm{Var}(\widetilde{X^{(N)}})$ will stay bounded from below by c. A similar result is valid for the length of the confidence interval.

4 Straightforward applications to the Behrens-Fisher problem

In this section we present two procedures proposed by Chapman(1950) and Ghosh(1975) and supplement them with a reasonable modification of the latter one.

Let $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ be *iid* normally distributed with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively. All procedures considered here start with preliminary samples $(X_i, \ldots, X_{n_0})$ and $(Y_i, \ldots, Y_{n_0})$ of equal sample size n_0. The variances σ_1^2 and σ_2^2 will be estimated by the usual estimates based on the preliminary samples $\widehat{\sigma}_{1;0}^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(X_i - \overline{X^{(n_0)}})^2$ and $\widehat{\sigma}_{2;0}^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(Y_i - \overline{Y^{(n_0)}})^2$.

Chapman(1950) determines the sample sizes by $N_i = \max([\widehat{\sigma}_{i;0}^2/z] + 1, n_0 + 1)$ and calculates $\widetilde{X^{(N_1)}}$ and $\widetilde{Y^{(N_2)}}$ individually according to Stein's first procedure. Then the distribution of $\widetilde{X^{(N_1)}} - \widetilde{Y^{(N_2)}} - (\mu_1 - \mu_2)$ is that of the difference of two independent t-distributed random variables with $n_0 - 1$ degrees of freedom each. Hence $E(\widetilde{X^{(N_1)}} - \widetilde{Y^{(N_2)}}) = \mu_1 - \mu_2$ and $\mathrm{Var}(\widetilde{X^{(N_1)}} - \widetilde{Y^{(N_2)}}) = 2z\frac{n_0-1}{n_0-3}$. Letting $z = \frac{1}{2}\frac{n_0-3}{n_0-1}c$ we obtain an estimate with prescribed accuracy $\mathrm{Var}(\widetilde{X^{(N_1)}} - \widetilde{Y^{(N_2)}}) = c$. Similarly $z = d^2/\tau^2$ yields a confidence interval for $\mu_1 - \mu_2$ with bounds $\widetilde{X^{(N_1)}} - \widetilde{Y^{(N_2)}} \pm d$ of bounded length $2d$ with confidence level $1-\alpha$ when τ is the $(1-\alpha/2)$-quantile of the above mentioned distribution (for a table of these quantiles see Chapman(1950)).

Alternatively Ghosh(1975) considered matched pairs of observations. This procedure makes special use of the obvious fact that $(X_i - Y_i)_{i\in\mathbb{N}}$ are *iid* normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Based on the first stage of n_0 observations

$\widehat{\sigma}_0^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(X_i - Y_i - (\overline{X^{(n_0)}} - \overline{Y^{(n_0)}}))^2$ is an estimate of the variance $\sigma_1^2 + \sigma_2^2$ of the matched observations $X_i - Y_i$. Now choose equal sample sizes $N_1 = N_2 = \max([\widehat{\sigma_0^2}/z]+1, n_0)$. The results concerning Stein's second procedure can be applied directly and we obtain $E(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) = \mu_1 - \mu_2$ for $n_0 \geq 3$, $\mathrm{Var}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) \leq z\frac{n_0-1}{n_0-3}$ for $n_0 \geq 4$, and

$$P(|(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2)| \leq t) \geq P(|T_{n_0-1}| \leq \tfrac{d}{\sqrt{z}}) .$$

By letting $z = \frac{n_0-3}{n_0-1}c$ and $z = d^2/t^2_{n_0-1;1-\alpha/2}$ we get again an estimate of $\mu_1 - \mu_2$ of prescribed accuracy $\mathrm{Var}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) \leq c$ or a confidence interval for $\mu_1 - \mu_2$ with bounds $\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} \pm d\sqrt{\widehat{\sigma}_0^2/(Nz)}$ of bounded length $2d$ with confidence level $1-\alpha$ respectively.

Ghosh(1975) compared these two procedures with respect to the task of generating a bounded length confidence interval with given confidence level and proposed the advantage of his procedure in many but not all situations. In particular Chapman's(1950) procedure suffers from the same drawbacks as Stein's first procedure compared to the second.

We will show, however, that the expected total sample sizes $N_1 + N_2$ needed to generate an estimate for $\mu_1 - \mu_2$ of prescribed accuracy are approximately the same for both methods. Just to give ideas we make crude approximations neglecting the influence of the size of the first stage. Then for Chapman's procedure we obtain

$$E(N_1 + N_2) \approx \tfrac{1}{z}(E(\widehat{\sigma}^2_{1;0}) + E(\widehat{\sigma}^2_{2;0})) = \tfrac{2}{c}\tfrac{n_0-1}{n_0-3}(\sigma_1^2 + \sigma_2^2)$$

and the same result for Ghosh's procedure

$$E(N_1 + N_2) \approx 2\tfrac{1}{z}E(\widehat{\sigma}_0^2) = \tfrac{2}{c}\tfrac{n_0-1}{n_0-3}(\sigma_1^2 + \sigma_2^2).$$

Since $N_1 = N_2$ in Ghosh's and N_1/N_2 approximately equal to σ_1^2/σ_2^2 in Chapman's procedure this result looks natural in view of the considerations following formula (1). However, a more careful investigation will probably show that Ghosh's procedure will be slightly more advantageous.

Unfortunately Ghosh's procedure has the unpleasant property that the estimate $\widehat{\sigma}_0^2$ of $\sigma_1^2 + \sigma_2^2$, and hence the sample size, heavily depends on the ordering of the observations. In particular, two different randomizations of the observations in the first stage may yield two totally different total sample sizes - a property which is very undesirable in practical applications (see Scheffé's(1970) comment on *"An Impractical Solution"*).

We now modify Ghosh's procedure in such a way that it is independent of any ordering of the observations. So if we choose $N_1 = N_2 = \max([(\widehat{\sigma}^2_{1;0} + \widehat{\sigma}^2_{2;0})/z] + 1, n_0)$, then $E(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) = \mu_1 - \mu_2$ for $n_0 \geq 3$, and we obtain as in Theorem 1 that the Behrens-Fisher statistic is less dispersed than a t-distribution with $n_0 - 1$ degrees of freedom (in particular, this means that the Behrens-Fisher statistic is less dispersed for the modified than for Ghosh's orginal procedure):

Theorem 2

With $N_1 = N_2$ chosen according to the modified rule we have for all t:

$$P\left(\left|\frac{(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2))}{\sqrt{\frac{1}{N_1}\widehat{\sigma}^2_{1;0} + \frac{1}{N_2}\widehat{\sigma}^2_{2;0}}}\right| \geq t\right) \leq P(|T_{n_0-1}| \geq t).$$

□

Theorem 2 can be derived from the result of Mickey and Brown(1966) because $\sqrt{N_1}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2 + \sigma_2^2}$ is standard normal and independent of $\widehat{\sigma}_{i;0}^2, i = 1, 2$, and the present Behrens-Fisher statistic is of their form V_γ with $\gamma = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$. By Theorem 2 we obtain

$$\mathrm{Var}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) \leq \mathrm{Var}\left(\frac{\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{N_1}\widehat{\sigma}_{1;0}^2 + \frac{1}{N_2}\widehat{\sigma}_{2;0}^2}} \cdot \sqrt{z}\right) \leq z\frac{n_0 - 1}{n_0 - 3},$$

for $n_0 \geq 4$, and

$$P(|\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2)| \leq d) \geq 2P(|T_{n_0-1}| \leq \tfrac{d}{\sqrt{z}}),$$

such that the procedure performs at least as good as that of Ghosh(1975). In particular the same values of $z = \frac{n_0 - 3}{n_0 - 1}c$ resp. $z = d^2/t_{n_0 1;1-\alpha/2}^2$ will give an estimate of prescribed accuracy and a confidence interval of bounded length.

If we calculate the expected total sample size for generating an estimate of prescribed accuracy with the help of the present modified procedure approximately we obtain the same value as before:

$$E(N_1 + N_2) \approx 2\tfrac{1}{z}(E(\widehat{\sigma}_{1;0}^2 + \widehat{\sigma}_{2;0}^2)) = \tfrac{2}{c}\tfrac{n_0 - 1}{n_0 - 3}(\sigma_1^2 + \sigma_2^2),$$

but the variance

$$\mathrm{Var}(N_1 + N_2) \approx 4\tfrac{1}{z^2}(\mathrm{Var}(\widehat{\sigma}_{1;0}^2) + \mathrm{Var}(\widehat{\sigma}_{2;0}^2)) = \tfrac{8}{c^2}\tfrac{n_0 - 1}{(n_0 - 3)^2}(\sigma_1^4 + \sigma_2^4)$$

is reasonably smaller in the present case than $4\frac{1}{z^2}\mathrm{Var}(\widehat{\sigma}_0^2) = \frac{8}{c^2}\frac{n_0 - 1}{(n_0 - 3)^2}(\sigma_1^2 + \sigma_2^2)^2$ which is the variance for Ghosh's procedure. This indicates additional advantages of the modified procedure.

5 "Optimum" allocation

The considerations on optimum allocation made in section 1 suggest that the sample sizes N_1 and N_2 should be chosen in such a way that their ratio N_1/N_2 is close to $\sqrt{\varrho} = \sigma_1/\sigma_2$. Such a procedure has been proposed by Banerjee(1967):

Choose $N_i = \max([\widehat{\sigma}_{i;0}(\widehat{\sigma}_{1;0} + \widehat{\sigma}_{2;0})/z] + 1, n_0)$, where $\widehat{\sigma}_{i;0} = \sqrt{\widehat{\sigma}_{i;0}^2}$ is the estimate of σ_i based on the observations $(X_1, \ldots, X_{n_0})$ and $(Y_1, \ldots, Y_{n_0})$ resp. of the first stage $(i = 1, 2)$. The methods used to show the results of Theorems 1 and 2 do not carry over directly to the present situation. However, with a refinement of the arguments used, Banerjee(1967) could obtain an upper bound for the dispersion probabilities:

Theorem 3

With N_1, N_2 chosen according to the rule of Banerjee we have for all t:

$$\begin{aligned} &P(|\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2)| \geq t) \\ &\leq \; 2P\left(|T_{n_0-1}| \geq \frac{t}{\sqrt{z}}\right) - P\left(|T_{n_0}| \geq \sqrt{\frac{n_0}{n_0 - 1}} \cdot \frac{t}{\sqrt{z}}\right). \end{aligned}$$

For a proof we refer to the results of Banerjee(1967). By symmetry $E(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) = \mu_1 - \mu_2$ for $n_0 \geq 3$, and (since $\frac{1}{N_1}\widehat{\sigma}^2_{1;0} + \frac{1}{N_2}\widehat{\sigma}^2_{2;0} \leq z$) :

$$\begin{aligned}
\mathrm{Var}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) &= \int_0^\infty P(((\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) - (\mu_1 - \mu_2))^2 \geq s)ds \\
&\leq 2\int_0^\infty P(T^2_{n_0-1} \geq \tfrac{s}{z})ds - \int_0^\infty P(T^2_{n_0} \geq \tfrac{n_0}{n_0-1}\tfrac{s}{z})ds \\
&= 2z\mathrm{Var}(T_{n_0-1}) - z\tfrac{n_0-1}{n_0}\mathrm{Var}(T_{n_0}) \\
&= z(2\tfrac{n_0-1}{n_0-3} - \tfrac{n_0-1}{n_0}\cdot\tfrac{n_0}{n_0-2}) \\
&= z\tfrac{(n_0-1)^2}{(n_0-2)(n_0-3)},
\end{aligned}$$

for $n_0 \geq 4$. Letting $z = \frac{(n_0-2)(n_0-3)}{(n_0-1)^2}c$ we can obtain an estimate for $\mu_1 - \mu_2$ of prescribed accuracy $\mathrm{Var}(\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}) \leq c$. With the same crude approximations as in the previous section we get

$$\begin{aligned}
E(N_1 + N_2) &\approx \tfrac{1}{z}E((\widehat{\sigma}_{1;0} + \widehat{\sigma}_{2;0})^2) \\
&= \tfrac{1}{c}\tfrac{(n_0-1)^2}{(n_0-2)(n_0-3)}(\sigma_1^2 + \sigma_2^2 + 2E(\widehat{\sigma}_{1;0})E(\widehat{\sigma}_{2;0})) \\
&= \tfrac{1}{c}\tfrac{(n_0-1)^2}{(n_0-2)(n_0-3)}\left(\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\tfrac{2}{(n_0-1)}\Gamma(\tfrac{n_0}{2})^2/\Gamma(\tfrac{n_0-1}{2})^2\right).
\end{aligned}$$

In Figure 1 we exhibit the approximate expected total sample size $E(N_1 + N_2)$ compared to that of the modified procedure of the previous section $\frac{2}{c}\frac{n_0-1}{n_0-3}(\sigma_1^2 + \sigma_2^2)$ in dependence on the ratio $\sqrt{\varrho} = \sigma_1/\sigma_2$ of the standard deviations.

As equal sample sizes are preferable in case of equal variances, it is not astonishing that Figure 1 exhibits a better performance of the modified procedure of section 4 with respect to the expected total sample size for $\varrho = \sigma_1^2/\sigma_2^2$ close to 1, whereas for considerably differing variances the present procedure manages with smaller sample sizes.

The same behaviour can be observed when generating confidence intervals of bounded length. Denote by $b_{n_0;\alpha}$ the "critical value" such that

$$2P(|T_{n_0-1}| \geq b_{n_0;\alpha}) - P(|T_{n_0}| \geq \sqrt{\tfrac{n_0}{n_0-1}}b_{n_0;\alpha}) = \alpha.$$

Then by setting $z = d^2/b^2_{n_0;\alpha}$ we obtain

$$P(|\overline{X^{(N_1)}} - \overline{Y^{(N_2)}} - (\mu_1 - \mu_2)| \leq d) \geq 2P(|T_{n_0-1}| \leq \tfrac{d}{\sqrt{z}}) - P(|T_{n_0}| \leq \sqrt{\tfrac{n_0}{n_0-1}}\tfrac{d}{\sqrt{z}}) = 1 - \alpha$$

TABLE 1: Critical values $b_{n_0;\alpha}$ for Banerjee's procedure ($\alpha = 0.01, 0.05, 0.10$)

n_0		2	3	4	5	6	7	8	9	10
$b_{n_0;\alpha}$	$\alpha = 0.01$	126.93	13.78	7.17	5.33	4.52	4.07	3.78	3.59	3.45
	$\alpha = 0.05$	25.05	5.88	3.87	3.20	2.87	2.68	2.55	2.46	2.40
	$\alpha = 0.10$	12.30	3.95	2.85	2.45	2.25	2.13	2.04	1.99	1.94
n_0		15	20	25	30	35	40	45	50	∞
$b_{n_0;\alpha}$	$\alpha = 0.01$	3.09	2.94	2.86	2.80	2.77	2.74	2.72	2.71	2.58
	$\alpha = 0.05$	2.23	2.15	2.11	2.08	2.06	2.05	2.04	2.03	1.96
	$\alpha = 0.10$	1.83	1.78	1.75	1.73	1.72	1.71	1.70	1.69	1.65

and hence a confidence interval of bounded length $2d$. Some values of $b_{n_0;\alpha}$ are given in Table 1.

In Table 2 we present the approximated bounds ϱ^* of the regions of preference with respect to the expected total sample sizes for the modified resp. Ghosh's procedure of section 4 on one hand side and Banerjee's present procedure on the other hand side, i.e. Banerjee's procedure is preferable if $\sigma_1^2/\sigma_2^2 > \varrho^*$ or $\sigma_1^2/\sigma_2^2 < 1/\varrho^*$.

6 Concluding remarks

As has been discussed exhaustively in he previous section neither Banerjee's procedure with adjusted sample sizes $N_1/N_2 \approx \hat{\sigma}_1/\hat{\sigma}_2$ nor the modified procedure of section 4 having equal sample sizes can be considered to be more preferable uniformly in $\varrho = \sigma_1^2/\sigma_2^2$.

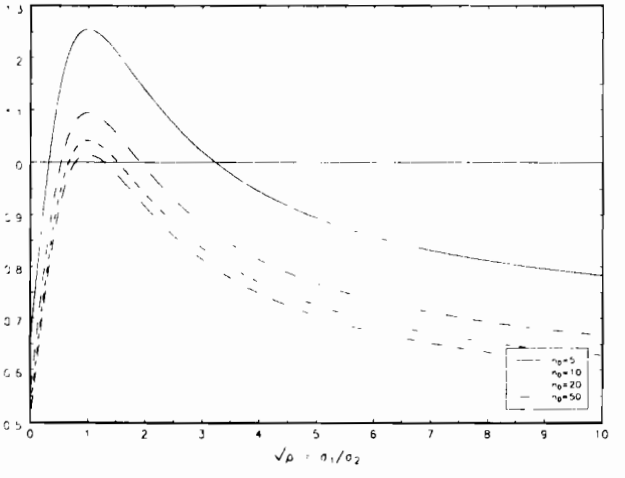


FIGURE 1: Ratio of the approximate expected sample sizes of Banerjee's procedure and the modified procedure of section 4

TABLE 2: Bounds ϱ^* of the region of preference for the modified procedure:
(1) estimate of prescribed accuracy;
(2) bounded length confidence intervals with confidence level $1-\alpha$ ($\alpha = 0.01, 0.05, 0.10$)

n_0		2	3	4	5	10	20	30	40	50	100
(1)		—	—	23.90	10.40	3.64	2.31	1.94	1.77	1.66	1.42
(2)	$\alpha = 0.01$	∞	1728.78	25.15	10.92	3.69	2.31	1.94	1.77	1.66	1.43
	$\alpha = 0.05$	∞	467.57	21.06	10.00	3.63	2.31	1.94	1.76	1.66	1.42
	$\alpha = 0.10$	∞	300.90	19.79	9.76	3.62	2.31	1.95	1.77	1.67	1.43

However, if there is no indication that the variances are far from equality the modified procedure of section 4 shows some advantages and, in particular, it only involves the known quantiles of the t-distribution. To perform a more detailed comparison more thorough investigations have to be made.

A further topic which will only be touched here is the question how to distribute the observations among the two groups for a given fixed sample size N. Again crude approximations show that for $N_1/N_2 \approx \hat{\sigma}_{1;0}/\hat{\sigma}_{2;0}$ the variance of $\overline{X^{(N_1)}} - \overline{Y^{(N_2)}}$ is approximately equal to

$\frac{1}{N}(\sigma_1^2+\sigma_2^2+2\sigma_1\sigma_2\Gamma(\frac{n_0}{2})\Gamma(\frac{n_0}{2}-1)/\Gamma(\frac{n_0-1}{2})^2)$ which in most cases is substantially smaller than the variance $\frac{2}{N}(\sigma_1^2+\sigma_2^2)$ obtained for equal sample sizes. This indicates that experimental design and sequential approach both considered as tools for improving the performance of an experiment can and should be successfully combined.

References

Bandemer, H. (Ed.) (1977). *Theorie and Anwendung der optimalen Versuchsplanung I. Handbuch zur Theorie.* Akademie-Verlag, Berlin.

Banerjee, S. (1967). Confidence interval of preassigned length for the Behrens-Fisher problem. *Ann. Math. Statist.* **38**, 1175-1179.

Chapman, D. G. (1950). Some two sample tests. *Ann. Math. Statist.* **21**, 601-606.

Chatterjee, S. K. (1991). Two-stage and multistage procedures. In: B. K. Ghosh and P. K. Sen (Eds.): *Handbook of Sequential Analysis.* Marcel Dekker, New York, 21-45.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

Ghosh, B. K. (1975). A two-stage procedure for the Behrens-Fisher problem. *J. Amer. Statist. Ass.* **70**, 457-462.

Mickey, M. R. and M. B. Brown (1966). Bounds on the distribution functions of the Behrens-Fisher statistic. *Ann. Math. Statist.* **37**, 639-642.

Pázman, A. (1986). *Foundations of Optimum Experimental Design.* Reidel, Dordrecht.

Scheffé, H. (1970). Practical solutions to the Behrens-Fisher problem. *J. Amer. Statist. Ass.* **65**, 1501-1508.

Silvey, S. D. (1980). *Optimal Design.* Chapman and Hall, London.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16**, 243-258.

Stuart, A. and J. K. Ord (1991). *Kendall's Advanced Theory of Statistics, Vol. 2, 5th Ed.* Edward Arnold, London.

A Useful Set of Multiple Orthogonal Polynomials on the q-Simplex and its Application to D-optimal Designs

Ralf-Dieter Hilgers

"[Wenn] wir aus viel hundert Stoffen
Durch Mischung - denn auf Mischung kommt es an -
Den [Stoff] gemächlich komponieren,[. . .]
So ist das Werk im stillen abgetan." Goethe, Faust II.

D-optimality of a design can be proven by the equivalence to G-optimality. In this context the calculation of the maximal variance of the prediction can be simplified by transforming the set of regression functions to a set of orthogonal regression functions with respect to a design. Given regression functions of a general type, an orthogonal regression function set is calculated. The properties of this set is investigated and applied to construct D-optimal designs on the q-simplex.

1 Introduction

In a standard (linear) regression model $\eta(\mathbf{x}) = \sum_{r=1}^{m} \vartheta_r f_r(\mathbf{x})$, the $f_1, \ldots, f_m$ are continuous, real valued, linear independent regression functions on a compact set $\mathcal{X}$ and $\vartheta_1, \ldots, \vartheta_m$ are m unknown parameters. The outcome $\eta(\mathbf{x})$ of an experiment $\mathbf{x} \in \mathcal{X}$ is considered as a realization of the random variable $\mathcal{Y}$ with mean $\sum_{r=1}^{m} \vartheta_r f_r(\mathbf{x})$ and variance σ^2 independent of $\mathbf{x}$. Here the observations in different experimental points are assumed to be uncorrelated. To estimate the coefficient vector $\vartheta = (\vartheta_1, \ldots, \vartheta_m)^t$ observations at various values of the experimental point $\mathbf{x}$ are taken. The generalization of the proportions of observations in the pairwise different experimental points to probabilities leads to the (discrete) probability measures ξ on $\mathcal{X}$ called designs.
Designs ξ are classified by concave functions of the information matrix $\mathbf{M}(\mathbf{f};\xi) = \int \mathbf{f}\,\mathbf{f}^t d\xi$. In particular maximizing the determinant subject to the design is called D-optimality. It is well known from Kiefer and Wolfowitz (1960) that D-optimality is equivalent to G–optimality, i.e. minimizing the maximum variance of the prediction. Moreover a design ξ^* is G–optimal if for all $\mathbf{x} \in \mathcal{X}$

$$\mathbf{f}^t(\mathbf{x})\,\mathbf{M}^{-1}(\mathbf{f};\xi^*)\,\mathbf{f}(\mathbf{x}) = d(\mathbf{x},\xi^*) \leq m \tag{1}$$

is satisfied. Notice, that ξ^* can have points of support only where $d(\mathbf{x},\xi^*) = m$.

Often the calculation of $d(\mathbf{x},\xi^*)$ is an unpleasant problem. However, the generalized variance is invariant under regular transformation, c.f. Fedorov (1972).
Assume the design ξ^* to investigate has *minimal support*. This means the number of supporting points is equal to the number of parameters in the regression. Then a transformation of the original regression functions $\mathbf{f}$ by a regular matrix $\mathbf{A}$ can be found, such that for $\mathbf{g} = \mathbf{A}\mathbf{f}$,

$$\mathbf{M}(\mathbf{g};\xi^*) = \int \mathbf{g}\,\mathbf{g}^t d\xi^* = \frac{1}{m}\mathbf{I}_m\,.$$

With these transferred regression functions $\mathbf{g}^t(\mathbf{x}) = \big(g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})\big)$ (1) becomes

$$\mathbf{g}^t(\mathbf{x})\mathbf{g}(\mathbf{x}) \leq 1 \,. \tag{2}$$

The regression functions $g_1, \ldots, g_m$ are called *orthogonal* with respect to the support $\mathrm{supp}(\xi^*) = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ of ξ^*, because they fulfill

$$g_i(\mathbf{z}_j) = \begin{cases} 1, & \text{for } i = j, \\ 0, & \text{else} \,. \end{cases}$$

It is well known, that a D-optimal design with minimal support is equal weighting.
In the following regression models with and without intercept are considered separately. Look at the model

$$\eta(\mathbf{x}) = \sum_{\ell=1}^{\nu} \sum_{1 \leq s_1 < \cdots < s_\ell \leq q} \vartheta_{s_1 \cdots s_\ell} h_{s_1 \cdots s_\ell}(x_{s_1}, \ldots, x_{s_\ell}) \tag{3}$$

defined on the factor space $\mathcal{X} \subset \mathbb{R}^q$. This model will be called *generalized multiple ν-tic polynomial.* Extending the regression model (3) by an intercept term the *extended generalized multiple ν-tic polynomial*

$$\eta(\mathbf{x}) = \vartheta_0 + \sum_{\ell=1}^{\nu} \sum_{1 \leq s_1 < \cdots < s_\ell \leq q} \vartheta_{s_1 \cdots s_\ell} h_{s_1 \cdots s_\ell}(x_{s_1}, \ldots, x_{s_\ell}) \tag{4}$$

defined on the factor space $\mathcal{X} \subset \mathbb{R}^q$ is obtained.

If the design to examine has minimal support and the regression functions $h_{s_1 \cdots s_\ell}$ satisfy some moderate restrictions a general set of orthogonal regression functions can be found. In section 2.1 the generalized multiple ν-tic polynomial (3) is investigated while in section 2.2 the model (4) with intercept is treated. Properties of the obtained orthogonal regression functions are discussed under more specific assumptions. Further on some applications are given to derive results in section 3 followed by some new applications for D-optimal designs on the q-dimensional simplex.

2 Orthogonal regression functions

2.1 Regression model without intercept

2.1.1 Supporting points

First let us assume that the points of support to estimate the coefficients of a regression model without intercept (3) are of the following structure:

Given $\nu \leq q$ non vanishing, real numbers $z_1, \ldots, z_\nu$. Then the support consists of the points

$$\mathcal{A} = \left\{ z_k \sum_{r=1}^{k} \mathbf{e}_{u_r} \,, 1 \leq u_1 < \cdots < u_k \leq q \,, 1 \leq k \leq \nu \right\} . \qquad (5)$$

Here $\mathbf{e}_r$ denotes the r-th unitvector of $\mathbb{R}^q$, so that $\mathbf{e}_1^t = (1, 0, \ldots, 0); \mathbf{e}_2^t = (0, 1, 0, \ldots, 0)$ and so on.
A point $\mathbf{z}_k = z_k \sum_{r=1}^{k} \mathbf{e}_{u_r}$ is called a *generalized barycenter of depth k*.

Example: Choosing $z_k = 1/k$ then $\mathcal{A}$ is the set of all barycenters up to depth ν of the unit q-simplex. The uniform weighting design with this support is called *simplex centroid design (of depth ν)*, c.f. Scheffé (1963).

2.1.2 A set of orthogonal regression functions without intercept

Let us start consideration with some normalization conditions of the regression functions $h_{s_1 \ldots s_\ell}(\mathbf{x})$ of model (3). Assume, that

(6a) each regression function $h_{s_1 \ldots s_\ell}(\mathbf{x})$ contains no parameters,

(6b) each regression function $h_{s_1 \ldots s_\ell}(\mathbf{x})$ is only a function of the component x_{s_1} to x_{s_ℓ} of $\mathbf{x} = (x_1, \ldots, x_q)^t$,

(6c) each regression function $h_{s_1 \ldots s_\ell}(\mathbf{x})$ becomes zero whenever one of its variables—c.f. (6b)—is zero,

(6d) each regression function vanishes on a generalized barycenter $\mathbf{z}_k \in \mathcal{A}$ if one of its variables is zero (c.f. 6c) or $h_{s_1 \ldots s_\ell}$ becomes

$$h_{s_1 \ldots s_\ell}(\mathbf{z}_k) = h_{s_1 \ldots s_\ell}\left(z_k \sum_{r=1}^{k} \mathbf{e}_{u_r} \right) = c_\ell(k) \neq 0 \,.$$

Remarks:
i) A regression function $h_{s_1 \ldots s_\ell}$ of ℓ components is called a *regression function of order ℓ*. If the regression model includes all regression functions up to order $\nu < q$ then it will be called *saturated* while it is called *fully saturated*, if it contains all regression functions up to order q.
ii) Condition (6a) is a natural assumption in linear models and like (6b) not very restrictive.
iii) From (6c) it follows, that $h_{s_1 \ldots s_\ell}(\mathbf{z}_k) = 0$ if the order ℓ of the function is higher than the depth k of the generalized barycenter (for all $\mathbf{z}_k \in \mathcal{A}$, where $\mathbf{z}_k$ is a generalized barycenter of depth $k < \ell$). The function also vanishes if its index set is not contained in the index set

defining the generalized barycenter, i.e. $\{s_1,\ldots,s_\ell\} \not\subseteq \{u_1,\ldots,u_k\}$.
iv) If, e.g., the regression function $h_{s_1\cdots s_\ell}(\mathbf{x})$ describes a kind of interaction between the variables x_{s_1} to x_{s_ℓ} it might be justified to accept (6c). The most restrictive condition seems to be the symmetry condition (6d). This means, that a regression function of the components x_{s_1} to x_{s_ℓ} of order ℓ becomes a non vanishing constant $c_\ell(k)$ at all generalized barycenters of depth k with indices $\{u_1,\ldots,u_k\} \supseteq \{s_1,\ldots,s_\ell\}$. Let us look for example at the choices

$$h_{s_1\cdots s_\ell}(\mathbf{x}) = \begin{cases} \min\{x_{s_1},\ldots,x_{s_\ell}\} \\ \max\{x_{s_1},\ldots,x_{s_\ell}\} \\ \left(x_{s_1}^{t_{s_1}}\cdots x_{s_\ell}^{t_{s_\ell}}\right)^w \end{cases} \qquad t_r > 0, w > 0.$$

Then (6a,b) are satisfied by all these regression functions. The first regression function in addition fulfills (6c) and (6d). However, the second regression function $h_{s_1\cdots s_\ell}(\mathbf{x}) = \max\{x_{s_1},\ldots,x_{s_\ell}\}$ is zero only if the largest component is zero. So in general (6c) is not valid. On the other hand, the third regression function fulfills (6c). But (6d) holds only, if $\ell = \nu = 1$ and $z_1 = 1$ or if $t_1 = \cdots = t_q$. Taking $t_1 = \cdots = t_q = w = 1$, the resulting regression model is the tic polynomial of Scheffé (1963).

Theorem 1: *Let the set of linear independent regression functions be given by*

$$\left\{h_{s_1\cdots s_\ell}(\mathbf{x})\,, 1 \le s_1 < \cdots < s_\ell \le q\,, 1 \le \ell \le \nu\right\}$$

defined on a compact set $\mathcal{X} \subset \mathbb{R}^q$, satisfying (6a) to (6d). Let the supporting points of interest be given by $\mathcal{A} \subset \mathcal{X}$. Then the set of regression functions

$$\left\{g_{s_1\cdots s_\ell}(\mathbf{x}), 1 \le s_1 < \cdots < s_\ell \le q, 1 \le \ell \le \nu\right\} \tag{7a}$$

where

$$g_{s_1\cdots s_\ell}(\mathbf{x}) = a_{0,\ell}\left[h_{s_1\cdots s_\ell}(\mathbf{x}) + \sum_{r=1}^{\nu-\ell}(-1)^r a_{r,\ell} \sum_{\substack{t_1,\ldots,t_r \in I^c \\ 1\le t_1<\cdots<t_r\le q}} h_{s_1\cdots s_\ell,t_1\cdots t_r}(\mathbf{x})\right] \tag{7b}$$

*with $I^c = \{1,\ldots,q\} \setminus \{s_1,\ldots,s_\ell\}$, $1 \le s_1 < \cdots < s_\ell \le q$ and $1 \le \ell \le \nu$ is orthogonal on $\mathcal{A}$. In (7b) the coefficients are given by**

$$a_{r,\ell} = \begin{cases} \frac{1}{c_\ell(\ell)}\,, & \text{if } r = 0,\ 1 \le \ell \le \nu, \\ \frac{(-1)^{r-1}}{c_{r+\ell}(r+\ell)}\left[c_\ell(r+\ell) + \sum_{u=1}^{r-1} a_{u,\ell}(-1)^u \binom{r}{u} c_{\ell+u}(r+\ell)\right], & \\ \qquad \text{if } r \in \{1,\ldots,\nu-\ell\},\ 1 \le \ell \le \nu-1\,. \end{cases} \tag{7c}$$

Proof: To show the orthogonality on $\mathcal{A}$ consider for a given index set $\{u_1,\ldots,u_k\} \subseteq \{1,\ldots,q\}$ with $k \le \nu$ the point

$$\mathbf{z}_k = z_k \sum_{r=1}^{k} \mathbf{e}_{u_r} \in \mathcal{A}\,.$$

* *The summation over no entries like $\sum_{r=1}^{0}\cdots$ is defined as zero.*

If at least one of the components x_{s_1} to x_{s_ℓ} vanishes, then $g_{s_1 \cdots s_\ell}(\mathbf{x}) = 0$. So $g_{s_1 \cdots s_\ell}(\mathbf{z}_k) = 0$, if the generalized barycenter $\mathbf{z}_k$ is of depth k less than ℓ, or the set of indices $\{s_1, \ldots, s_\ell\}$ is not included in $\{u_1, \ldots, u_k\}$. Furthermore, if $k = \ell$ and the sets of indices—$\{s_1, \ldots, s_\ell\} = \{u_1, \ldots, u_\ell\}$—are equal, then the second sum in (7b) is zero by (6c), so that $g_{s_1 \cdots s_\ell}(\mathbf{z}_k) = a_{0,\ell} c_\ell(\ell) = 1$ by (7c). If $k > \ell$ and $\{s_1, \ldots, s_\ell\}$ is included in $\{u_1, \ldots, u_k\}$, then $\binom{k-\ell}{r}$ of the regression functions of order $(\ell + r), 1 \le r \le k - \ell$, in the second sum become $c_{\ell+r}(k)$. So

$$\sum_{r=1}^{\nu-\ell} (-1)^r a_{r,\ell} \sum_{\substack{t_1,\ldots,t_r \in I^c \\ 1 \le t_1 < \cdots < t_r \le q}} h_{s_1 \cdots s_\ell, t_1 \cdots t_r}(\mathbf{z}_k) = \sum_{r=1}^{k-\ell} (-1)^r a_{r,\ell} \binom{k-\ell}{r} c_{\ell+r}(k) ,$$

and by (7c) it holds

$$\begin{aligned} g_{s_1 \cdots s_\ell}(\mathbf{z}_k) &= \frac{1}{c_\ell(\ell)} \left[c_\ell(k) + \sum_{r=1}^{k-\ell} (-1)^r a_{r,\ell} \binom{k-\ell}{r} c_{\ell+r}(k) \right] \\ &= \frac{1}{c_\ell(\ell)} \left[c_\ell(k) + \sum_{r=1}^{k-\ell-1} (-1)^r a_{r,\ell} \binom{k-\ell}{r} c_{\ell+r}(k) + (-1)^{k-\ell} c_k(k) a_{k-\ell,\ell} \right] \\ &= \frac{1}{c_\ell(\ell)} \left[(-1)^{k-\ell-1} c_k(k) a_{k-\ell,\ell} + (-1)^{k-\ell} c_k(k) a_{k-\ell,\ell} \right] \quad = \quad 0 . \end{aligned}$$

■

With a special choice of the constants it follows:

Lemma 1: *Let in the theorem 1 the regression functions fulfill (6d) with*

$$c_\ell(k) = \left(\frac{1}{k} \right)^\ell , \quad \ell \le k \le \nu .$$

a) Then the coefficients (7c) are given by

$$a_{r,\ell} = \begin{cases} \ell^\ell , & \text{if } r = 0, 1 \le \ell \le \nu, \\ \ell^{r-1}(\ell + r) , & \text{if } 1 \le r \le \nu - \ell, 1 \le \ell \le \nu - 1 , \end{cases} \tag{8}$$

and the orthogonal regression functions sum to

$$\sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g_{s_1 \cdots s_\ell}(\mathbf{x}) = \sum_{r=1}^{q} h_r(\mathbf{x}). \tag{9}$$

b) If $4 \le \nu < q$ and there exists a barycenter $\mathbf{z}_{\nu+1}$ of depth $(\nu + 1)$ so that

$$c_\ell(\nu + 1) = \left(\frac{1}{\nu + 1} \right)^\ell , \quad 1 \le \ell \le \nu ,$$

then the following inequality holds true

$$\sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) > 1 . \tag{10}$$

Proof: To prove the form of the coefficients in part a) of the lemma notice that from (7c) it follows $a_{0,\ell} = \frac{1}{c_\ell(\ell)} = \ell^\ell$. Further assume $a_{r,\ell}$ is correct, then from (7c)

$$
\begin{aligned}
&a_{r+1,\ell} = \\
&\frac{(-1)^r}{c_{r+1+\ell}(r+1+\ell)}\left[c_\ell(r+1+\ell) + \sum_{u=1}^{r}\binom{r+1}{u}(-1)^u a_{u,\ell}\, c_{\ell+u}(r+1+\ell)\right] \\
&= (-1)^r\left[(r+1+\ell)^{r+1} + \sum_{u=1}^{r}\binom{r+1}{u}(-1)^u \ell^{u-1}(\ell+u)(r+1+\ell)^{r+1-u}\right] \\
&= (-1)^r\left[(r+1+\ell)^{r+1} + \sum_{u=1}^{r}\binom{r+1}{u}(-1)^u \ell^u (r+1+\ell)^{r+1-u}\right. \\
&\qquad\left. + \sum_{u=1}^{r}\binom{r+1}{u}(-1)^u u\ell^{u-1}(r+1+\ell)^{r+1-u}\right] \\
&= (-1)^r\left[(r+1)^{r+1} + (-1)^r\ell^{r+1} - (r+1)^{r+1} + (-1)^r(r+1)\ell^r\right] \\
&= \ell^r(r+1+\ell).
\end{aligned}
$$

To prove equation (9), note that each $g_{s_1\cdots s_\ell}(\mathbf{x})$ is a linear combination of the regression functions of order greater or equal to ℓ having indices s_1 to s_ℓ. So on the left side of (9) each of the regression functions of the same order is added equally frequently. Rearranging the sum subject to the type of regression functions leads to

$$
\sum_{\ell=1}^{\nu}\sum_{1\le s_1<\cdots<s_\ell\le q} g_{s_1\cdots s_\ell}(\mathbf{x}) = \sum_{\ell=1}^{\nu} B_\ell \sum_{1\le s_1<\cdots<s_\ell\le q} h_{s_1\cdots s_\ell}(\mathbf{x})
$$

where for $1 \le \ell \le \nu$

$$
B_\ell = a_{0,\ell} + \sum_{r=1}^{\ell-1}\binom{\ell}{r}(-1)^r a_{0,\ell-r}a_{r,\ell-r}\,.
$$

Obviously $B_1 = a_{0,1} = 1$. For $\ell > 1$

$$
B_\ell = \ell\Big[\sum_{r=1}^{\ell}\binom{\ell}{r}(-1)^{\ell-r}r^{\ell-1}\Big] = 0\,,
$$

because the expression in brackets is the ℓ-th difference of the function $f(x) = x^{\ell-1}$ in the points $0, 1, \ldots, \ell$.

To show (10) use the formulas derived for a barycenter of depth $k \ge \ell$ in the proof of theorem 1. It follows with $r + \ell = \nu + 1$ in (7c) and then using (8)

$$
\begin{aligned}
g_{s_1\cdots s_\ell}(\mathbf{z}_{\nu+1}) &= \frac{1}{c_\ell(\ell)}\left[c_\ell(\nu+1) + \sum_{u=1}^{\nu+1-\ell}(-1)^u a_{u,\ell}\binom{\nu+1-\ell}{u}c_{\ell+u}(\nu+1)\right] \\
&= \frac{1}{c_\ell(\ell)}(-1)^{\nu-\ell}c_{\nu+1}(\nu+1)\,a_{\nu+1-\ell,\ell} \\
&= (-1)^{\nu-\ell}\left(\frac{\ell}{\nu+1}\right)^\nu\,.
\end{aligned}
$$

As outlined by Atwood (1969), then the following inequality holds true

$$\sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) = \sum_{\ell=1}^{\nu} \binom{\nu+1}{\ell} \left(\frac{\ell}{\nu+1}\right)^{2\nu} > 1 .$$

■

Remark:
Notice that, using the equivalence theorem in the special form (2), equation (10) can be used to disprove D–optimality.

In the discussion above the regression model consists of all regression functions up to order ν, so the set of supporting points does. Generalizations according to the non saturated case—not all the regression functions of order ℓ are included in the model—can be constructed similarly, but need a more complex computation.

2.2 Regression model with intercept

2.2.1 Supporting points

To extend the concept introduced so far to regression models with intercept (4) let us assume now that the support is of the form:

Given the set $\mathcal{A}$ as defined by (5). Then the support consists of the points

$$\mathcal{A}_0 = \mathcal{A} \cup \{\mathbf{0}\} . \qquad (11)$$

Examples: Choosing $z_k = 1/k$ then $\mathcal{A}_0$ is the set of all barycenters up to depth ν of the q–simplex. The uniform weighting design with this support is called *extended simplex centroid design (of depth ν)*, c.f. Hilgers (1991), Hilgers and Bauer (1992).
Obviously choosing $z_k = 1$ for all k, and $\nu = q$, then $\mathcal{A}_0$ is the set of all vertices of the cube $\{\mathbf{x} \in \mathbb{R}^q : 0 \le x_r \le 1, 1 \le r \le q\}$.

2.2.2 A set of orthogonal regression functions with intercept

As in section 2.1.2 let the regression functions $h_{s_1 \cdots s_\ell}$ fulfill the normalization conditions (6a) to (6d).

Remark:
Notice that in section 2.1.2 the model equation has $\sum_{\ell=1}^{\nu} \binom{\nu}{\ell}$ parameters while in (4) it has $1 + \sum_{\ell=1}^{\nu} \binom{\nu}{\ell}$ parameters.

Theorem 2: *Let the set of linear independent regression functions be given by*

$$\{1; h_{s_1 \cdots s_\ell}(\mathbf{x}) , 1 \le s_1 < \cdots < s_\ell \le q , 1 \le \ell \le \nu\}$$

defined on a compact set $\mathcal{X} \subset \mathbb{R}^q$, *satisfying (6a) to (6d). Let the supporting points of interest be given by* $\mathcal{A}_0 \subset \mathcal{X}$. *Then the set of regression functions*

$$\left\{1 - \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g_{s_1 \cdots s_\ell}(\mathbf{x}); \right. \\ \left. g_{s_1 \cdots s_\ell}(\mathbf{x}), 1 \le s_1 < \cdots < s_\ell \le q, 1 \le \ell \le \nu \right\} \tag{12}$$

with $g_{s_1 \cdots s_\ell}$ *defined in (7b) and (7c) is orthogonal on* $\mathcal{A}_0$.

Proof: Notice that $\mathcal{A}_0 = \mathcal{A} \cup \{\mathbf{0}\}$ and $g_{s_1 \cdots s_\ell}$ are orthogonal to $\mathcal{A}$ as shown in theorem 1. Obviously $g_{s_1 \cdots s_\ell}(\mathbf{z}) = 0$ by (6c), if $\mathbf{z} = \mathbf{0}$. Otherwise the first regression function in (12) vanishes, if $\mathbf{z} \in \mathcal{A}$. On the other hand, if $\mathbf{z} = \mathbf{0}$ the first regression function in (12) becomes 1. ∎

With a special choice of the constants it follows:

Lemma 2: *Let in the theorem 2 the regression functions fulfill (6d) with*

$$c_\ell(k) = \left(\frac{1}{k}\right)^\ell, \quad \nu \ge k \ge \ell \,.$$

If $4 \le \nu < q$ *and there exists a barycenter* $\mathbf{z}_{\nu+1}$ *of depth* $(\nu+1)$ *so that*

$$c_\ell(\nu+1) = \left(\frac{1}{\nu+1}\right)^\ell, \quad 1 \le \ell \le \nu \,,$$

then

$$\sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) - 2\sum_{r=1}^{q} h_r(\mathbf{z}_{\nu+1}) + \Big(\sum_{r=1}^{q} h_r(\mathbf{z}_{\nu+1})\Big)^2 > 0 \,. \tag{13}$$

Proof: It is easy to see that (8) and (9) hold true for the regression functions $g_{s_1 \cdots s_\ell}(\mathbf{x})$. Adding a positive term to the left side of (10) and then applying (9) it follows (13)

$$\begin{aligned} 1 &< \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) \\ &\le \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) + \Big(1 - \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1})\Big)^2 \\ &= \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} g^2_{s_1 \cdots s_\ell}(\mathbf{z}_{\nu+1}) + 1 - 2\sum_{r=1}^{q} h_r(\mathbf{z}_{\nu+1}) + \Big(\sum_{r=1}^{q} h_r(\mathbf{z}_{\nu+1})\Big)^2 \,. \end{aligned} \tag{14}$$

∎

Remark:
Like in the case of the regression model without intercept (13) can be used to disprove D-optimality by the equivalence theorem 2 as can be seen by (14).

3 Applications

3.1 ν–tic polynomials

Scheffé (1963) introduced the regression model

$$\eta(\mathbf{x}) = \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} \vartheta_{s_1 \cdots s_\ell} x_{s_1} \cdots x_{s_\ell}$$

defined on the $(q-1)$ simplex $\mathcal{U}_q = \left\{ \mathbf{x} \in \mathbb{R}^q : 0 \le x_r \le 1,\ 1 \le r \le q,\ \sum_{r=1}^{q} x_r = 1 \right\}$ to describe mixture experiments. As an appropriate design he suggested the uniform weighted *simplex centroid design (of depth ν)*. The support of this design results by taking $z_k = \frac{1}{k}, 1 \le k \le \nu$, in (5) above. Further for $h_{s_1 \ldots s_\ell}(\mathbf{x}) = x_{s_1} \cdots x_{s_\ell}, 1 \le s_1 < \cdots < s_\ell \le q, 1 \le \ell \le \nu$, (6a,b,c) hold true and with constants $c_\ell(k) = (1/k)^\ell$ in (6d) the assumptions of lemma 1 are satisfied. Hence the set of orthogonal regression functions is given by $\{g_{s_1 \ldots s_\ell}(\mathbf{x}),\ 1 \le s_1 < \cdots < s_\ell \le q,\ 1 \le \ell \le \nu\}$ with (7b) and (8). This form is found by Atwood (1969). He disproved D–optimality for $4 \le \nu < q$—c.f. lemma 1—and proved D–optimality for $\nu = q$ by other arguments. For $\nu = 2$ Kiefer (1961) and $\nu = 3$ Uranisi (1964) proved the D–optimality by using the orthogonal regression functions.

If the regression model is extended by an intercept and defined on the q simplex $\mathcal{S}_q = \left\{ \mathbf{x} \in \mathbb{R}^q : 0 \le x_r \le 1,\ 1 \le r \le q,\ \sum_{r=1}^{q} x_r \le 1 \right\}$, Hilgers and Bauer (1992) found the extended simplex centroid design (of depth ν) to be D–optimal. This approach seems to be appropriate to describe mixture amount experiments, where the response does not only depend on the proportions but also on the total amount of the components.

3.2 Other examples

Becker (1968) introduced special regression models on $\mathcal{U}_q$. Hilgers (1991) extended these models to the mixture amount factor space $\mathcal{S}_q$. Call the model

$$\eta(\mathbf{x}) = \vartheta_0 + \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} \vartheta_{s_1 \cdots s_\ell} \min\left\{ x_{s_1}, \ldots, x_{s_\ell} \right\} \tag{15}$$

the *minimum model (of order ν)* and

$$\eta(\mathbf{x}) = \vartheta_0 + \sum_{\ell=1}^{\nu} \sum_{1 \le s_1 < \cdots < s_\ell \le q} \vartheta_{s_1 \cdots s_\ell} \sqrt[\ell]{x_{s_1} \cdots x_{s_\ell}} \tag{16}$$

the *root model (of order ν)* defined on $\mathcal{S}_q$. An appropriate design may be given again by the extended simplex centroid design. Obviously the regression functions fulfill (6a,b,c) and (6d) with $c_\ell(k) = 1/k$. Then by using theorem 2 to get the orthogonal polynomials D–optimality can be proved for model (15) with $\nu = q$ and non D–optimality can be proved for $2 \le \nu < q$ for models (15) and (16), c.f. Hilgers (1991). Analogous results hold true for the models without intercept on $\mathcal{U}_q$, c.f. Hilgers (1991). Further results are not known up to now, but research is going on to get more general conditions for D–optimality.

3.3 Quasi linear regression on S_q

In the following section models perhaps not of particular interest are treated. The results may help to understand more about behaviour of D–optimality by changing the regression functions.
If the dose variable in the linear relationship is transferred by a power transformation, the *extended quasi linear regression model (of power* $w_r, 1 \le r \le q$*)*

$$\eta(\mathbf{x}) = \vartheta_0 + \sum_{r=1}^{q} \vartheta_r x_r^{w_r} \tag{17}$$

defined on S_q results as an extension of Hilgers and Bauer (1992). Other transformations like taking logarithms of the doses are quite common in dose response relationships. In the following restrictions of the power w_r are established to guarantee D–optimality of the (extended) simplex centroid design, applying the results of section 2.

Theorem 3: *Let the regression be of the form (17) defined on* S_q*. If* $w_r \ge 1$ *for all* $1 \le r \le q$*, then the extended simplex centroid design is the unique D–optimal design to estimate the coefficients by their* BLUE.

Proof: Here $\nu = 1, \mathcal{A}_0 = \{0, \mathbf{e}_1, \ldots, \mathbf{e}_q\}$ and $c_\ell(k) = 1$ so the orthogonal regression functions are given by $\left\{1 - \sum_{r=1}^{q} x_r^{w_r};\ x_r^{w_r}, 1 \le r \le q\right\}$ applying theorem 2. Now with (2) D–optimality is equivalent to

$$\begin{aligned} 1 &\ge \left(1 - \sum_{r=1}^{q} x_r^{w_r}\right)^2 + \sum_{r=1}^{q} x_r^{2w_r} \\ &= 1 + \left(\sum_{r=1}^{q} x_r^{w_r}\right)^2 - \sum_{r=1}^{q} x_r^{w_r} + \sum_{r=1}^{q} x_r^{2w_r} - \sum_{r=1}^{q} x_r^{w_r}. \end{aligned}$$

This inequality holds true, because $w_r \ge 1$, so that $0 \le x_r^{w_r} \le x_r \le 1$ and

$$0 \le \left(\sum_{r=1}^{q} x_r^{w_r}\right)^2 \le \sum_{r=1}^{q} x_r^{w_r} \le \sum_{r=1}^{q} x_r \le 1.$$

∎

For the model without intercept a similar result holds true:

Theorem 4: *Let the regression be of the form*

$$\eta(\mathbf{x}) = \sum_{r=1}^{q} \vartheta_r x_r^{w_r}$$

defined on S_q*. If* $w_r \ge 1/2$ *for all* $1 \le r \le q$*, then the simplex centroid design (of depth 1) is the unique D–optimal design to estimate the coefficients by their* BLUE.

Proof: Like in the theorem above, D–optimality is equivalent to

$$1 \ge \sum_{r=1}^{q} x_r^{2w_r}, \quad \mathbf{x} \in S_q,$$

which holds true, if $w_r \geq 1/2$ for all $1 \leq r \leq q$. ∎

Remarks:
i) Theorem 4 holds true even if the factor space is changed to $\mathcal{U}_q$.
ii) Whereas the D–optimality result for regression functions of the form $\sqrt{x_r}$ holds true in theorem 4, it does not in theorem 3. This can be seen by inserting a barycenter of depth 2 $(\mathbf{e}_1 + \mathbf{e}_2)/2$ in the right side of (2) calculated in the proof of theorem 3.

$$\left(\sum_{r=1}^{q}\sqrt{x_r}\right)^2 + \sum_{r=1}^{q} x_r - 2\sum_{r=1}^{q}\sqrt{x_r} = \left(2\sqrt{\frac{1}{2}}\right)^2 + 2\frac{1}{2} - 2\left(\sqrt{\frac{1}{2}} + \sqrt{\frac{1}{2}}\right)$$
$$= 3 - 2\sqrt{2} > 0\,.$$

iii) Results for other values of w_r in theorem 3 and 4 are not known up to now.
iv) The regression functions $\sqrt{x_r}$ were used to model an extreme change in the response behaviour as the value of one or more components tends to a boundary of the simplex region, c.f. Draper and John (1977).
v) Hilgers (1991) considered the extended factor space

$$\mathcal{S}_q^{(\kappa)} = \left\{\mathbf{x} \in \mathbb{R}^q : 0 \leq x_1, \ldots, x_q \leq 1,\ 0 \leq \sum_{r=1}^{q} x_r^{\kappa} \leq 1\right\} \subset \mathbb{R}^q$$

and formulated analogous results. However, the condition for D–optimality in theorem 3 has to be modified to $w_r \geq \kappa > 0$ whereas in theorem 4 one has to choose $w_r \geq \kappa/2 > 0$ for all r. Let, e.g., $\kappa = 2$, then this factor space is the "positive subspace" of the unit ball $\mathcal{B}_q = \{\mathbf{x} \in \mathbb{R}^q, \sum_{r=1}^{q} x_r^2 \leq 1\}$. So some D–optimality results on $\mathcal{S}_q$ respectively $\mathcal{S}_q^{(2)}$ and $\mathcal{B}_q$ can be obtained. But these results may be of interest only from the theoretical point of view.

3.4 Quasi tic regression on $\mathcal{S}_q$

A first approach of extending the results of the previous section to higher degree of quasi multiple polynomials is given by the following theorem:

Theorem 5: *Let the regression be of the form*

$$\eta(\mathbf{x}) = \sum_{r=1}^{q} \vartheta_r x_r^w + \sum_{1 \leq r < s \leq q} \vartheta_{rs} x_r^w x_s^w\,,$$

defined on $\mathcal{S}_q$. Then the simplex centroid design (of depth 2) is D–optimal to estimate the coefficients by their BLUE *if $w \geq 1/2$.*

Proof: With $\nu = 2, \mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_q, (\mathbf{e}_r + \mathbf{e}_s)/2, 1 \leq r < s \leq q\}$ and (6a to d, 7a to c) the orthogonal polynomials are given by

$$\left(x_r^w\left[1 - \sum_{\substack{s \neq r \\ s=1}}^{q} (2x_s)^w\right], 1 \leq r \leq q;\ (2x_r 2x_s)^w, 1 \leq r < s \leq q\right)\,.$$

So with (3) the D–optimality is equivalent to

$$
\begin{aligned}
1 &\geq 2^{4w} \sum_{1\leq r<s\leq q} [x_r x_s]^{2w} + \sum_{r=1}^{q} x_r^{2w} \Big[1 - 2^w \sum_{\substack{s\neq r \\ s=1}}^{q} x_s^w\Big]^2 \\
&= \sum_{r=1}^{q} x_r^{2w} + \left(2^{4w-1} + 2^{2w}\right) \sum_{1\leq r\neq s\leq q} x_r^{2w} x_s^{2w} \\
&\qquad - 2^w \sum_{1\leq r\neq s\leq q} \left(x_r^w x_s^{2w} + x_r^{2w} x_s^w\right) + 2^{2w} \sum_{1\leq r\neq s\neq t\leq q} x_r^{2w} x_s^w x_t^w \\
&= \sum_{r=1}^{q} x_r^{2w} + \left(2^{2w} - 2\right) \sum_{1\leq r\neq s\leq q} x_r^w x_s^w \\
&\qquad - \sum_{1\leq r\neq s\leq q} x_r^w x_s^w \Big[2^{2w} - 2 - \left(2^{4w-1} + 2^{2w}\right) x_r^w x_s^w \\
&\qquad + 2^w (x_r^w + x_s^w) - 2^{2w} x_r^w \sum_{t\neq r,s} x_t^w\Big] .
\end{aligned}
$$

Assume now without loss of generality, that $x_r > 0$ for all r. Then it holds* :

$$
1 \geq \left(\sum_{r=1}^{q} x_r\right)^{2w} \geq \sum_{r=1}^{q} x_r^{2w} + [2^{2w} - 2] \sum_{1\leq r\neq s\leq q} (x_r x_s)^w .
$$

Furthermore it holds by arithmetic-geometric mean inequality

$$
x_r^w \sum_{t\neq r} x_t^w \leq \left(\frac{1}{2} \sum_{t=1}^{q} x_t^w\right)^2 \quad \leq \quad 2^{-2}
$$

and $(x_r x_s)^w \leq 2^{-2w}$ for $\mathbf{x} \in S_q$. With $x_r^w + x_s^w - 4x_r^w x_s^w \geq (x_r^w + x_s^w)^2 - 4x_r^w x_s^w \geq (x_r^w - x_s^w)^2 \geq 0$ the expression in brackets gives:

$$
\begin{aligned}
&2^{2w} - 2 - \left(2^{4w-1} + 2^{2w}\right) x_r^w x_s^w + 2^w (x_r^w + x_s^w) - 2^{2w} x_r^w \sum_{t\neq r,s} x_t^w \\
&= 2^{2w} - 2 - 2^{4w-1} x_r^w x_s^w + 2^w (x_r^w + x_s^w - 4x_r^w x_s^w) + 2^{w+2} x_r^w x_s^w - 2^{2w} x_r^w \sum_{t\neq r} x_t^w \\
&\geq 2^{2w} - 2 - 2^{2w-1} + 2^{-w+2} - 2^{2w-2} \quad \geq \quad 0 .
\end{aligned}
$$

■

---

* In the inequality the following identity is used, expanding the sum on the right side for $k_1 = m$ and $k_1 + k_2 = m$, c.f. Bronstein (1985, p. 106 formula 2.6).

$$
\left(\sum_{r=1}^{q} x_r\right)^m = \sum_{k_1+\cdots+k_q=m} \binom{m}{k_1, \ldots, k_q} x_1^{k_1} \cdots x_q^{k_q} \quad , x_r \neq 0, 0 \leq k_r \leq m, 1 \leq r \leq q
$$

Remarks:
i) Clearly the result of theorem 5 holds true if the factor space is reduced to $\mathcal{U}_q$.
ii) In Hilgers (1991) the same model is considered on the extended simplex $S_q^{(\kappa)}$. Two interesting points arise. First one has to modify the assumption concerning the w. On $S_q^{(\kappa)}$ the condition has to be of the form $w \geq \kappa$. The proof is more algebraically than the one outlined above. Secondly from D–optimality it follows $w \geq \kappa/2$ which can be proved by inserting a corresponding barycenter of depth 3. Further let us consider the support of the corresponding design. Here one chooses $z_1 = 1$ and $z_2 = 1/\sqrt[\kappa]{2}$ so that $\mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_q, (\mathbf{e}_r + \mathbf{e}_s)/\sqrt[\kappa]{2}, 1 \leq r < s \leq q\}$ and by theorem 1 with $c_\ell(k) = 1$ the corresponding orthogonal regression functions are given by

$$\left(x_r^w \left[1 - \sum_{\substack{s=1 \\ s \neq r}}^{q} \left(\sqrt[\kappa]{2} x_s \right)^w \right], 1 \leq r \leq q;\ \left(\sqrt[\kappa]{2} x_r \sqrt[\kappa]{2} x_s \right)^w, 1 \leq r < s \leq q \right) .$$

Unfortunately computational difficulties increase, so that D–optimality considerations are restricted to the described cases. No results for the model with intercept respectively for the model with different exponents are available up to now.

4 Concluding remarks

The presented results of section 2 are shown to be useful to prove D–optimality in some cases. Although the considerations of the application in section 3 are restricted to the simplex, there may be other applications as mentioned by some comments.

Important points of practical interest follow from the results of section 3. Once the researcher has chosen the appropriate model, the D–optimal design is fixed in advance. After performing the trials he only can estimate the coefficients of the regression model fixed in advance without loss in the efficiency of the design. The results of section 3, however, suggest, that he can choose between a class of regression models of the same kind and select the model with the best fit. These result may give a further justification to use D-optimal designs and are investigated further on.

Acknowledgement:

This paper partly includes results of the author's PhD thesis accepted by the University of Dortmund. The author is greatful to Prof. P. Bauer for useful discussions and the referee for his constructive criticism.

References

Atwood, C.L. (1969). Optimal and efficient designs of experiments. *Ann. Statist.* **40**, 1570-1602.

Becker, N.G. (1968). Models for the response of a mixture. *J. Roy. Statist. Soc. B* **30**, 349-358.

Bronstein, I.N. and K.A. Semendjajew (1985). Taschenbuch der Mathematik. Harri Deutsch, Thun und Frankfurt/ Main.

Draper, N.R. and St.R.C. John (1977). A mixture model with inverse terms. *Technometrics* **19**, 37-46.

Farrell, R.H., J. Kiefer and A. Walbran (1967). Optimum multivariate designs. *Proc. Fifth Berkeley Symp. on Math. Statist. Probab.* **Vol. 1**, 113-138.

Fedorov, V.V. (1972). Theory of optimal experiments. Academic Press, New York.

Hilgers, R.-D. (1991). Optimale Versuchsplanung in Mischungs-Mengen-Experimenten. Ph.D. Thesis, University of Dortmund.

Hilgers, R.-D. and P. Bauer (1992). Optimal designs for mixture amount experiments. *submitted to J. Statist. Plann. Inf.*

Kiefer, J. (1961). Optimum designs in regression problems II. *Ann. Math. Statist.* **32**, 298-325.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canad. J. Math.* **12**, 363-366.

Scheffé, H. (1958). Experiments with mixtures. *J. Roy. Statist. Soc. B* **20**, 344-360.

Scheffé, H. (1963). The simplex centroid design for experiments with mixtures. *J. Roy. Statist. Soc. B* **25**, 235-263.

Uranisi, H. (1964). Optimum design for the special cubic regression on the q-simplex. *Mathematical Reports* (General Ed. Dept., Kuyushu University), **1**, 7-12.

On Designs with a Non-Orthogonal Row-Column-Structure

Joachim Kunert

When searching for optimal row-column designs, it is a good strategy to try to find a design which is a balanced incomplete block design when columns are considered as blocks, and for which the treatments are orthogonal to rows. In the usual setting where each row intersects with each columm exactly once, the treatments are orthogonal to rows if and only if they are proportional, i.e. each treatment appears in each row equally often. In that case the orthogonality condition does not depend on how the treatments are distributed over the columns, see also Kurotschka (1978).
Recently, there has been considerable interest in a non-orthogonal row column setting, where not each row intersects with each column. It was observed by Steward and Bradley (1991) and others that then the orthogonality condition can be fulfilled by designs which are not proportional. We treat this orthogonality condition in detail and show that it can be fulfilled by designs which are clearly non-optimal in the model without column effects.
We also show that the usual two way block model for non-orthogonal row and column structures in general cannot be justified with the help of randomization arguments.

1 Introduction

Recently, there has been considerable interest in optimal design for a non-orthogonal row-column setting, where, for instance certain combinations of rows and columns are not possible, see Saharay (1986), Mukhopadhyay and Saharay (1988), Shah and Sinha (1990), Steward and Bradley (1991), Jacroux and Saha Ray (1991), Baksalary and Pukelsheim (1992) and Kunert (1990).
We assume that we have a fixed arrangement of n experimental units into p rows and u columns. Not all combinations of rows and columns need to have a unit. Assume the unit which is in row i and column j receives treatment r. Then we assume for the measurement y_{ij} on this unit that

$$y_{ij} = \tau_r + \alpha_i + \beta_j + e_{ij} \tag{1}$$

where τ_r is the effect of treatment r, α_i the effect of row i and β_j the effect of column j. The errors e_{ij} are assumed to be uncorrelated with constant variance σ^2 and expectation 0. Let P and U be the design matrices for rows and columns, respectively. Then each unit is exactly in one row and one column and hence $P1_p = U1_u = 1_n$, where 1_x is the x-vector of ones.
We assume the row and column structure to be fixed, *i.e.* P and U are fixed. The experimenter can only choose which of the t treatments he applies to which unit. A rule which determines which treatment is applied to which unit is called a design. The set of all designs with t treatments and the given P and U is called Ω_{tPU}. Each design d determines a treatment design matrix T_d, where $T_d 1_t = 1_n$.
For a matrix M, define M^+ as the Moore-Penrose pseudoinverse, $pr(M) = M(M'M)^+M'$ and $pr^\perp(M) = I_n - pr(M)$. We are interested in the estimation of the vector with entries $\tau_r - \sum \tau_s / t$. It is well-known that the covariance matrix of the best linear unbiased estimate for this vector in model (1) equals $\sigma^2 C_d^+$, where

$$C_d = T_d' pr^\perp([P, U]) T_d.$$

A design d^* for which there is a real number z, such that $C_{d^*} = z(I_t - \frac{1}{t}1_t 1_t')$ and for which tr C_{d^*} is maximal over Ω_{tPU} is universally optimal. See Kiefer (1975) for a definition and details.
It follows from the results of Kunert (1983) that

$$C_d \leq T_d' pr^{\perp}(U) T_d = \tilde{C}_d,$$

say,
with equality holding if and only if

$$T_d' pr^{\perp}(U) P = 0.$$

Obviously, this is equivalent to

$$T_d' P = T_d' U (U'U)^{+} U' P. \tag{2}$$

Note that $U'U$ is a $u \times u$ diagonal-matrix. The $i-th$ diagonal-element equals the number c_i of units in column i. If all columns have the same number of units c, say, then (2) transforms to

$$T_d' P = \frac{1}{c} T_d' U \; U' P.$$

We call $T_d' P$ the treatment $\times$ row incidence matrix, $T_d' P$ the treatment $\times$ column and $P'U$ the row $\times$ column incidence matrix. It is well-known, see *e.g.* Kunert (1983), that a design d^* for which

$$\tilde{C}_{d^*} = z(I_t - \frac{1}{t}1_t 1_t'), \tag{3i}$$

$$tr\tilde{C}_{d^*} = \max_{d \in \Omega_{tPU}} tr\tilde{C}_d \tag{3ii}$$

and

$$T_{d^*}' P = T_{d^*}' U (U'U)^{+} U' P \tag{3iii}$$

is universally optimal in model (1). If there is a design $f \in \Omega_{tPU}$ which is a balanced block design with columns as blocks, then this design fulfills conditions (3i) and (3ii). Kunert (1990) has shown that in a model with non-orthogonal row-column structure, there are such designs which additionally fulfill (3iii) but are *not* balanced block designs with rows as blocks.

2 An evaluation of the orthogonality condition

We want to consider one special instance in detail. We explore the same experimental situation as the one considered in Kunert (1990). For some numbers of rows, columns and treatments, we will determine the set of all designs to fulfill conditions (3i), (3ii) and (3iii). Before we do that, we make sure that the case of a non-orthogonal row-column structure is really different from the usual setting of an orthogonal structure: in the orthogonal case conditions (3i), (3ii) and (3iii) can only be fulfilled by designs where each treatment appears in each row equally often.
The row-column structure is orthogonal if each combination of rows and columns receives exactly 1 unit. It follows that $U'P = 1_u 1_p'$. Hence, the right hand side of condition (3iii) can be written as

$$\frac{1}{p}T_d'U\ 1_u1_p' = \frac{1}{p}T_d'1_n1_p' = \frac{1}{p}[r_{d1},...,r_{dt}]'1_p',$$

where r_{di} is the number of appearances of treatment i in the design, $1 \leq i \leq t$. As the $(i,j) - th$ entry of $T_d'P$ gives the number of appearances of treatment i in row j, it follows that (3iii) is fulfilled if and only if treatment i appears r_{di}/p times in each row.
As was pointed out by Kunert (1990) this needs not be the case in a non-orthogonal setting. In what follows we consider one special instance of such a setting. Assume, we have t rows and kt columns, $k \in \mathbf{N}$. However, each column has only $t-1$ units, such that there is no unit in row j of column $\ell t + j$, $0 \leq \ell \leq k-1$, $1 \leq j \leq t$. Then $U'P = 1_k \otimes (1_t1_t' - I_t)$, where '$\otimes$' denotes the Kronecker-product of matrices. The matrix $U'U$ equals $(t-1)I_u$. If $d \in \Omega_{tUP}$ for this setting is a balanced incomplete block design with columns as blocks then in each block there must be exactly $t-1$ treatments, *i.e.* one is missing. Additionally, each treatment must be missing in the same number of blocks, namely in k blocks. This means, however, that there must be a $kt \times kt$ permutation matrix Π_d, such that $T_d'U = \{1_k' \otimes (1_t1_t' - I_t)\}\Pi_d$. Consequently, the right-hand side of (3iii) can be written as

$$\frac{1}{t-1}\{1_k' \otimes (1_t1_t' - I_t)\}\Pi_d\{1_k \otimes (1_t1_t' - I_t)\}.$$

Since $1_k \otimes 1_t$ is permutation-invariant, this equals

$$\frac{1}{t-1}\{k(t-2)1_t1_t' + (1_k' \otimes I_t)\Pi_d(1_k \otimes I_t)\}.$$

Now, write $\Pi_d = [\Pi_{ij}^{(d)}]_{1 \leq i,j \leq k}$, where $\Pi_{ij}^{(d)} \in \mathbf{R}^{t \times t}$. Then

$$(1_k' \otimes I_t)\Pi_d(1_k \otimes I_t) = \sum_{i=1}^{k}\sum_{j=1}^{k}\Pi_{ij}^{(d)} = Q_d,$$

say.
The matrix Q_d has constant row- and column-sums k and its elements are nonnegative integers. Condition (3iii) thus reads

$$T_d'P = \frac{1}{t-1}\{k(t-2)1_t1_t' + Q_d\}. \tag{4}$$

Condition (4) can only be fulfilled if the elements of the right hand side are integers. Obviously, $\frac{1}{t-1}k(t-2)1_t1_t'$ can have integral entries only if k is divisible by $t-1$. So a necessary condition for the existence of designs in our setting which fulfill conditions (3i), (3ii) and (3iii) is that either k is divisible by $t-1$ or that Q_d has only nonzero entries. The smallest k for which Q_d can have only nonzero entries, is $k = t$, when $Q_d = 1_t1_t'$ is possible. However, if $Q_d = 1_t1_t'$, then (4) reads as

$$T_d'P = \frac{1}{t-1}\{k(t-2)1_t1_t' + 1_t1_t'\} = t\ 1_t1_t'.$$

This is not very interesting as it means that every treatment appears in every row equally often.
Kunert (1990) considered the case $k = a(t-1)$, $a \in \mathbf{N}$, where for all $a \geq 2$ and all $t \geq 3$ there are designs fulfilling (3i), (3ii) and (3iii), which are not balanced block designs with rows as blocks. One example is

$$d = \begin{bmatrix} & 1 & 1 & 1 & & 4 & 2 & 3 & & 4 & 2 & 3 & & 1 & 1 & 1 & & 4 & 2 & 3 & & 4 & 2 & 3 \\ 2 & & 2 & 2 & 3 & & 4 & 1 & 3 & & 4 & 1 & 2 & & 2 & 2 & 3 & & 4 & 1 & 3 & & 4 & 1 \\ 3 & 3 & & 3 & 4 & 1 & & 2 & 4 & 1 & & 2 & 3 & 3 & & 3 & 4 & 1 & & 2 & 4 & 1 & & 2 \\ 4 & 4 & 4 & & 2 & 3 & 1 & & 2 & 3 & 1 & & 4 & 4 & 4 & & 2 & 3 & 1 & & 2 & 3 & 1 & \end{bmatrix}$$

with $t = 4$ and $k = 6$. For a general construction, see Kunert (1990).

We now examine the case $t = 3$ for small k.

Case 1: $k = 1$.

Then, except for a permutation of rows and columns $Q_d = I_3$. Hence, $\frac{1}{2}(k\, 1_3 1_3' + Q_d)$ cannot have only integral entries.

Case 2: $k = 2$.

Then, except for a permutation of rows and columns, there are only three possibilities for Q_d, namely,

$$Q_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \text{ or } Q_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \text{ or } Q_3 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Only $\frac{1}{2}(k\, 1_3 1_3' + Q_1)$ has integral entries. An example with $Q_d = Q_1$ and $T_d' P = \frac{1}{2}(2\, 1_3 1_3' + Q_d)$ is

$$d = \begin{bmatrix} & 1 & 1 & & 3 & 2 \\ 2 & & 2 & 3 & & 1 \\ 3 & 3 & & 2 & 1 & \end{bmatrix},$$

which is a member of the class considered by Kunert (1990). It is, however, not very interesting as it is a balanced block design with rows as blocks.

Case 3: $k = 3$.

As said before, in the case $k = t$ there is only one Q_d for which $\frac{1}{2}(k\, 1_3 1_3' + Q_d)$ has only integral entries, namely $Q_d = 1_3 1_3'$. An example with this Q_d which fulfills (3iii) is

$$d = \begin{bmatrix} & 1 & 1 & & 3 & 3 & & 2 & 2 \\ 2 & & 2 & 1 & & 1 & 3 & & 3 \\ 3 & 3 & & 2 & 2 & & 1 & 1 & \end{bmatrix}$$

for which each treatment appears twice in each row.

Case 4: $k = 4$.

Here are three different matrices Q_d for which the matrices $\frac{1}{2}(k\, 1_3 1_3' + Q_d)$ have only integral entries. These are

$$Q_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \text{ or } Q_2 = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix} \text{ or } Q_3 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

For $f = 1$, 2 and 3 there is a design $d(f)$ with $Q_{d(f)} = Q_f$ fulfilling (3iii). These are

$$d(1) = \begin{bmatrix} & 1 & 1 & & 3 & 2 & & 1 & 1 & & 3 & 2 \\ 2 & & 2 & 3 & & 1 & 2 & & 2 & 3 & & 1 \\ 3 & 3 & & 2 & 1 & & 3 & 3 & & 2 & 1 & \end{bmatrix},$$

$$d(2) = \begin{bmatrix} & 1 & 1 & & 2 & 3 & & 2 & 1 & & 3 & 2 \\ 2 & & 3 & 3 & & 1 & 1 & & 2 & 3 & & 1 \\ 3 & 2 & & 2 & 1 & & 3 & 3 & & 1 & 2 & \end{bmatrix}$$

and

$$d(3) = \begin{bmatrix} & 1 & 1 & & 1 & 1 & & 2 & 3 & & 2 & 3 \\ 2 & & 2 & 3 & & 2 & 3 & & 1 & 3 & & 1 \\ 3 & 3 & & 2 & 3 & & 2 & 1 & & 2 & 1 & \end{bmatrix}.$$

Here, $d(1)$ is a member of the series constructed by Kunert (1990) and not a balanced block design with rows as blocks, while $d(2)$ is a balanced block design with rows as blocks. The design $d(3)$ is the most interesting as for $d(3)$ the design matrix in the simple block model with rows as blocks is not even of the form $c(I_3 - \frac{1}{3} 1_3 1_3')$.

3 On the non-validity of the usual row-column model under a randomization-theory viewpoint

In the case of a non-orthogonal row-column structure there is very little freedom for randomization. This creates the suspicion that model (1) may not be justifyable with the usual randomization argument. We will show for an example that indeed it cannot. Hence, it might be appropriate to consider different, more complicated models to analyse designs with a non-orthogonal row-column structure.

The example considered is the simple case that $u = p = 4$ and $t = 3$. Assume that each row has only 3 units, the unit in row i and column i is missing, $1 \leq i \leq 4$. It is easily seen that each design for which each treatment appears exactly once in each row and column, is universally optimal in model (1) for this situation. Except for permutation of treatment labels, there are only four different such designs. These are

$$d_1 = \begin{bmatrix} & 1 & 2 & 3 \\ 3 & & 1 & 2 \\ 2 & 3 & & 1 \\ 1 & 2 & 3 & \end{bmatrix}, d_2 = \begin{bmatrix} & 1 & 2 & 3 \\ 1 & & 3 & 2 \\ 2 & 3 & & 1 \\ 3 & 2 & 1 & \end{bmatrix}, d_3 = \begin{bmatrix} & 1 & 2 & 3 \\ 1 & & 3 & 2 \\ 3 & 2 & & 1 \\ 2 & 3 & 1 & \end{bmatrix}$$

and $d_4 = \begin{bmatrix} & 1 & 2 & 3 \\ 2 & & 3 & 1 \\ 1 & 3 & & 2 \\ 3 & 2 & 1 & \end{bmatrix}$.

Hence, if we want to run such an experiment, then the only randomization left to do is to select one of designs d_1, d_2, d_3 and d_4 according to some prespecified probability measure and then to randomize treatment labels. We want to show that this cannot lead to validity of model (1). The technical device used is very similar to the one used in Bailey, Kunert and Martin (1991).

A necessary condition for the validity of model (1) is that in absence of treatment effects the expected mean square for treatments should be the same as the expected mean square for error. Now assume that the errors in model (1) are not uncorrelated, but have an unknown correlation structure Σ. Then, in general for one fixed design d, the expected mean squares will not be the same. A valid randomization, however, will create a probability distribution on a set of designs, such that the *average* over this distribution of the expected mean squares for error will equal the *average* expected mean square for treatments. Since the covariance structure is unknown, we have to find a distribution having this property for every positive definite matrix Σ. Such a distribution can indeed be found for model (1) in the case of an orthogonal row-column structure. In that case we start with any row-column design d. We consider the set of all designs which can be created from d by a permutation of rows and a permutation of columns. The probability distribution gives equal probability to each of those designs.

Assume we want to create such a distribution over the set $\mathcal{D} = \{d_1, d_2, d_3, d_4\}$. Note that the error mean square and the treatment mean square are not changed by permutation of treatment labels. Hence, this set $\mathcal{D}$ is the largest possible set over which the randomization can be extended, if we restrict attention to designs for which treatments are orthogonal to rows and columns in model (1).

For a fixed design $d_f \in \mathcal{D}$ the treatment mean squares, SST_f, equals

$$SST_f = Y' pr^{\perp}([P,U]) T_{d_f} (T'_{d_f} pr^{\perp}([P,U]) T_{d_f})^{-} T'_{d_f} pr^{\perp}([P,U]) Y.$$

It is easy to see that for all d_f we have

$$T'_{d_f} pr^{\perp}([P,U]) = T'_{d_f} pr^{\perp}(1_{12})$$

and

$$T_{d_f}' pr^{\perp}([P,U])T_{d_f} = 4(I_3 - \frac{1}{3}1_3 1_3').$$

We also have

$$P'U = 1_4 1_4' - I_4.$$

Denote the measurement on the unit in row i and column j by Y_{ij}. Remember that Y_{ii} is not defined as there aren't such units. Further, let $Y_{i\cdot}$ and $Y_{\cdot j}$ denote the marginal sums, while $\sum_{[r]} Y_{ij}$ is the sum of all measurements on units receiving treatment r. Then

$$SST_f = \frac{1}{4}\{\sum_{r=1}^{3}(\sum_{[r]} Y_{ij})^2 - \frac{1}{3}Y_{\cdot\cdot}^2\}.$$

Note that the units receiving treatment r are determined by the design d_f. Similarly, we have for the error sum of squares, SSE_f, that

$$\begin{aligned} SSE_f &= Y'pr^{\perp}([P,U,T_{d_f}])Y = Y'pr^{\perp}([P,U])Y - SST_f \\ &= Y'Y - \frac{3}{8}Y'PP'Y - \frac{3}{8}Y'UU'Y - \frac{1}{4}Y'UP'Y + \frac{1}{6}Y'1_{12}1_{12}'Y - SST_f \\ &= \sum_{i=1}^{4}\sum_{j=1}^{4} Y_{ij} - \frac{3}{8}\sum_{i=1}^{4}(Y_{i\cdot})^2 - \frac{3}{8}\sum_{j=1}^{4}(Y_{\cdot j})^2 - \frac{1}{4}\sum_{i=1}^{4} Y_{i\cdot}Y_{\cdot i} + \frac{1}{6}Y_{\cdot\cdot}^2 - SST_f. \end{aligned}$$

We calculate the expected sum of squares of treatments and of error for three special covariance matrices.

Assume Σ_1 is such that var $Y_{ij} = \sigma^2$ for all i and j. Further, let Cov $(Y_{14}, Y_{23}) = \rho$ while all other covariances are assumed zero. Then

$$\begin{aligned} E(SST_f) &= \frac{1}{4}\{(12\sigma^2 + 2x_f\rho) - \frac{1}{3}(12\sigma^2 + 2\rho)\} \\ &= 2\sigma^2 + (x_f/2 - 1/6)\rho, \end{aligned}$$

where E indicates the expectation and $x_f \in \{0, 1\}$ indicates whether design d_f applies the same treatment to units (1,4) and (2,3). We also have

$$\begin{aligned} E(SSE_f) &= 12\sigma^2 - \frac{3}{8}12\,\sigma^2 - \frac{3}{8}12\,\sigma^2 - \frac{1}{4}0 + \frac{1}{6}(12\sigma^2 + 2\rho) - E(SST_f) \\ &= 3\sigma^2 + (\frac{1}{2} - x_f/2)\rho. \end{aligned}$$

There are 3 degrees of freedom for error in model (1) and 2 degrees of freedom for treatments. Hence, the expected treatment mean square equals the expected error mean square if and only if

$$\sigma^2 + (x_f/4 - 1/12)\rho = \sigma^2 + (1/6 - x_f/6)\rho.$$

This can only be true for all σ^2 and ρ if

$$x_f = 3/5.$$

There is no d_f such that $x_f = 3/5$. For a valid randomization, however, the mean of x_f over the randomization distribution must be 3/5. Since x_f equals 0 for d_1 and 1 for d_2, d_3 and d_4, a randomization can only be valid if it has probability 2/5 of selecting d_1 and 3/5 of selecting one of d_2, d_3 or d_4.
Now consider a second covariance structure Σ_2, where var $Y_{ij} = \sigma^2$ for all i and j, but Cov $(Y_{12}, Y_{42}) = \rho$ and all other covariances are zero. Then

$$E(SST_f) = 2\sigma^2 + (z_f/2 - 1/6)\rho,$$

where $z_f \in \{0, 1\}$ indicates whether units (1,3) and (4,2) receive the same treatment. We also have that

$$\begin{aligned} E(SSE_f) &= 12\ \sigma^2 - \frac{3}{8}12\ \sigma^2 - \frac{3}{8}12\ \sigma^2 - \frac{1}{4}0 + \frac{1}{6}(12\sigma^2 + 2\rho) - E(SST_f) \\ &= 3\sigma^2 + (1/2 - z_f/2)\rho. \end{aligned}$$

Hence, for a valid randomization we also need a randomization mean for z_f of 3/5. Since z_f equals 0 for d_3 and 1 for d_1, d_2 and d_4 the valid randomization must have a probability of 2/3 of selecting d_3 as well.
Finally, we consider Σ_3 such that var $Y_{ij} = \sigma^2$ for all i and j and $Cov(Y_{13}, Y_{24}) = \rho$ while all other covariances are zero. Again

$$E(SST_f) = 2\sigma^2 + (u_f/2 - 1/6)\rho,$$

where $u_f \in \{0, 1\}$ indicates whether units (1,3) and (2,4) receive the same treatment. Like before, we get

$$E(SSE_f) = 3\sigma^2 + (\frac{1}{2} - u_f/2)\rho.$$

This implies that for a valid randomization the average of u_f must be 3/5. Since $u_f = 0$ for d_4 and 1 for d_1, d_2 and d_3, a valid randomization must have a probability of 2/5 of selecting d_4, as well.
Altogether, we have shown that it is impossible to find a randomization such that the expected mean squares for treatments equals the expected mean squares for error for all covariance matrices. This shows that model (1) cannot be justified by randomization arguments.

Acknowledgement

Research for this paper was sponsored by a Heisenberg Scholarship of the Deutsche Forschungsgemeinschaft.

References

Bailey, R.A., Kunert, J. and Martin, R.J. (1991): Some Comments on Gerechte Designs II. Randomization Analysis, and Other Methods that Allow for Inter-Plot Dependence. *J. Agronomy & Crop Science* 166, 101 — 119.

Baksalary, J.K. and Pukelsheim, F. (1992): Adjusted Orthogonality Properties in Multiway Block Designs. In: *Data Analysis and Statistical Inference. Festschrift in Honour of Friedhelm Eicker* (S. Schach & G. Trenkler, ed.) Eul Verlag, Köln, 413 — 420.

Jacroux, M. & Saha Ray, R. (1991): On the determination and construction of optimal row-column designs having unequal row and column sizes. *Ann. Inst. Statist. Math.* 43, 377 — 390.

Kiefer, J. (1975): Construction and optimality of generalized Youden designs. In: *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) North Holland, Amsterdam, 333 — 353.

Kunert, J. (1983): Optimal design and refinement of the linear model with application to repeated measurements designs. *Ann. Stat.* 11, 247 — 257.

Kunert, J. (1990): A note on optimal designs with a non-orthogonal row-column structure. *J. Statist. Plann. Inf.* to appear

Kurotschka, V. (1978): Optimal design of complex experiments with qualitative factors of influence. *Comm. Stat. — Theor. Meth.* A7, 1363 — 1378.

Mukhopadhyay, A.C. and Saharay, R. (1988): Three way elimination of heterogeneity with non-orthogonal incidence structure for every two directions. *SankhyāB* 50, 181 — 194.

Saharay, R. (1986): Optimal designs under a certain class of non orthogonal row-column structure. *SankhyāB* 48, 44 — 67.

Shah, K.R. and Sinha, B.K. (1990): Optimality Aspects of Row-Column Designs with Non-orthogonal Structure. Paper presented at the *Workshop on Linear Models, Experimental Designs, and Related Matrix Theory* in Tampere, Finland, August 6 — 8, 1990.

Steward, F.P. and Bradley, R.A. (1991): Some universally optimal row-column designs with empty nodes. *Biometrika* 78, 337 — 348.

Optimal Simulation Design by Branching Technique

V.B. Melas

The paper is devoted to a theory of branching simulation technique for general Markov chains. A representation for variance of linear functional estimators is obtained. It allows to receive optimal designs interpreted as functions determining branching numbers. The optimality criterion is multiplication of the variance by summing length of simulated trajectories. For the important partical case of random walk simulation as well as for the case of finite Markov chains the theory is applied and optimal designs are obtained. The efficiency of the designs is investigated and seems rather substantial.

1 Introduction

Simulation experiment is interpreted here as construction of random process trajectories by computer to obtain parameters of interest estimators. The random process can describe dynamic behaviour of a real system or of an ideal object. Let $\{Z_t\}$ be a family of random values defined on a measurable space $\{X, \mathcal{A}\}$ where X is a set (state space) and $\mathcal{A}$ is a σ-field of its subsets. In many applications X is a finite or countable space or an interval and $\mathcal{A}$ is introduced by the standard way, t can be from an interval or a discrete set.

Consider behaviour of a queueing system with a single service device denoted by $GI/G/1/\infty$ in the Kendall classification. It can be desribed by the following way. Demands come in moments $t_1, t_2, \ldots$ and need sevice times $u_1, u_2, \ldots$ respectively. Demands are servied in the order of coming. Let random values $X_i = u_i - v_i$ where $v_i = t_{i+1} - t_i$ be independent and identically distributed with a distribution density f. Let $W_1, W_2, \ldots$ be the time spending in the queue by first, second,... demand. Suppose that in the initial moment $t = t_0 = 0$ the system is free of demands. Then

$$W_1 = 0,\ W_{n+1} = \max(0, W_n + X_n),\ n = 1, 2, \ldots$$

Consider the process

$$S_1 = 0,\ S_{n+1} = S_n + X_n,\ n = 1, 2, \ldots$$

and define

$$M_n = \max\{S_1, \ldots, S_n\},\ n = 1, 2, \ldots$$

The process $\{S_n = n \geq 1\}$ is the random walk process connected with the waiting process $\{W_n;\ n \geq 1\}$. It is known (see 1,Ch.XII) that W_n and M_n have the same distribution and $W_n \to W,\ M_n \to M$ in distribution where W and M are random values if it is supposed that

$EX_1 < 0$. Denote $\theta = P\{W \geq v\} = P\{M \geq v\}$. This quantity is the probability of exiding a given value (v) by the time spending in the queue by a demand in the stationary limit. The closed analytical expression for θ is generally unknown and the problem is to estimate θ by simulation $\{W_n; n \geq 1\}$ or $\{S_n; n \geq 1\}$. This problem has many important interpretations. We mention only the problem of sequential test power estimation (see 2).

The considered problem is only one of a great variaty of system simulation examples (see 3, 4). The simulation method is greatly universal but can require tremendous computer time. So simulation design is needed. For this purpose can be used general optimal design methods since different trajectories can be considered as independent ones (see 5). But for optimization of a single trajectory performance special methods are desirable. Such methods are usually called variance reduction techniques. There are a number of classical methods to be found in known monographies or review papers 3, 4, 6, 7. They are: importance sampling, conditional Monte Carlo method, control variates, random cubatura formulas, antithetic variate and others. But almost all this methods require rather substantial prior information on the system under consideration. Such information is often absent and we need a more universal method.

Such method can be developed on the base of von Neumann idea of splitting and Russian roulette 8. This idea consists in parallel simulation of independent trajectories while each trajectory can be dublicated or vanished depending on its importance for estimation exactness. There are a number of papers considering partical problems 8-11 but they do not contain enough developed theory. Such theory is developed in author's papers 12, 13 for finite Markov chains and in 14, 15 for random walks. Here we generalise the approach of this papers to general Markov chains. It allows to improve previous results. Note that many of system simulation problems can be reduced to enclosed Markov chain simulation (see, f.e., 7). Therefore the developed theory seems rather universal.

2 Formulation of the problem

Consider the problem of a linear functional estimation of a general Markov chain. Let us introduce necessary notations and definitions. Theirs full exposition can be found in 16. Let random values $X_1, X_2, \ldots$ are defined on a probability space $(X, \mathcal{A}, P)$ and form a homogeneous Markov chain with transition function $P(x, dy) = P\{X_1 \in dy | X_0 = x\}$ measurable in the first argument and such that $\int_X P(x, dy) = 1$. Further $\int$ will be placed for $\int_X$. Suppose that σ-field $\mathcal{A}$ is countable generated. Define the function

$$h_B^\infty(x) = P\{X_n \in B \quad \text{infinitely often} \quad | X_0 = x\}$$

where "$X_n \in B$ infinitely often" means $\cap_{m=0}^\infty \cup_{n=m}^\infty \{X_n \in B\} \neq \emptyset$, $\emptyset$ is the empty set.

Let us say that a Markov chain is Harris recurrent if there exists a measure μ on $(X, \mathcal{A})$ such that $\mu(X) > 0$ and

$$h_B^\infty(x) \equiv 1 \quad \text{for} \quad x \in X, B \in \{B \in \mathcal{A}; \mu(B) > 0\}.$$

It is known that for Harris recurrent Markov chains there exist a nontrivial σ-finite measure π such that

$$\pi(dy) = \int P(x, dy)\pi(dx).$$

If $\pi(X) < \infty$ then the chain will be called positively recurrent. In this case put $\pi(X) = 1$. Such measure is probability measure and is called stationary one. Arbitrary Harris recurrent Markov chain is irredusable and there exist a function $s(x)$ and a probability measure ν such that $\int_B s(x)\mu(dx) > 0$ if $\mu(B) > 0$ and

$$P^{m_0}(x, B) \geq s(x)\nu(B) \tag{1}$$

for every $x \in X, B \in \mathcal{A}$.

Markov chain will be called Harris positively recurrent if it is positively recurrent and Harris recurrent as well. The Markov chain will be called ergodic if it is Harris positive recurrent and aperiodic. It is known that if $\{X_n; n \geq 0\}$ is ergodic then for arbitrary function h integrable in measure π and any probability measure ν on $(X, \mathcal{A})$

$$\nu p^n h^{\mathrm{df}} \to = \int \nu(dx)p^n(x, dy)h(y) \to (\pi, h)^{\mathrm{df}} \to = \int \pi(dx)h(x)$$

with $n \to \infty$.

Denote $J = (\pi, h)$ and consider the problem of its estimation by simulation of $\{X_n; n \geq 1\}$ trajectories.

Suppose that $h(x)$ is a measurable on $(X, \mathcal{A})$ function, $\tilde{h}(x) = h(x) - J$ and

$$\int |\tilde{h}(x)|\pi(dx) < \infty,$$

$$\int |\tilde{h}(x)| E\left\{\sum_{n=1}^{\tau(B)} h(X_n)|X_0 = x\right\} \pi(dx) < \infty \tag{2}$$

for any $B \in \mathcal{A}$ where $\tau(B) = \min\{n; X_n \in B\}$. Suppose that $m_0 = 1$ for the sake simplicity (general case can be considered similarily).

3 Parameter estimators

Consider two estimators of J: the direct estimator and the regeneration method estimator. The direct estimator can be determined merely as the sum of the function $h(x)$ values along the trajectory:

$$J_N = \sum_{i=0}^{N} h(X_i)/N$$

where $\{X_n; n \geq P\}$ is a Markov chain with arbitrary initial distribution. The following result is proved in 16,theorem 7.6.

Theorem 1

Let $\{X_n; n \geq 0\}$ be an ergodic Markov chain satisfying the condition (1) with $m_0 = 1, h(x)$ be measurable function on $(X, \mathcal{A})$ satisfying the condition (2). Then the distribution of the random value $\sqrt{N}(\hat{J}_N - J)$ tends to normal one with zero expectation and the variance σ^2 where

$$\sigma^2 = (\pi, 2\tilde{\varphi}\tilde{h} - \tilde{h}^2),$$

$\tilde{\varphi}$ is iterative solution of the equation

$$\tilde{\varphi} = (P - s \otimes \nu)\tilde{\varphi} + \tilde{h}$$

For determination and investigation of the regeneration method estimator it is usefull the following construction (see, also, 16,Ch.4). Let $Z_n (n = 0, 1, \ldots)$ takes values one or zero and $z_n = 1$ with the probability $S(X_n)$ and $Z_n = 0$ with the probability $1 - S(X_n)$. Let the random value X_0 distribution is $\nu(dx), \alpha_0 = 0, \alpha_1 = \min\{n \geq 0, Z_n = 1\}, \alpha_2 = \min\{n > \alpha_1, Z_n = 1\}, \ldots$ It is known that $\alpha_i, i = 1, 2, \ldots$ are random values and the distribution of X_{α_i} coinsides with $\nu(dx), i = 1, 2, \ldots$ Therefore pairs $(Y(k), b(k)), k = 1, 2, \ldots$ where

$$Y(k) = \sum_{i=\alpha_k - 1}^{\alpha_k - 1} \tilde{h}(X_i), b_k = \alpha_k - \alpha_{k-1}$$

are independent and have the same distribution. The values collection $\{X_{\alpha_{k-1}}, \ldots, X_{\alpha_k - 1}\}$ is called k-th circle or tour. Note that $Eb(1)$ is the expected length of one tour. Consider the estimator

$$\tilde{J}_k = (1/k) \sum_{i=1}^{k} Y(i) / (\sum_{i=1}^{k} b(i)/k).$$

Such estimators were introduced in 17. They were considered in a number of papers (see 8). Nevertheless the following result is seemingly new. Besides the approach introducing for its prove is applicable for branching technique as well.

Theorem 2

Let the conditions of theorem 1 are fulfilled. Then $\tilde{J}_k$ tends to J with probability one and the distribution of the random value $\sqrt{k}(\tilde{J}_k - J)$ tends to the normal one $(k \to \infty)$ with zero expectation and the variance σ_1^2. Besides $Eb(1) = 1/\int s(x)\pi(dx)$ and the magnitude $Eb(1) \cdot \sigma_1^2$ equals

$$(\pi, 2\tilde{\varphi}\tilde{h} - \tilde{h}^2) = (2DY(1) - 2JCov(Y(1), b(1) + J^2 Db(1))/Eb(1).$$

Proofs of this and the following theorems are to be found in Appendix.

Remark 1

Note that the quantity $Eb(1)\sigma_1^2$ can be considered as the asymptotic efficiency of the estimator $\tilde{J}_k$. The corresponding value for the estimator $\hat{J}_N$ is σ^2. So we see that the asymptotic efficiency of the direct estimator and the regeneration one coincide and do not depend on the choose of ν and s in the condition (1).

Both estimators can be improved by branching technique to which the remainder of the paper is devoted.

4 Branching technique

Let at an n-th step η_n copies of the chain $\{X_n, Z_n\}$ are simulated: $\{X_n^{(\gamma)}, Z_n^{(\gamma)}; \gamma \in 1 : \eta_n\}$. The procedure regulating numbers of the copies can be introduced in the following way.

Definition. Arbitrary measurable on $(X, \mathcal{A})$ function $\beta(x)$ such that

$$\inf \beta(x) > 0 \beta(x) = 1 \quad \text{for such } x \text{ that } \quad s(x) \neq 0 \tag{3}$$

will be called branching simulation design (BSD). Determine η_0 as the random value:

$$\eta_0 = \begin{cases} 1 & \text{if } z_0 = 1, \\ \lfloor \beta(x) \rfloor \text{ with probability } (1 - \overline{\beta(x)}) & \text{if } z_0 = 0, X_0 = x \\ \lfloor \beta(x) \rfloor \text{ with probability } \overline{\beta(x)} & \text{if } z_0 = 0, X_0 = x \end{cases}$$

where $\lfloor \beta(x) \rfloor$ is the integer part of $\beta(x), \overline{\beta(x)} = \beta(x) - \lfloor \beta(x) \rfloor$. If $X_n^{(\gamma)} = x$ and $X_{n+1}^{(\gamma)} = y (n = 0, 1, 2, \ldots; \gamma = 1, 2, \ldots, \eta_n)$ add $r^{(\gamma)}(x, y) - 1$ chains beginning at $X_{n+1} = X_{n+1}^{(\gamma)}$ if $r^{(\gamma)}(x, y) \geq 1$ where

$$r^{(\gamma)}(x, y) = \begin{cases} 1 & \text{if } z_0^{(\gamma)} = 1 \\ \lfloor \beta(y)/\beta(x) \rfloor \text{ with probability } 1 - \overline{(\beta(y)/\beta(x))} & \text{if } z_0^{(\gamma)} = 0 \\ \lfloor \beta(y)/\beta(x) \rfloor + 1 \text{ with probability } \overline{\beta(y)/\beta(x)} & \text{if } z_0^{(\gamma)} = 0 \end{cases}$$

If $r^{(\gamma)}(x, y) < 1$ and $z_0^{(\gamma)} = 0$ then simulation of the copy is abrubt with probability $1 - \overline{(\beta(y)/\beta(x))}$ and is continued with probability $\overline{\beta(y)/\beta(x)}$. After simulation of the current step of all η_n copies newly arising and remaining copies are numerated over again. All copies are simulated up to hit in $X \otimes \{1\}$.

The collection of values $\{X_n^{(\gamma)}; \gamma \in 1 : \eta_n\}$ obtained in the above procedure will be called circle. Each circle abrubts with probability one since below theorem 2 shows that $E \sum_{n=1}^{\infty} \eta_n < \infty$.

Consider results of k circles simulation. Denote them by

$$\{X_n^{(\gamma)}(1); \gamma \in 1 : \eta_n(1)\}, \ldots, \{X_n^{(\gamma)}(k); \gamma \in 1 : \eta_n(k)\}.$$

Determine the estimator

$$\widetilde{J_{\beta,k}} = \left(\sum_{i=1}^{k} Y_\beta(i)/k \right) / \left(\sum_{i=1}^{k} b_\beta(i)/k \right)$$

where

$$Y_\beta(i) = \sum_{n=0}^{\infty}\sum_{\gamma=1}^{\eta_n(i)} h(X_n^{(\gamma)}(i))/\beta(X_n^{(\gamma)}(i)),$$

$$b_\beta(i) = \sum_{n=0}^{\infty}\sum_{\gamma=1}^{\eta_n(i)} 1/\beta(X_n^{(\gamma)}(i)), \quad i = 1,2,\ldots$$

Set

$$T_{\beta,k} = \sum_{i=1}^{k}\sum_{n=0}^{\infty}\eta_n(i)$$

This magnitude equals the summing number of steps of all simulated in k circles trajectories. Denote

$$Q(x,dy) = P(x,dy) - s(x)v(dy),$$

$$D_Q\tilde{\varphi}(x) = D[\overline{\varphi}(X_2)|X_1] = \int \tilde{\varphi}^2(y)Q(x,dy) - (\int \tilde{\varphi}(y)Q(x,dy))^2,$$

$$r_Q\tilde{\varphi}(x) = \int [\tilde{\varphi}(y)\beta(x)/\beta(y)]^2[\overline{\beta(y)/\beta(x)} - (\overline{\beta(y)/\beta(x)})^2]Q(x,dy),$$

$$D_\nu\tilde{\varphi} = \int \tilde{\varphi}^2(y)\nu(dy),$$

$$r_\nu = \int [\tilde{\varphi}(y)/\beta(y)]^2[\overline{\beta(y)} - (\overline{\beta(y)})^2]\nu(dy)$$

where $\tilde{\varphi}$ is defined in theorem 1.

Let

$$T_\beta = \lim_{k\to\infty} \overline{T}_{\beta,k}/k,\ \sigma_{1\beta}^2 = \lim_{k\to} k D\tilde{J}_{\beta,k}. \tag{4}$$

Theorem 3

Let the conditions of theorem 1 are satisfied. Then for arbitrary BSD $\beta(x)$ the limits in (4) exist and are finite, $\tilde{J}_{\beta,k} \to J$ $(k \to \infty)$ with probability one and the distribution of the random value $\sqrt{k}(\tilde{J}_{\beta,k} - J)$ tends to the normal one with zero expectation and the variance $\sigma_{1\beta}^2$ and

$$T_\beta\sigma_{1\beta}^2 = \int \beta(x)\pi(dx) \times \{\int \beta^{-1}(x)\pi(dx)[D_Q\tilde{\varphi}(x) + r_Q(x)] + \int s(x)\pi(dx)[D_\nu\tilde{\varphi} + r_\nu]\}$$

$$= [E\sum_{n=0}^{\infty}\eta_n/(Eb_\beta(1))^2] \times [EY_\beta^2(1) - 2JE(Y_\beta(1),b_\beta(1)) + J^2Eb_\beta^2(1)].$$

Remark 2

It is obvious that taking $s_1(x), s_1(x) = 1$ if $s(x) = 1, s_1(x) = 0$ if $s(x) < 1$ and $Q_1(x,dy) = P(x,dy) - s_1(x)\nu(dy)$ instead of $s(x)$ and $Q(x,dy)$ we can only reduce the quantity $T_\beta\sigma_{1\beta}^2$

since the value will be the same and it arise more freedom in choosing $\beta(x)$. So we assume further that $s(x) = 1, x \in S$ and $s(x) = 0, x \in \mathcal{X} \setminus S$ and get

$$T_\beta \sigma_{1\beta}^2 = \int \beta(x)\pi(dx0 \int \beta^{-1}(x)\pi(dx)[D_Q\tilde{\varphi}(x) + r_Q(x) + D_\nu\tilde{\varphi}(x) + r_Q(x) + D_\nu\tilde{\varphi} + r_\nu].$$

Note that the last expression does not depend on the choose of (s, ν) in (1). It can be proved by a direct calculation.

Remark 3

Note that the magnitudes $r_Q(x)$ and r_ν depend on $\beta(x)$. They connected with randomness of branching numbers. But they can be ignored in typical situations. Particularly they can be substantially reduced by simulation N circles simultaneously if we will use the rotation sampling method (see, f.e.18) for chosing branching numbers. It can be realysed in the following way. Let $\mathcal{X} = \cup_{i=1}^m \mathcal{X}_i$, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$, $i \neq j$, $\beta(x) = c_i$ for $x \in \mathcal{X}_i$. Suppose that at n-th step k copies transit from $x \in \mathcal{X}_i$ to $y \in \mathcal{X}_j$. Then the branching number for copy $R(R = 1, 2, \ldots, k)$ can be determined by the formula

$$[c_j/c_i] + \overline{[\alpha l/m]}$$

where α is a random value with the uniform distribution on $[0, 1]$. Remarks 2 and 3 allow to define the following optimality criterion

$$\int \beta(x)\pi(dx) \left[\int \beta^{-1}(x)\pi(dx)(D_Q\tilde{\varphi}(x) + D_\nu\tilde{\varphi}) \right] \to \min_\beta$$

BSD minimising the left side will be called quasy optimal. Such designs can be obtained by Schwarz inequality.

Theorem 4

Let the conditions of theorem 1 are satisfied and $s(x) = 0, x \in \mathcal{X} \setminus S$. Then quasy optimal BSD $(\beta^*(x))$ is unique and

$$\beta^*(x) = \sqrt{D_Q\varphi(x)/D_\nu\varphi}, x \in \mathcal{X} \setminus S.$$

Note that the magnitudes $D_Q\varphi(x)$ and $D_\nu\varphi$ can be estimated in the simulation experiments by the standard way. Thus quasy optimal designs can be found in the sequential approach style.

Remark also that the branching analog of the direct estimator can be investigated in the similar way.

5 Random walks simulation

For the problem formulated in the introduction we can obtain additional results. Assume that $EX_1 < 0$. It is straightforward that the pair (s, ν):

$$s(x) = 1, x = 0, s(x) = 0, x > 0,$$

$$\nu(dx) = af(x)dx, x > 0, a = 1/(1 - F(0)), 1 - F(0) = \int_0^\infty f(t)dt$$

satisfies the condition (1) with $m_0 = 1$. For $\pi(dx) = P\{W \in dx\}$ it is known that 1,Ch.XII $\pi(dx) = M'(x)dx, x > 0, P\{W = 0\} = M(0)$ where $M(x) = P\{M < x\}$. The function

$$h(x) = 0,\ x < v,\ h(x) = 1,\ x \geq v$$

fulfills the conditions of theorem 1.

The explicit form of the asymptotically ($v \to \infty$) quasy optimal BSD and its asymptotic efficiency are given by the following theorem.

Theorem 5

If $EX_1 < 0$ and for a positive value λ_0

$$\int_{-\infty}^{\infty} e^{\lambda_0 t} f(t)dt = 1, \int_{-\infty}^{\infty} e^{2\lambda_0 t} f(t)dt = K < \infty$$

then for arbitrary BSD $\beta(x)$ and for $v \to \infty$ we have $\theta \to 0$ and

$$\lim_{v\to\infty} T_\beta \sigma_{1\beta}^2/(\theta^2 \ln^2(1/\theta)) \geq C$$

where C is a constant.

Asymptotically quasy optimal BSD $\beta^*(t)$ is

$$\beta^*(t) = e^{\lambda_0 t}, 0 < t < v, \beta^*(t) = e^{\lambda_0 v}, t \geq v$$

and for BSD $\beta(t) = \exp((\lambda_0 + \varepsilon)t), t < v, \beta(t) = \beta(v), t \geq v$ it holds

$$\lim_{v\to\infty} T_\beta \sigma_{1\beta}^2/\theta^{2+\epsilon} = C_1,$$

$$\lim_{\epsilon\to 0}\lim_{v\to 0} T_\beta \sigma_{1\beta}^2/(\theta^2 \ln^2(1/\theta)) = C_1$$

where C_1 is a constant and $C \leq C_1 \leq \frac{K-1}{K-3/4}C$.

Note that the asymptotic efficiency of the regeneration estimator is $O(\theta \ln(1/\theta))$ and of the importance sampling estimator 2 is $O(\theta^2 \ln(1/\theta))$. So the efficiency of the branching technique estimator is very close to that of the importance sampling one. However branching technique is more universal than importance sampling. This result is an improvement of the main result in 15.

6 Finite Markov chains

Let $\{X_n; n \geq 0\}$ be a homogeneous positively recurrent aperiodic Markov chain with the finite state space $\{0, 1, \ldots, m\}$ and a transition matrix $P = (p_{ij})_{i,j=0}^m$. Functions $s_{i_0} = 1, s_i = 0, i \neq i_0$ with $i_0 = 0, 1, \ldots, m$ play the role of $s(x)$. Take $i_0 = 0$. Then obviously $\nu(dx)$ becomes $\{\nu_i\}_{i=0}^m, \nu_i = p_{0i}, i = 0, 1, \ldots, m, Q(x, dy)$ becomes $Q = (\overline{p_{ij}})_{i,j=0}^n$ where $\overline{p_{ij}} = p_{ij}$ if $i \neq 0$ and $\overline{p_{0i}} = 0 (i = 0, 1, \ldots, m)$. Let $\tilde{\varphi} = (\tilde{\varphi}_0, \ldots, \tilde{\varphi}_m)^\top$ be the unique solution of the equation

$$\tilde{\varphi} = Q\tilde{\varphi} + \tilde{h}$$

where $\tilde{h}_i = h_i - J, h = (h_0, \ldots, h_m)$ is arbitrary vector with real components, $J = \sum_{i=0}^n h_i \pi_i$, $(\pi_0, \ldots, \pi_n)$ is the stationary distribution.

Results of §4 take the following form.

Theorem 6

For the above described Markov chain theorems 3 and 4 hold and

$$T_\beta \sigma_{1\beta}^2 = \sum_{i=0}^n \pi_i \beta_i \times \sum_{i=0}^n \pi_i \beta_i^{-1} \{[\sum_{j=1}^n p_{ij} \tilde{\varphi}_j^2 - (\sum_{j=1}^n p_{ij} \tilde{\varphi}_j)^2] + \sum_{j=1}^n (\tilde{\varphi}_j \beta_i / \beta_j)^2 [(\overline{\beta_i / \beta_j}) - [(\overline{\beta_i / \beta_j})]^2] p_{ij}\}$$

where $\beta_0 = 1$. Quasy optimal BSD is

$$\beta_i = \sqrt{\left(\sum_{i=1}^n p_{ij} \tilde{\varphi}_j^2 - (\sum p_{ij} \tilde{\varphi}_j)^2\right) / \sum_{j=1}^n p_{0j} \tilde{\varphi}_j^2}$$

Investigation of different BSD closed to optimal in the sense of magnitude $T_\beta \sigma_{1\beta}^2$ and another criterion for a special type of finite Markov chains can be found in the author's paper 13.

7 Appendix

Proof of theorem 2

The approach of the proof base on the study of two dual equations.

Define the random functions

$$\overline{\varphi}(x) = \left\{\sum_{i=0}^{\alpha_1} h(X_i) | X_0 = x\right\}, \quad \overline{\varphi}^{(n)}(x) = \left\{\sum_{i=0}^{n_1} h(X_i) | X_0 = x\right\}$$

where $\alpha_1 = \min\{k \geq 0, Z_k = 1\}$, $n_1 = \min\{n, \alpha_1\}$.

According to the condition (1) expectations of $\overline{\varphi}(x)$ and $\overline{\varphi}^{(n)}(x)$ exist and finite. Denote $\varphi(x) = E\overline{\varphi}(x)$, $\varphi^{(n)}(x)$. Since $E\{\sum_{i=0}^{n_1} |h(x_i)| \quad | X_0 = x\}$ is finite by (1) the limit $\varphi_n(x)$ $(n \to \infty)$ exist and equal to $\varphi(x)$. Obviously $\overline{\varphi}^{(0)}(x) = h(x)$ and

$$\overline{\varphi}^{(n+1)}(x) = \begin{cases} \overline{\varphi}^{(n)}(y) + h(x) & \text{with probability } Q(x, dy) \\ h(x) & \text{with probability } s(x)\nu(dy) \end{cases} \tag{5}$$

Calculating the expectation in (6) receive

$$\varphi^{(0)}(x) = h(x), \varphi^{(n+1)}(x) = \int Q(x, dy)\varphi^{(n)}(y) + h(x),\ n \geq 0. \tag{6}$$

By limit transition with $n \to \infty$ we obtain

$$\varphi(x) = \int Q(x, dy)\varphi(y) + h(x). \tag{7}$$

Consider the dual equation

$$\psi(dx) = \int Q(y, dx)\psi(dy) + \nu(dx) \tag{8}$$

where $\psi(dx)$ is a measure on $(\mathcal{X}, \mathcal{A})$. This equation has the explicit solution

$$\psi(dy) = \pi(dy) \Big/ \int s(x)\pi(dx) \tag{9}$$

where $\pi(dy)$ is the stationary distribution of the chain $\{X_n; n \geq 0\}$. Solutions of equations (7) and (8) are connected by the following result. Set

$$\psi^{(0)}(dx) = \nu(dx), \psi^{(n+1)}(dx) = \int Q(y, dx)\psi^{(n)}(dy) + \nu(dx),\ n = 0, 1, \ldots \tag{10}$$

Lemma 1

Let $P(x, dy)$ is the transition function of an ergodic Markov chain satisfying the condition (1) with $m_0 = 1$, $Q(x, dy) = P(x, dy) - s(x)\nu(dy)$. The following assertions hold

(i) there exist the limit

$$\lim_{n \to \infty} \psi^{(n)}(B) = \psi(B)$$

for arbitrary $B \in \mathcal{A}$ and $\psi(dy)$ is determined by the formula (10);

(ii) if $h(x)$ is a measurable function on $(\mathcal{X}, \mathcal{A})$ fulfilling conditions (2) then

$$\lim \int \varphi^{(n)}(x)\nu(dx) = \int \pi(dx)h(x) \Big/ \int \pi(dx)s(x).$$

Proof of the lemma 1 is straightforward.

Thus we have using lemma 1

$$EY(1) = \int \varphi(x)\nu(dx) = \int \psi(dx)h(x) = \int \pi(dx)h(x) \Big/ \int \pi(dx)s(x).$$

Taking $h(x) \equiv 1$ receive

$$Eb(1) = \int \pi(dx) \Big/ \int \pi(dx)s(x) = 1 \Big/ \int \pi(dx)s(x).$$

By the strong law of large numbers

$$\tilde{J}_k = \frac{1}{k}\sum_{i=1}^{k} Y(i) \Big/ \left(\sum_{i=1}^{k} b(i)/k\right)_{k\to\infty} \longrightarrow EY(1)/Eb(1) = J \quad \text{a.s.}$$

Note that since the condition 1 hold σ^2 in theorem 1 is finite. Calculating the expectation of $(\overline{\varphi}^{(n+1)}(x))^2$ with the help of relation (6) and using the limit transition with $n \to \infty$ we receive

$$E\overline{\varphi}^2(x) = \int E\overline{\varphi}^2(y)Q(x,dy) + 2h(x)\varphi(x) - h^2(x),$$

$$EY^2(1) = \int E\varphi^2(x)\nu(dx) = \int \psi(dx)\{2h(x)\varphi(x) - h^2(x)\} = (\pi, 2h\varphi - h^2)/(\pi, s).$$

Set $\tilde{Y}(1) = Y(1) - Jb(1)$. Then

$$E\tilde{Y}(1) = 0,\ D\tilde{Y}(1) = EY^2(1) - 2JE(Y(1), b(1)) + J^2Eb^2(1).$$

Besides

$$D\tilde{Y}(1) = (\pi, 2\tilde{\varphi}\tilde{h} - \tilde{h}^2)/(\pi, s)$$

where $\tilde{h}(x) = h(x) - J$, $\tilde{\varphi}$ is the iterative solution of the equation

$$\tilde{\varphi} = Q\tilde{\varphi} + \tilde{h}.$$

Since by the strong law of large numbers

$$\sum_{i=1}^{k} b(i)/k \longrightarrow Eb(1) = (\pi, s) \quad (k \to \infty)$$

we have

$$\begin{aligned} \lim_{k\to\infty} kD\tilde{J}_k &= D\tilde{Y}(1)/(Eb(1))^1 = \\ &= (\pi, 2\tilde{\varphi}\tilde{h} - \tilde{h}^2)/(s, \pi). \end{aligned}$$

The distribution of the random value $\sqrt{k}(\widetilde{K}_k - J)$ tends to the normal one by the central limit theorem.

Proof of theorem 3

The proof is similar to the preceding one.

Define the random functions

$$\tilde{\varphi}_{\beta}^{(n)}(x) = \left\{ \sum_{i=0}^{n_1} \sum_{j=1}^{\eta_i} h(X_i^{(\gamma)}) | X_0 = x \right\}, \quad n = 0, 1, \ldots$$

where $n_1 = \min\{n, \alpha\}$, $\alpha = \min\{k; k \geq 0, \eta_{k+1} = 0\}$. Obviously

$$\tilde{\varphi}_{\beta}^{(0)}(x) = h(x), \tilde{\varphi}_{x,\beta}^{(n+1)}(x) = \begin{cases} \sum_{\gamma=1}^{r_{x,y}} \tilde{\varphi}_{\beta}^{(n)}(y) + h(x) & \text{with probability } Q(x, dy) \\ h(x) & \text{with probability } s(x)\nu(dy) \end{cases} \tag{11}$$

$n = 0, 1, \ldots$

Calculating the expectation of relation (12) both sides and using the known Wald's lemma and the limit transition ($n \to \infty$) receive

$$\varphi_{\beta}(x) = \int \frac{\beta(y)}{\beta(x)} Q(x, dy) \varphi_{\beta}(y) + h(x)/\beta(x)$$

and

$$\beta(x)\varphi_{\beta}(x) = \int Q(x, dy)\beta(y)\varphi_{\beta}(y) + h(x)$$

where $\varphi_{\beta}(x) = \lim_{n \to \infty} E\tilde{\varphi}_{x,\beta}^{(n)}(x)$.

Using the dual equation (9) and lemma 1 we obtain

$$\begin{aligned} EY_{\beta}(1) &= \int \nu(dx)\beta(x)\varphi_{\beta}(x) = \int \psi(dx)h(x) = \\ &= \int \pi(dx)h(x)/(\pi, s) = EY(1) \end{aligned}$$

and similarly

$$b_{\beta}(1) = Eb(1).$$

Denote $\hat{\varphi}_{\beta}(x) = \beta(x)\overline{\varphi}_{\beta}(x)$. Then $E\hat{\varphi}_{\beta}(x) = \varphi(x)$ and calculating the expectation of squares of relation (12) both sides we receive by limit transition

$$\begin{aligned} E(\hat{\varphi}_{\beta}(x))^2 &= \int \frac{\beta(x)}{\beta(y)} Q(x, dy) E\hat{\varphi}_{\beta}^2(y) \\ &+ \int Q(x, dy)\varphi^2(y)\{1 - \frac{\beta(x)}{\beta(y)} + \frac{\beta^2(x)}{\beta^2(y)}[\overline{\beta(y)/\beta(x)} - [\overline{\beta(y)/\beta(x)}]^2]\} \\ &+ 2h(x)\varphi(x) - h^2(x) \end{aligned}$$

Since

$$EY_\beta^2(1) = \int \nu(dx) E\left[\sum_{j=1}^{\eta_0} \overline{\varphi}_\beta^{(\gamma)}(x)\right]^2$$

where $\overline{\varphi}_\beta^{(\gamma)}(x)$ are independent realisations of $\overline{\varphi}_\beta(x)$ we receive by the Wald's lemma

$$EY_\beta^2(1) = \int \nu(dx)\beta^{-1}(x) E\hat{\varphi}_\beta^2(x) + \int \nu(dx) u(x)\varphi^2(x)$$

where

$u(x) = 1 - \beta^{-1}(x) + \beta^{-2}(x)(\overline{\beta(x)} - \overline{\beta(x)}^2)$.

Placing $\tilde{Y}_\beta(1) = Y_\beta(1) - Jb_\beta(1)$ instead $Y_\beta(1)$ we receive similarly to the proof of theorem 2 that

$$\sqrt{k} D\tilde{J}_{k,\beta k \to \infty} \longrightarrow D\tilde{Y}_\beta(1)/(Eb_\beta(1))^2 = D\tilde{Y}_\beta(1)/(Eb_\beta(1))^2$$

and the estimator $\tilde{J}_{\beta,k}$ is strongly consistent and asymptotically normal. Note that $\tilde{T}_{\beta,1} = Y_\beta(1)$ if we take $h(x) = \beta(x)$ and $E\tilde{T}_{\beta,1} = (\beta,\pi)/(\pi,s)$. Now applying lemma 1 obtain

$$\begin{aligned} D\tilde{Y}_\beta(1) = E\tilde{Y}_\beta^2(1) &= \int \nu(dx)\tilde{\varphi}^2(x) + r_\nu \\ &+ \int \beta^{-1}(x)\pi(dx)[D_Q\tilde{\varphi}(x) + r_Q(x)]/(s,\pi). \end{aligned}$$

Since $T_\beta = E\tilde{T}_{\beta,1}$ the formula for $T_\beta\sigma_{1\beta}^2$ can be obtained by multiplication of the expressions for $D\tilde{Y}_\beta(1)/(Eb_\beta(1))^2$ and $E\tilde{T}_{\beta,1}$.

Proof of theorem 4

The result follows from the Schwarz inequality in the following form

$$\int \beta(x) g_1(x)\pi(dx) \int \beta^{-1}(x) g_2(x)\pi(dx) \le [\int (g_1(x)g_2(x))^{1/2}\pi(dx)]^2$$

Proof of theorem 5

Let $h_x(t) = 1$, $t \ge x$, $h_x(t) = 0$, $t < x$ where x is a fixed value. Denote the functions $\hat{\varphi}_\beta(t)$ and $\varphi(t)$ corresponding to $h(t) = h_x(t)$ by $\hat{\varphi}_{x,\beta}(t)$ and $\varphi_x(t)$, respectively. Using (13) we get that the function $\varphi_1(t) = D\hat{\varphi}_{x,\beta}(t)$ satisfies the equation

$$\varphi_1(t) = \int_0^\infty \varphi_1(u)(\beta(t)/\beta(u)) f(u-t)du + h_1(t)$$

where

$$h_1(t) = F_f\varphi(t) + r(t),$$

$$D_f\varphi(t) = D(\varphi_x(W_2)(W_1 - t) = \int_0^\infty \varphi_x^2(u) f(u-t)du - \left(\int_0^\infty \varphi_x(u) f(u-t)du\right)^2,$$

$$r(t) = \int_0^\infty (\varphi_x(u)\beta(t)/\beta(u))^2 \gamma_{u,t} f(u-t)du,$$

$$\gamma_{u,t} = \overline{\beta(u)/\beta(t)} - (\overline{\beta(u)/\beta(t)})^2.$$

In the paper 15 it is shown that

$$\varphi_x(t) \sim \begin{cases} c_1\theta e^{\lambda_0 t} & t < x,\ t, x \to \infty \\ c_1\theta e^{\lambda_0 t} + (t-x)c_2\theta & t \geq x,\ x \to \infty \end{cases}$$

where c_1 and c_2 are constants,

$$h_1(t) = \begin{cases} e^{2\lambda_0 t} O(\theta^2) & t < x,\ x \to \infty \\ O(1) & t \geq x,\ x \to \infty \end{cases}$$

Therefore

$$D_f\varphi(t) \sim \begin{cases} e^{2\lambda_0 t}(k-1)c^2\theta^2 & t < x,\ x \to \infty \\ e^{2\lambda x}c^2\theta^2 DX_1 & t \geq x,\ x \to \infty \end{cases}$$

Now the results of theorem 5 can be got by immediate calculations.

Proof of theorem 6

The proof is very similar to that of theorem 3.

References

1. Feller, W. (1970). An introduction to probability theory and its appplications, v.2, Wiley, New York.
2. Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. Ann. Stat. 4, 673-684.
3. Kleinen, J.P.C. (1974). Statistical techniques in Simulation. Part I. Dekker, New York.
4. Bratley, P., Fox, B.L. and Schrage, L.E. (1987). A Guide to Simulation. 2nd ed., Springer, Berlin.
5. Fedorov, V.V. (1972). Theory of Optimal Experiments, Academic Press, New York.
6. Ermakov, S.M. (1975). The Monte Carlo method and related questions. Nauka, Moscow (in Russian)
7. Glynn, W.P., Iglehart, D.L. (1988). Simulation methods for queues: an overview. Queueing systems 3, 221-256.
8. Kahn, H. (1956). Use of different Monte Carlo sampling techniques. In: Simposium on Monte Carlo Methods, Wiley, New York, 146-190.

9. Ogybin, V.N. (1967). On applications of splitting and roulette method in Monte Carlo calculation of particles transfer. In: The Monte Carlo method in particles transfer problems. Atomizdat, Moscow, 72-82 (in Russian).

10. G.A. (1987). Weighing Monte Carlo methods optimization. Nauka. Moscow. Ch.7 (in Russian)

11. Ermakov, S.M., Melas, V.B. (1989). On optimal trajectory branching in the simulation of systems descibed by stationary stochastic processes. In: Izvesia Acad. Nauk SSSR. Techical Kybernet. 2, 64-69 (in Russian).

12. Melas, V.B. (1990). On branching technique for increasing efficiency of complex system simulation. Inst. of Kybern. Acad. Nauk Ukr.SSR, Kiev

13. Melas, V.B. (1992). Branching technique for Markov chain simulation (finite state case). Statistics, to be published.

14. Ermakov, S.M., Krivulin, N.K. and Melas, V.B. (1992). Efficient methods of queueing systems simulation. In: Modelling and Simulation, 1991. Proceedings of the 1991 European Simulation Multiconference. Ed. by Erik Mosenkilde, Copenhagen, 9-19.

15. Melas, V.B. (1992). On comparison of importance sampling and branching technique for queueing systems. Stochastic Optimization & Design 3, to be published.

16. Nummelin, E. (1984). General irreducible Markov chains and non-negative operators. Cambridge University Press, Cambridge, London.

17. Crane, A.M., Iglehart, D.L. (1974). Simulating stable stochastic systems. II: Markov chains. J. ACM 21, 114-123.

18. Fishman, G.S. (1983). Accelerated accuracy in the simulation of Markov chains. Oper. Res. 31. 466-487.

PART II. STATISTICAL APPLICATIONS

Estimates with Branching for a Functional of Stationary Distribution of Markov Chain

S.M.Ermakov and J.N.Kashtanov

In this paper some approaches to minimization of estimates variance in Monte-Carlo calculations of a functional of stationary distribution of Markov chain are developed. A computational example from the queueing theory, illustrating an application of the theoretical results, is presented.

1

The problem of calculation of a functional $J = (\pi, h)$ where π is the stationary distribution of some Markov chain with transition probabilities $p(x, dy)$ often meets in queueing theory, statistical physics; some problems of calculation eigenvalues and functionals of eigenfunctions of some intergral and differential operators can be reduced to it [5].

As a rule the analitical solution of problems pointed above is possible only in the simplest cases, therefore functional J is often calculated by Monte-Carlo method with estimate $J_N = \frac{1}{N}\sum_{i=1}^{N} h(\xi_i)$ where ξ_i - is a simulated Markov chain with transition probabilities $p(x, dy)$. In the cases when J is small the estimate J_N becomes not effective and it's necessary to use some methods of reduction of estimate variance. As it was shown in [1,2] the straight use of importance sampling leads to inconsistant estimates and some method of improvement the efficiency was suggested. The essence of the method is that the main chain with transition probabilities $p(x, dy)$ is simulated and auxiliary chains with, generally speaking, arbitrary transition probabilities $r(x, dy)$ are branching from the main chain at some deterministic moments. On the paths of these auxiliary chains estimates of some functional are constructed and transition probabilities $r(x, dy)$ are chosen in some special way in order to decrease the variance. This method was generalized in [3], where, in particular, the estimate with zero variance was constructed. In [4] another approach to variance reduction for estimating J was developed. The essence of this method is that we must know or at least can simulate the distribution of the chain at some moments (for example at the moment of achieving some set). Starting from this distribution we can simulate the chain, generally speaking, with arbitrary transition probabilities until the next moment when we know the chain distribution.

In present work the results of [3,4] are generalized to the case of nonlinear functionals.

2

Let $p(x, dy)$ be transition probabilities in space $(X, \mathcal{B})$ with stationary distribution $\pi(dx)$. We assume that the condition $|p^{(i)}(x, B) - \pi(B)| \leq const\ \delta^i$ is valid, where $B \in \mathcal{B}$, $p^{(i)}$ is the i-th power of p, *const* is suffuciently large constant, $\delta < 1$. Let a measurable bounded function $h : X \to R^n$ and function $u : R^n \to R$, $u \in C^2$ be given and $J = (\pi, h)$. We consider estimates for $u(J)$ with minimal variance.

Let $\{r_i(x, dy)\}$ be substochastic transition probabilities, $m_i(x) = r_i(x, X)$, $g_i(x) = 1 - m_i(x)$,

$$\mathbf{X} = \bigcup_{1 \leq i \leq \infty} X^i,\ \tau_0(\mathbf{x}) = i,\ if\ \mathbf{x} \in X^i,$$

$\mathbf{R}_x$ be probabilities in $\mathbf{X}$ generated by $r_i(x, dy)$, $\mathbf{P}_x$ - those generated by $p(x, dy)$. Mathematical expectation according the measure $\mathbf{R}_x$ we assign as $\mathbf{M}_x$, that, according the measure $\mathbf{P}_x$, - by $\mathbf{E}_x$ and consider transition probabilities in $\mathbf{X}$:

$$\mathbf{R}(\mathbf{x}, \cdot) = \int_X p(x_0, dy_0)\mathbf{R}_{y_0}(\cdot),\ x_0 = \mathbf{x}(0).$$

Transition probabilities $\mathbf{R}(\mathbf{x}, \cdot)$ have a stationary disribution $\Pi(\cdot) = \int_X \pi(dx)\mathbf{R}_x(\cdot)$.

Let $\rho_i = \chi_{(i<\tau_o)}(\mathbf{x}) \prod_{j=1}^{i} \frac{p(x_{j-1}, dx_j)}{r_j(x_{j-1}, dx_j)}$, $\mathbf{A}h(\mathbf{x}) = \sum_{i<\tau_0} \rho_i(\mathbf{x})w_i(\mathbf{x})\alpha_i(x_i)h(x_i)$.

Lemma 1. Let the following conditions be valid:

1. $\alpha_i \geq 0$, $\sum_{i=0}^{\infty} \alpha_i(x) = 1$,
2. $\mathbf{E}_x(\chi_{\mathbf{B}}\alpha_i h(x_i)) \neq 0 \Rightarrow \mathbf{R}_x(\mathbf{B}) > 0$, $\mathbf{B} \subset \{i < \tau_0\}$,
3. $\mathbf{M}_x(w_i|B^i) = 1$,
4. $\sum_{i<\tau_0} |\rho_i(\mathbf{x})w_i(\mathbf{x})\alpha_i(x_i)h(x_i)| \in L^2(\mathbf{R}_{x_0})$,

then $\int_{\mathbf{X}} \Pi(dx)\mathbf{A}h(\mathbf{x}) = J$.

It should be noted that an estimate, satisfying condition 3, is called in Monte-Carlo method a "unite class" estimate, an estimate with $w_i(\mathbf{x}) = \delta_{i,\tau_0-1}/g_{\tau_0}(x_i)$, is called an esimate "by absorptions" and an one with $w_i(\mathbf{x}) = 1$ is called an estimate "by collisions".

It follows from Lemma 1 that as an estimate for $u(J)$ one can take $u(\mathbf{J}_N)$, where

$$\mathbf{J}_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{A}(h + c)(\mathbf{x}_k) - c,$$

$c \in R^n$, $\mathbf{x}_k$ is a Markov chain with transition probabilities $\mathbf{R}(\mathbf{x}, \cdot)$.

Let $v(x) = (u'(J), h(x) + c)$, $I = (u'(J), J + c)$, $\varphi(x) = \mathbf{M}_x\mathbf{A}v$, $\psi(x) = \mathbf{M}_x(\mathbf{A}\mathbf{v})^2$.

Lemma 2. If in neighbourhood of J $|\partial^2 u/\partial y_i \partial y_j| \leq const$, then

$$\mathbf{Prob}(|u(\mathbf{J}_N) - u(J)| < s/\sqrt{N}) \longrightarrow \frac{1}{\sqrt{2\pi d}} \int_{-s}^{s} e^{\frac{-t^2}{2d}} dt,$$

where **Prob** is the measure in the universal probabilistic space, $d = \sum_{k=-\infty}^{\infty} R(|k|)$, $R(k) = \int_X \pi(dx)(\varphi(x) - I)(P_x^k\varphi - I)$, $R(0) = \int_X \pi(dx)\psi(x) - I^2$.

One can note that only $R(0)$ depends on $\{r_i\}$. Thus the problem of variance reduction for the estimate $\mathbf{J}_N$ is reduced to that of minimization of $D = \int_X \pi(dx)\psi(x)$.

We assume that vector c is chosen so that $v(x) \geq v_0 > 0$. The next theorems can be proved in the same way as in [2].

Theorem 1. Let $\alpha_i = \delta_{i,l}$, then for the estimate "by absorptions" the minimum of D is achieving when $r_i(x, dy) = \chi_{(i\leq l)}p(x, dy)P_y^{l-i}v/P_x^{l-i+1}v$.

Theorem 2. Let $\alpha_i = \chi_{(i\leq l)}/(l+1)$, then for the estimate "by collisions" the minimum of D is achieving when

$$r_i(x, dy) = \chi_{(i\leq l)}p(x, dy)\sum_{j=0}^{l-i} P_y^j v/\sum_{j=0}^{l-i} P_x^{j+1}v.$$

Theorem 3. Let $0 < \gamma < 1$, $\alpha_i = \gamma^i(1-\gamma)$, then for the estimate "by absorptions" the minimum of D is achieving when $r_i(x, dy) = \gamma p(x, dy)\varphi(y)/\varphi(x)$.

Theorem 4. Let $0 < \gamma < 1$, $\alpha_i = \gamma^i(1-\gamma)$, $\gamma^2 < c \leq 1$, then for the estimate "by collisions" the minimum of D at condition $m_i(x) \leq c$ is achieving when $r_i(x, dy) = cp(x, dy)b(y)/P_x b$, where b is the nonnegative solution of the equation

$$b^2(x) = \frac{\gamma^2}{c}\Big(\int_X p(x, dy)b(y)\Big)^2 + (1-\gamma)h(x)(2\varphi(x) - (1-\gamma)h(x)).$$

Up to this we considered minimisation of d only by choise of r_i but we can also choose α_i in some special way in order to minimize d. Let us define $\sigma(x) = \sum_{i=0}^{\infty} P_x^i(v - I) + s$ and suppose that s and c are chosen so that $v(x) \geq \sigma(x) > 0$.

Theorem 5. Let $\alpha_1 = \sigma/v$, $\alpha_0 = 1 - \alpha_1$, $\alpha_i = 0$, $i \geq 2$, then for the estimate "by absorptions" $d = 0$.

3

Now we shall consider another approach to constructing the estimates for J.

Let τ be some Markov moment, $\theta^s\mathbf{x}(t) = \mathbf{x}(s+t)$, $\tau_1 = \tau$, $\tau_i = \tau_{i-1} + \tau(\theta^{\tau_{i-1}}(\mathbf{x}))$, $\mathbf{P}_x(\tau > i) \leq const\ \delta^i$. Let Markov chain $\mathbf{x}(\tau_n)$ have the stationary distribution $\pi_\tau(dx)$. We shall assign $H = \sum_{i<\tau} \rho_i w_i h(x_i)$, $H_0 = \sum_{i<\tau} \rho_i w_i$, $\mathbf{M}$ - a sign of mathematical expectation according the measure $\mathbf{R}(d\mathbf{x}) = \int_X \pi_\tau(dx_o)\mathbf{R}_{x_o}(dx)$, $\mathbf{E}$ - that according the measure $\int_X \pi_\tau(dx_o)\mathbf{P}_{x_o}(dx)$.

Lemma 3. Under conditions stated above $J = \mathbf{M}H/\mathbf{M}H_0$.

So if we assume $\mathbf{J}_N = \sum_{i=1}^{N} H(\mathbf{x}_k) / \sum_{i=1}^{N} H_0(\mathbf{x}_k)$, where $\mathbf{x}_k$ are independent realizations with values in $\mathbf{X}$ and distribution $\mathbf{R}(dx)$ then one can take $u(\mathbf{J}_N)$ as an estimate for u(J). Let $f(x) = (u'(J), h(x) - J)$, $\varphi(x) = \mathbf{E}_x \sum_{i<\tau} f(x_i)$ then

$$\mathbf{Prob}(|u(\mathbf{J}_N) - u(J)| < s/\sqrt{N}) \longrightarrow \frac{1}{\sqrt{2\pi d}} \int_{-s}^{s} e^{\frac{-t^2}{2d}} dt,$$

where $d = (\mathbf{E}\tau)^{-2}\mathbf{M}(\sum_{i<\tau} \rho_i w_i f(x_i))^2$.

Let C be a subset of X and τ be the first value of $i > 0$ with $\mathbf{x}_i \in C$, $k(x, dy) = p(x, dy)\chi_{X \setminus C}(y)$.

Theorem 6. For the estimate "by absorptions" the minimum of d is achieving when $r(x, dy) = \chi_{(\bar{\varphi}(x)\neq 0)} k(x, dy)\bar{\varphi}(y)/\bar{\varphi}(x)$, where $\bar{\varphi}(x) = \mathbf{E}_x \sum_{i<\tau} |f(x_i)|$.

Theorem 7. Let $\delta^2 < c_0 \leq c(x) \leq 1$ then the minimum of d for the estimate "by collisions" under condition $m(x) \leq c(x)$ is achieving when $r(x, dy) = \chi_{(K_x b \neq 0)} c(x) k(x, dy) b(y)/K_x b$, where b is the nonnegative solution of the equation

$$b^2(x) = \frac{1}{c(x)}(K_x b)^2 + \varphi^2(x) - (K_x \varphi)^2.$$

4

The following example is taken from [1].

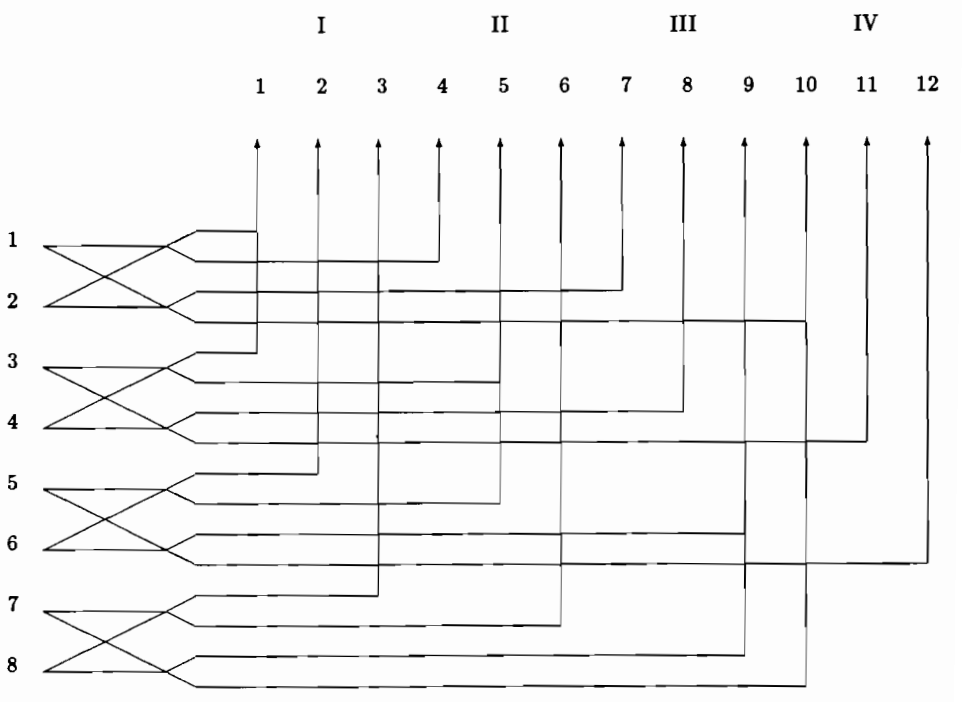
The Poisson flow of bids with summary rate equal to 7 is uniformly distributed between entries (1-8). Every bid chooses with equal probability one of four directions (I-IV) and occupies the corresponding channel (1-12). The attended time has the exponential distribution with mean equal to 1. If a channel is already occupied then the bid, requesting for this channel can not be served and leaves the system. One needs to calculate the probability p of the event that more than 8 channels are occupied.

Let us introduce the next assignments: x is a state of the system, $l(x)$ is a number of bids in system in the state x, $h_1(x) = 1/(l(x) + 7)$, $h_2(x) = \chi_{(l(x)>8)}(x)/(l(x) + 7)$, $h = (h_1, h_2)$, π is the stationary distribution, $J = \sum_x \pi(x)h(x)$, then $p = u(J) = J_2/J_1$.

In order to calculate approximate values of α_i let us integrate the states. We shall define an integrated state as a level set of function $l(x)$ and introduce the transition matrix $\bar{p}_{m,k}$ in the same way as in [2]. We can easily find $\bar{\alpha}_i[k]$, $\bar{b}[k]$ for $\bar{p}_{m,k}$ and then assume $\alpha_i(x) = \bar{\alpha}_i[l(x)]$, $b(x) = \bar{b}[l(x)]$.

The results of simulations are given below.

	Average	Spread of 95% confidence interval
straight sumulation	$4.63 \cdot 10^{-3}$	$2.42 \cdot 10^{-3}$
estimate of theorem 5	$5.50 \cdot 10^{-3}$	$2.98 \cdot 10^{-4}$
estimate of theorem 7	$5.67 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$



References

1. Linnik I.J. (1972) *'On Improvment of Convergence Rate of Monte-Carlo Method in Calculation of Queueing Systems' Parameters'* , in : Probabilistic Methods in Solving the Mathematical Physics Problems, Novosobirsk, pp. 156 - 164. (in Russian).

2. Linnik I.J. (1973) *'Improvment of Convergence of Monte-Carlo Method in Some Problems of Queueing Theory'* , Kibernetika, N 5, pp.129-132. (in Russian).

3. Kashtanov J.N. (1980) *'Improvement of Convergence in Calculation of Functionals of Stationary Distribution of Markov Chain by Monte-Carlo Method'* , Vestnik of Leningrad University, N 13, pp. 34 - 40. (in Russian).

4. Kashtanov J.N. (1981) *'Some methods of variance reduction in Monte-Carlo Estimating of Functionals of Stationary Distribution of Markov Chain'* . Vestnik of Leningrad University, N 7, pp. 42 - 49. (in Russian).

5. Kashtanov J.N. (1987) *'A Rate of Convergence to Eigenmeasure of Integral Operator in Generations Method'* , Vestnik of Leningrad University, N 1, pp.17-21.(in Russian).

Optimized Moving Local Regression: Another Approach to Forecasting

Valery V. Fedorov, Peter Hackl and Werner G. Müller

This paper empirically demonstrates the relative merits of the optimal choice of the weight function in a moving local regression as suggested by Fedorov et al. (1993) over traditional weight functions which ignore the form of the local model. The discussion is based on a task that is imbedded into the smoothing methodology, namely the forecasting of business time series data with the help of a one-sided moving local regression model.

1 Introduction

In the moving local regression approach parameters are estimated by weighting down the observations so that the weights reflect the "distance" of the observations from the forecast point. This gives the flexibility to parametrize the model depending on local conditions. Given that the true model is locally approximated and a certain form of the approximation error (such as the remainder term of a local series expansion) is suspected to be relevant at times, it is possible to choose the weights such that optimal forecasting power is achieved. Such models are particularly useful for describing or forecasting time series that are generated by time-varying processes.

In the literature several suggestions for the choice of the weight function in moving local regression models can be found [e.g. McLain (1971), Cleveland (1979)]. A common feature of these weighting schemes is that they are chosen taking no regard of the model specification. The approach presented here aims at maximizing the forecast accuracy and takes a possible model misspecification into account.

In Section 2 the model and the estimation method are introduced. Section 3 presents three weight functions that are to be compared in Section 4. This comparison is based on a time series from bank business that is a typical candidate for nonparametric analysis.

2 The method

Let $\{x_1, \ldots, x_T\}$ be a given set of supporting points, i.e., points where observations $\{y_1, \ldots, y_T\}$ are available, and let $d_t = x_{T+1} - x_t, \quad t = 1, \ldots, T$, be the "distances" from the point of interest x_{T+1}. Then

$$y_t = \theta^T f(d_t) + \delta^T \varphi(d_t) + \varepsilon_t, \quad t = 1, \ldots, T \tag{2.1}$$

will be called a one-sided regression model. It consists of a main term $\theta^T f(d_t)$ describing the model, that locally approximates the true model, a "nuisance term" $\delta^T \varphi(d_t)$ describing

the approximation error, and an error term ε_t following the usual assumptions $E[\varepsilon_t] = 0$ and $E[\varepsilon_t \varepsilon_{t'}] = \sigma_\varepsilon^2$ if $t = t'$ and 0 otherwise. The number of components of the parameter vector θ is determined by the structure of the approximating model. For φ, an appropriate function has to be specified; the "nuisance parameter" δ is unknown.

Setting $t = T + 1$ in (2.1) allows us to calculate a forecast for y_{T+1}. If we make the reasonable

> **Assumption:** $f_1(d) \equiv 1$, $f_j(d) \to 0$ for $d \to 0$ and $j \geq 2$, and all components of $\varphi(d)$ also vanish [usually faster than $f_j(d)$] for $d \to 0$,

the forecast

$$\hat{y}_{T+1} = \hat{\theta}_1 \tag{2.2}$$

is the first component of the (weighted least squares) estimator

$$\hat{\theta} = M^{-1} Y,$$

with $M = \sum_{t=1}^{T} \lambda(d_t) f(d_t) f^T(d_t)$ and $Y = \sum_{t=1}^{T} \lambda(d_t) f(d_t) y_t$.

The mean squared error matrix of the estimator $\hat{\theta}$ is

$$R = \mathrm{E}\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\} = M^{-1} M_{12} \delta \delta^T M_{12}^T M^{-1} + \sigma^2 M^{-1} \mathcal{M} M^{-1}, \tag{2.3}$$

where $M_{12} = \sum_{t=1}^{T} \lambda(d_t) f(d_t) \varphi^T(d_t)$ and $\mathcal{M} = \sum_{t=1}^{T} \lambda^2(d_t) f(d_t) f^T(d_t)$. The choice of the weight function $\lambda(d_t)$ which reflects the reliability of the local approximation is discussed in the subsequent section.

$\hat{\theta}$ is generally biased:

$$\mathrm{E}\{\hat{\theta}\} = \theta + M^{-1} M_{12} \delta, \tag{2.4}$$

A detailed treatment of the estimation properties is given in the nonparametric regression literature such as Cleveland & Devlin (1988) or Buja et al. (1989).

Models of type (2.1) used in local fitting are particularly helpful for time series whose characteristics change over time. For cases where higher order terms reflected by $\delta^T \varphi(d_t)$ are suspected to have some effect, Fedorov et al. (1993) suggest choosing the weight function $\lambda(d_t)$ so that a suitably chosen scalar function of the m.s.e. matrix R is minimized. Adapted to the forecasting problem, this means direct minimization of the mean square error of the forecast $\hat{\theta}_1$. It is performed under the restriction $\lambda(d_t) \geq 0$ for all d_t and $\sum_t \lambda(d_t) = 1$. The weight function depends on the nuisance parameter δ and the location of observations (i.e. d_t, $t = 1, \ldots, T$). Therefore, in its derivation in a particular situation, δ has to be estimated in a preliminary step. The weight function is entirely determined by the model specification and the data.

In a forecasting context this method will be sequentially applied, i.e., forecasts are calculated for time points $T + 1, T + 2, \ldots$, each estimate being based on the currently available amount of information. This implies that the weight function is derived in each forecast point anew. This generalization of the estimation process is straightforward and so we do not record the corresponding formulae.

Example 1 *As an illustration, the optimal weight function is derived for the model* $y_t = \theta + \delta d_t^2 + \varepsilon_t$, *i.e. the moving average specification with a quadratic "nuisance" term. We consider a collection of* T *equally spaced points in the interval* $[-1, \frac{T-2}{T}]$ *and derive the value*

of the weight function for point 1. *For the average quadratic distance* $\bar{d}^2$ *and its variance we obtain* $\frac{2T(T+1)(2T+1)}{3T^3}$ *and* $\frac{4T^2(2T+1)(T+1)(8T^2+3T-11)}{45T^6}$ *respectively. The optimal weights are:*

$$\lambda(d_t) = \frac{1}{T} - \frac{\frac{\delta^2}{\sigma_t^2}[d_t^2 - \bar{d}^2]\bar{d}^2}{1 + T\frac{\delta^2}{\sigma_t^2}\,\mathrm{var}(d_t^2)}, \tag{2.5}$$

c.f Fedorov et al. (1993). Note that they are linear in d_t^2.

The form of the weight function and the number of supporting observations that have nonzero weights (the "window width"), and consequently the degree of smoothing crucially affects the estimate $\hat{\theta}$. A weight function that is too concentrated around the forecast point results in undue variation as it allows reaction to local time series characteristics; a too flat weight function smoothes out local tendencies.

The use of moving averages, i.e., application of the model from Example 1, is suitable for the description of the long wave changes in a time series but smoothes away short term effects. Using a linear moving regression that includes the term θd allows us to identify changes which occur within the period covered by the weight function.

3 Comparison of weight functions

When applying moving regression to a set of time series that differ considerably with respect to its characteristics, the smoothing interval has to be long enough to cover the longest period of changes in these characteristics.

In the literature several recommendations for the choice of the weight function are given. Out of practical considerations McLain (1971) suggested

$$\lambda(d) = \exp\frac{-\|d\|^2/{d_n}^2}{\|d\|^2 + \rho}, \tag{3.1}$$

where d_n is the average distance between neighbouring data points and the constant $\rho = 10^{d_n} - 1$ prevents numerical accuracy problems. A computationally simpler function, the so-called tricube,

$$\lambda(d) = \begin{cases} [1 - (\|d\|/d_q)^3]^3 & 0 \le \|d\|/d_q \le 1 \\ 0 & \text{else} \end{cases} \tag{3.2}$$

with d_q being the distance of the $q.n$ nearest point to x, is used by Cleveland (1979). This function smoothly decreases from 1 to 0 with increasing $\|d\|$. The weight functions (3.1) and (3.2) have in common that they are chosen without regard of the local model, and the possibility of a nuisance term is neglected.

Following the recommendations by Fedorov et al. (1993) the weight function can be chosen such that the mean squared error matrix R [see (2.3)] is minimized in a certain sense. In general, this approach should be clearly superior to techniques that are based on weight functions such as (3.1) or (3.2). A demonstration of the relative capabilities in applications will be given in the following section by means of an example in which a forecast of bank account data is required.

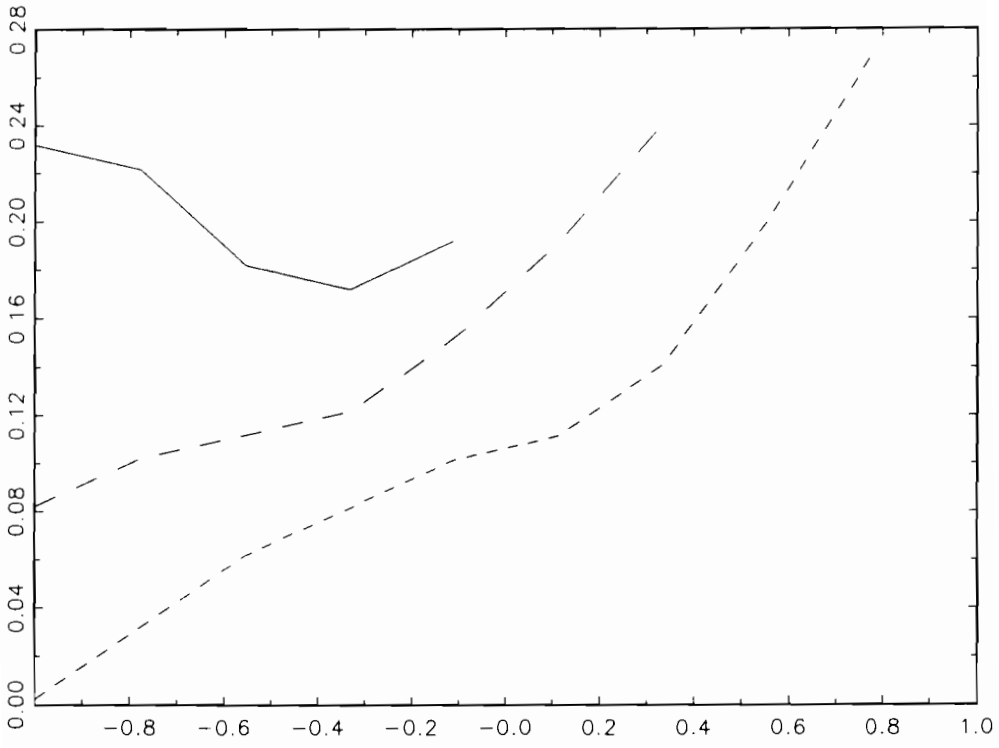


Figure 1: Optimal weights for forecasting at $x = 0$ (solid), $x = 0.5$ (dashed) and $x = 1$ (short dashed).

Example 2 *Let $y_1, \ldots, y_T$ be observations from locations $-1 \leq x_1 < \ldots < 0 < \ldots < x_T \leq 1$ symmetrically arranged around 0. The aim is to get a prediction $\hat{y}$ at the forecast point x. If a linear model with a quadratic nuisance term (cf. next section) is assumed, the optimal weights λ^* for $T = 10$ and $x = 0$, $x = 5$ and $x = 1$ are shown in Figure 1.*

4 A case study

For comparing various weight functions the model

$$y_t = \theta_1 + \theta_2 d_t + \delta d_t^2 + \epsilon_t, \quad t = 1, \ldots, T \tag{4.1}$$

was chosen. It implies that linearity is considered as a suitable description of the local behaviour, and that a possible effect of a quadratic term is allowed to be corrected via the weights of the local regression.

The comparison is based on a time series from the bank business, which is given in Table 1. The data analyzed in the example are the fractions, to which the creditline of a typical small Austrian enterprise is used, observed weekly over a period of 14 months, a 100% exhausted

1	1.092860	46	1.136138	91	1.173454	136	0.881794	181	0.709575
2	1.113694	47	1.146138	92	1.173454	137	1.131522	182	0.801418
3	1.092860	48	1.146138	93	1.173454	138	1.131522	183	0.859931
4	1.092860	49	1.146138	94	1.190121	139	1.131522	184	0.859931
5	1.092860	50	1.146192	95	1.200121	140	1.131522	185	0.826418
6	1.092860	51	1.149839	96	1.200121	141	1.131522	186	0.801418
7	0.821910	52	1.149839	97	1.172621	142	1.131522	187	0.801418
8	1.251459	53	1.149839	98	1.172621	143	1.131522	188	0.837006
9	0.914887	54	1.149839	99	1.171991	144	1.131522	189	0.921416
10	0.948220	55	1.174839	100	1.171991	145	1.168189	190	0.834855
11	0.914887	56	1.149839	101	1.171991	146	1.168189	191	0.836679
12	0.914887	57	1.174839	102	1.297045	147	1.168189	192	0.836679
13	0.914887	58	1.312339	103	1.172045	148	1.131522	193	0.837929
14	0.914887	59	1.174947	104	1.072271	149	1.133346	194	0.837929
15	0.914887	60	1.174947	105	1.172153	150	0.885651	195	0.837929
16	0.914887	61	1.174947	106	1.173977	151	1.133346	196	0.837929
17	0.935720	62	1.174947	107	1.173977	152	1.133346	197	0.837929
18	0.935720	63	1.205920	108	1.173977	153	1.133346	198	0.895419
19	1.131808	64	1.205920	109	1.173977	154	0.805013	199	0.883762
20	1.131808	65	1.205920	110	1.173977	155	0.805013	200	0.837929
21	1.131808	66	1.241076	111	1.173977	156	0.805013	201	0.754596
22	1.131808	67	1.207743	112	1.173977	157	0.805013	202	0.754596
23	1.256808	68	1.207743	113	1.270315	158	0.843346	203	0.754596
24	1.131808	69	1.207743	114	1.174031	159	0.805013	204	0.754596
25	1.131808	70	1.207410	115	1.174031	160	0.841679	205	0.775429
26	1.131808	71	1.207410	116	1.174031	161	0.841679	206	0.754596
27	1.131808	72	1.207410	117	1.174031	162	0.841679	207	0.754596
28	1.131808	73	1.207410	118	1.174031	163	0.805013	208	0.754596
29	1.131808	74	1.207410	119	1.260698	164	0.905013	209	0.754596
30	1.131808	75	1.207410	120	1.210698	165	0.805013	210	0.754596
31	1.131808	76	1.234076	121	1.210698	166	0.805013	211	0.789596
32	1.131808	77	1.234076	122	1.199281	167	0.805013	212	0.756419
33	1.131808	78	1.234076	123	1.174031	168	0.805013	213	0.847993
34	1.131808	79	1.171576	124	1.325726	169	0.805013	214	0.768086
35	1.131808	80	1.171576	125	1.212815	170	0.805013	215	0.756419
36	1.152641	81	1.197622	126	1.212815	171	0.806836	216	0.824753
37	1.055290	82	1.171631	127	1.214639	172	0.806836	217	0.824753
38	1.177804	83	1.171631	128	1.214639	173	0.806836	218	0.758086
39	1.136138	84	1.171631	129	1.131306	174	0.806836	219	0.758086
40	1.136138	85	1.171631	130	1.131468	175	0.806836	220	0.658086
41	1.136138	86	1.173454	131	1.131468	176	0.870484	221	0.685583
42	1.136138	87	1.173454	132	1.131468	177	0.806836	222	0.699753
43	1.136138	88	1.200121	133	1.131468	178	0.862669	223	0.711774
44	1.136138	89	1.173454	134	1.131522	179	0.834751	224	0.658086
45	1.261138	90	1.173454	135	1.131522	180	0.834751	225	0.658086

Table 1: Fractions y_t to which a creditline is used at time t.

226	0.658086	239	0.659909	252	0.686790	265	0.415400	278	0.419733
227	0.658086	240	0.659909	253	0.688613	266	0.415400	279	0.466630
228	0.658086	241	0.659909	254	0.688613	267	0.415400	280	0.468936
229	0.658086	242	0.659909	255	0.688613	268	0.415983	281	0.481679
230	0.658086	243	0.659909	256	0.688613	269	0.445566	282	0.481679
231	0.658086	244	0.659909	257	0.688613	270	0.445566	283	0.481679
232	0.741419	245	0.659909	258	0.563613	271	0.435733	284	0.511262
233	0.658086	246	0.659909	259	0.563613	272	0.414900	285	0.444596
234	0.659909	247	0.659909	260	-0.06570	273	0.414900	286	0.444596
235	0.659909	248	0.660623	261	0.416113	274	0.417909	287	0.444596
236	0.659909	249	0.661669	262	0.415400	275	0.419733	288	0.444596
237	0.659909	250	0.661669	263	0.415400	276	0.419733	289	0.450180
238	0.659909	251	0.686790	264	0.415400	277	0.419733	290	0.450180
								291	0.450180

Table 2: Fractions y_t to which a creditline is used at time t, continued.

creditline gives a value of 1 in the corresponding series. The bank utilizes these fractions to decide whether the credit should be prolonged or not.

As a first step moving averages were constructed for all possible window lengths (from 5 to 290 days) and all past time points. They can be interpreted as the simplest one step ahead forecasts. The forecasts with the lowest average squared forecast error, corresponding to a window length of 65 days, were used as a reference point for the comparison, as well as for the preestimation of the residual variance $\hat{\sigma}_\epsilon^2$, which gave 0.0172.

Next, minimal average squared forecast errors were found for weight functions (3.1) and (3.2). In applying the numerical algorithm for weight optimization from Fedorov et al. (1993), for simplification of the calculation process we firstly assumed that δ is constant over time. The estimate $\hat{\delta}$ obtained in a preliminary analysis turned out to be 5.8×10^{-4}. For comparison of the results from the three weighting regimes one has to define a common measure of smoothness. Simple to calculate is the sum of squared second differences as an estimate of the local curvature, which is commonly used for penalizing in spline regression.

The data and optimal forecasts are displayed in Figure 4. Figure 2 presents the average squared forecast error for the alternative weighting procedures.

The proposed method with the "optimal" weight function is clearly superior to alternative weighting schemes. The average squared forecast error over all time points lies uniformly below the respective errors for the forecasts using weight functions (3.1) or (3.2) for comparable smoothness levels greater than 0.15. Moreover, its minimum value is 0.0151, which is considerably below (around 6%) the minimum values of 0.0160 and 0.0165 for (3.1) and (3.2), respectively.

Alternatively, to avoid the assumption of constancy in δ,we applied a two step procedure. In the first step a moving quadratic regression was performed to preestimate δ for each forecast point. Using those estimates in the weight optimizing procedure result in an average squared forecast error of 0.01440, another improvement of around 5%.

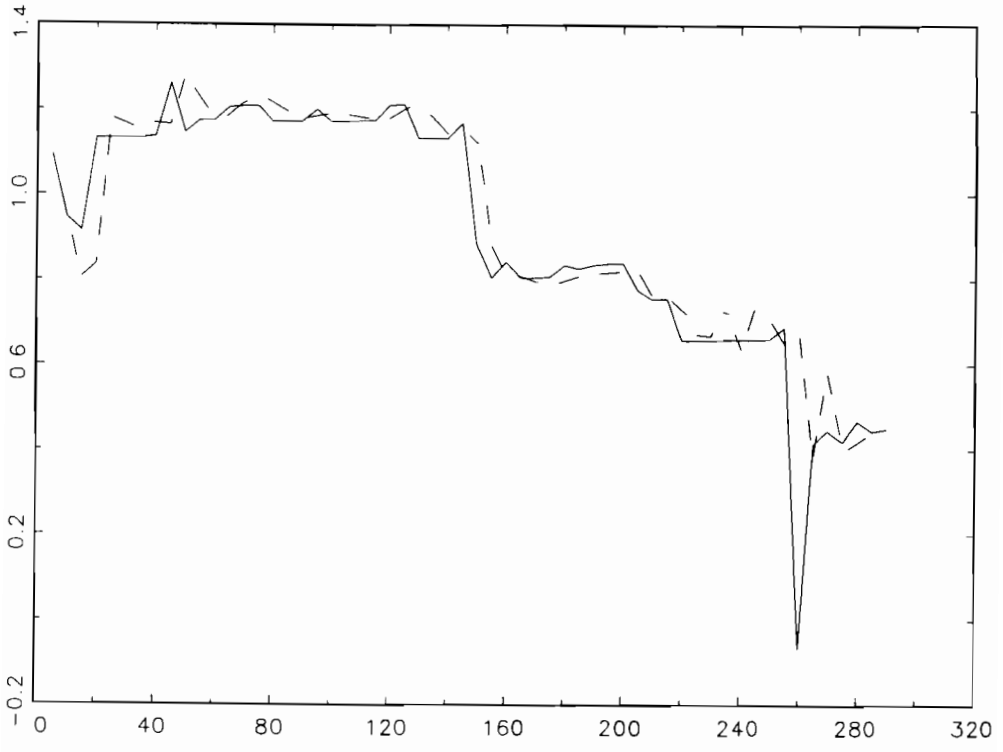


Figure 2: Fractions y_t to which a creditline was used (solid) and optimal forecasts $\hat{y}_t$ by time t (in days).

5 Conclusions

The comparison of forecast errors obtained by the optimized moving local regression approach and two traditional weighting schemes indicates a clear superiority of the former technique. This superiority strongly supports the choice of this technique in this and similar applications. Of course it has to be noted that for cases where the assumed model does not hold the different weighting schemes compete on the same level and one might then perform accidentally better than another.

In addition to the forecasts a lot of valuable information can be gained from the estimators. A continuously performed discriminant analysis for instance allows various enterprises to be distinguished by their economic status. A related example utilizing such an approach is presented by Müller (1992).

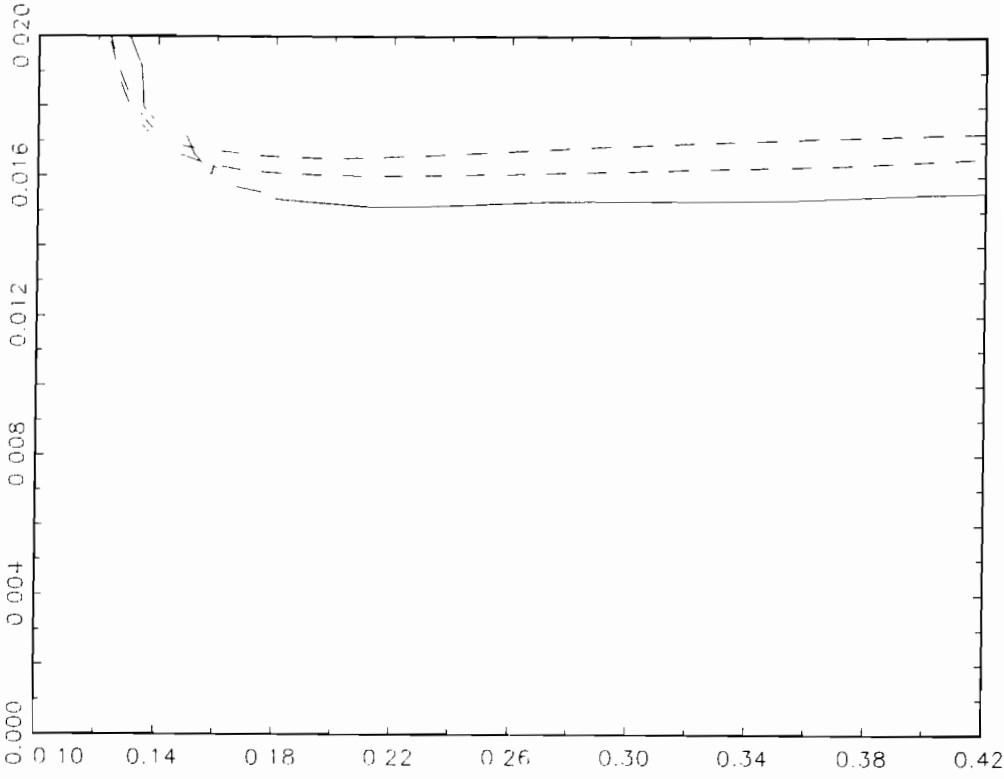


Figure 3: Average forecast error vs. smoothing level: dot-dashed - for weight function (3.1), dashed - (3.2), solid - optimized weight function.

References

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models with discussion. *The Annals of Statistics*, 17(2):453–555, 1989.

W.S. Cleveland and S. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

V.V. Fedorov, P.Hackl, and W.G. Müller. Moving local regression: the weight function. *Journal of Nonparametric Statistics*, in preparation, 1993.

D.H. McLain. Drawing contours from arbitrary data points. *The Computer Journal*, 17(4):318–324, 1971.

W.G. Müller. The evaluation of bank accounts using optimized moving local regressions. In L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz, editors, *Advances in GLIM and Statistical Modelling*, Springer Verlag, 1992.

Sliding Window Polynomial Smoothing of Correlated Data

A.V.Makshanov

New algorithm of adaptive polynomial smoothing of the regular component of time series with on-line identification of local noise spectrum is proposed.

1 Fixed-point fixed-memory filtering

Assume that given empirical dependence locally, on the segment $[\tau - l, \tau + l]$, satisfies a model relation:

$$y(s) = P_m(s) + \xi(s), s = \tau - l, \tau - l + 1, ..., \tau + l,$$

$P_m(s)$ being a polynomial of degree m :

$$P_m(s) = \sum_{j=0}^{m} a_j t^j, t = s - \tau,$$

$\xi(s)$ being a stationary ergodic discrete random process that satisfies autoregressive model of given order $p << l$:

$$\xi(s) = \rho_1 \xi(s-1) + ... + \rho_p \xi(s-p) + \eta(s),$$

$$\eta(s) \in WN(0, \sigma^2)$$

Coefficients $a_0, ..., a_m$, autoregressive parameters $\rho_1, ..., \rho_p$ and intensity of income white noise σ^2 are unknown but constant parameters to be estimated. Further on we are to be interested not with the parameters themselves but with the value of smoothing polynomial $P_m(t)$ at some inner point of the segment $[\tau - l, \tau + l]$, as a rule – at its centre $s = \tau (t = 0)$

Assume that autoregressive process $\xi(t)$ satisfies stability condition: all the roots of the polynomial $Q(z) = 1 - \rho_1 z - ... - \rho_p z^p$ are outside unit circle, so spectral density of $\xi(t)$ is of the form

$$S(\omega) = \frac{1}{2\pi} \frac{1}{|Q(e^{-i\omega})|^2} \tag{1}$$

For estimating unknown parameters $a_0, ..., a_m, \rho_1, ..., \rho_p$ metod of Mann and Wald [1,2] is available. We transform:

$$\eta(t) = \xi(t) - \rho_1 \xi(t-1) - ... - \rho_p \xi(t-1) =$$

$$y(t) - P_m(t) - \rho_1 [y(t-1) - P_m(t-1)] - ...$$

$$- \rho_p [y(t-p) - P_m(t-1)] = \sum_{i=1}^{p} \rho_i y(t-i) - \sum_{j=0}^{m} \gamma_j t^j. \tag{2}$$

Now coefficients $\rho_1, ..., \rho_p, \gamma_0, ..., \gamma_m$ may be found by least-squares algorithm applied to linear model

$$Y = A\Theta + H$$

where

$$Y = (y(\tau - l), y(\tau - l + 1), ..., y(l))^T;$$

$$\Theta = (\rho_1, ..., \rho_p, \gamma_0, ..., \gamma_m)^T;$$

$$H = (\eta(\tau - l), \eta(\tau - l + 1), ..., \eta(\tau + l))^T;$$

$$A = (A_1, A_2)^T;$$

$$A_1 = y(k - i), \tau - l <= k <= \tau + l, 1 <= i <= p;$$

$$A_2 = k^j, \tau - l <= k <= \tau + l, 0 <= j <= m. \quad (3)$$

The least-squares estimate $\hat{\Theta} = (\hat{\rho_1}, ..., \hat{\rho_p}, \hat{\gamma_0}, ..., \hat{\gamma_m})^T = (A^T A)^{-1} A^T Y$ is asymptotically-normal with mean Θ and covariance matrix $\sigma^2 \Sigma$,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} = (A_T A)^{-1} \quad (4)$$

The estimation of intensity σ^2 may be easily found as follows:

$$\hat{\sigma^2} = (2l + 1 - m - p)^{-1} \sum_{k=-l}^{l} \hat{\eta^2}(\tau + k),$$

$$\hat{\eta(t)} = y(t) - \sum_{i=1}^{p} \hat{\rho_i} y(t - i) - \sum_{j=0}^{m} \hat{\gamma_j} t^j. \quad (5)$$

Now it is easy to see that parameters $\gamma_0, ..., \gamma_m$ are connected with parameters $a_0, ..., a_m$ by the following system of equations:

$$\gamma_s = a_s \delta_0 + \sum_{k=1}^{m-s} (-1)^{k+1} C_{s+k}^k \delta_k, s = 0, ..., m, \quad (6)$$

where $\delta_0 = 1 - \rho_1 - ... - \rho_p; \delta_j = \rho_1 + 2^j \rho_2 + ... + p^j \rho_p, 0 < j <= m.$

The estimate $\hat{a}$ of the vector $a = (a_0, ..., a_m)$ is to be found from the system of equations that may be obtained from (6) by replacing γ_i, δ_j with their estimates $\hat{\gamma_i}, \hat{\delta_j}$ based on estimates $\hat{\rho_1}, \hat{\rho_p}$. The matrix of such a system coincides with the matrix B that is to be defined below.

2 Covariance matrix estimation

Let's introduce vectors $\rho = (\rho_1, ..., \rho_p)^T, \gamma = (\gamma_0, ..., \gamma_m)^T, a = (a_0, ..., a_m)^T, \delta = (\delta_0, ..., \delta_m)^T$ and their estimates $\hat{\rho}, \hat{\gamma}, \hat{a}, \hat{\delta}$. Regarding matrices $F = \frac{d\delta}{d\rho}(\hat{\rho}); B = \frac{d\gamma}{da}(\hat{\delta}) = (b_{sj})_{s,j=0}^m; C = \frac{d\gamma}{d\delta}(\hat{a}) = (C_{sk})_{s,k=0}^m$, we can see that

$$F = \begin{bmatrix} -1 & -1 & \dots & -1 \\ 1 & 2 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ 1 & 2^m & \dots & \rho^m \end{bmatrix};$$

$$b_{ss} = \hat{\delta}_0;$$

$$b_{sj} = \begin{cases} (-1)^{j-s+1} C_j^s \hat{\delta}_{j-s} & \text{if } j >= s+1 \\ 0 & \text{if } j < s \end{cases};$$

$$c_{s0} = \hat{a}_s;$$

$$c_{sk} = \begin{cases} (-1)^{k+1} C_{s+k}^k \hat{a}_{s+k} & \text{if } l \le k \le m-s \\ 0 & \text{if } k > m-s \end{cases}; \qquad (7)$$

In these notations we obtain from (6) that

$$d\hat{\gamma} = CFd\hat{\rho} + Bd\hat{a},$$

so

$$d\hat{a} = B^{-1}d\hat{\gamma} - B^{-1}CFd\hat{\rho} = Hd\hat{\gamma} - Gd\hat{\rho}$$

The latter relation supplies an expressionfor estimate of the covariance matrix of the vector a, see (4),(5):

$$cov\hat{a} = \hat{\sigma^2}(H\Sigma_{22}H^T + G\Sigma_{11}G^T - H\Sigma_{12}^T G^T - G\Sigma_{12}H^T). \qquad (8)$$

3 Sliding memory recurrent least squares

Assume that least squares estimate $\hat{\Theta}(n, n+r)$ in the model (3) is found on the basis of observations $y(n+1), ..., y(n+r)$ using rows of A from $(n+1)$ up to $(n+r)$. Denoting corresponding arrays as $Y(n, n+r), A(n, n+r)$ and regarding the matrix $\Sigma(n, n+r) = [A^T(n, n+r)A(n, n+r)]^{-1}$, we have

$$\hat{\Theta}(n, n+r) = \Sigma(n, n+r)A^T(n, n+r)Y(n, n+r) \qquad (9)$$

It's well known how to bring this estimate to recurrent form by organazing the new observation introducing operator:

$$\hat{\Theta}(n, n+r+1) = \hat{\Theta}(n, n+r) + K^{(1)}(n, n+r)[y(n+r+1) - a^T(n+r+1)\hat{\Theta}(n, n+r)], \quad (10)$$

$a^T(n+r+1)$ being the row number $(n+r+1)$ of $A, K^{(1)}$ the transition coefficient to be calculated as follows:

$$K^{(1)}(n,n+r) = \Sigma(n,n+r)a(n+r+1)[I + a^T(n+r+1)\Sigma(n,n+r)a(n+r+1)]^{-1}. \quad (11)$$

where

$$\Sigma(n,n+r+1) = \Sigma(n,n+r) - K^{(1)}(n,n+r)a^T(n+r+1)\Sigma(n,n+r) =$$

$$[I - K^{(1)}(n,n+r)a^T(n+r+1)]\Sigma(n,n+r). \quad (12)$$

Using similar demonstrations we may organize for the estimate (9) the old observation rejecting operator:

$$\hat{\Theta}(n+1,n+r+1) = \hat{\Theta}(n,n+r+1) - K^{(2)}(n,n+r+1)[y(n+1) - a^T(n+1)\hat{\Theta}(n,n+r+1)], \quad (13)$$

where

$$K^{(2)}(n,n+r+1) = \Sigma(n,n+r+1)a(n+1)[I - a_{n+1}^T\Sigma(n,n+r+1)a(n+1)]^{-1}. \quad (14)$$

Such estimates may be now transformed to sliding-window form by using the formula:

$$\Sigma(n+1,n+r+1) = \Sigma(n,n+r+1) + K^{(2)}(n,n+r+1)a^T(n+1)\Sigma(n,n+r+1) =$$

$$[I + K^{(2)}(n,n+r+1)a^T(n+1)]\Sigma(n,n+r+1), \quad (15)$$

that makes possible repeating the cycles of introducing new observations and rejecting (discounting) old observations as many times as is necessary. Similar procedures are regarded, for example, in [3].

References

[1] Rao C.R.(1968). Least squares theory using an estimated dispersion matrix and its application to the measurement of signals. In Proceedings of the 5th Berkeley symposium, 355 - 372.

[2] Mann H.B., Wald A.(1943). On the statistical treatment of linear stochastic difference equations. Econometrica 11, 173 - 181.

[3] Jazwinski A.H. (1970). Stochastic processes and filtering theory, 573.

The Extrapolation Problem of Stationary Time Series Correlation

V.N.Fomin

1 Introduction

The restoration of stochastic process characteristics with help of a finite process realization is an important problem of the mathematical statistics. To be more precise, let us formulate such a problem in the following manner: some lag values of the correlation function of the time series stationary component are known and we need to estimate the corresponding spectral density. Like this we have so called problem of spectrum estimation [1]. This problem usually may be solved in different ways, and every decision one may interpreted as some correlation function extrapolation.

To formulate problem statement we make some following assumptions about time series under consideration are made.

2 Assumption about time series

Let $y = \{y_t : t \in \mathbf{Z},\ y_t \in \mathbf{R}\}$ be stochastic time series with the following structure

$$y = \mu_t + \eta_t,\ t \in \mathbf{Z},\ \eta_t \in \mathbf{M}y_t.$$

Here $\mu = \{\mu_t, t \in \mathbf{Z}\}$ is the process drift (it is a deterministic time series), $\eta = \{\eta_t,\ t \in \mathbf{Z}\}$ is the stationary component of y : $\mathbf{M}\eta_t \equiv 0$, $\mathbf{M}\eta_t\eta_t, = \sigma^2 R(t - t')$, $R(.)$ is the normalized $(R(0) = 1)$ covariance (or correlation) function of the time series y stationary component η. We'll suppose that $R(.)$ is square summable function:

$$\sum_{t=-\infty}^{+\infty} [R(t)]^2 < \infty.$$

It allows us to introduce the function

$$G(\lambda) = \sum_{t=-\infty}^{+\infty} \lambda^t R(t) \tag{1}$$

which is known as the spectral density (of the time series y stationary component η).

3 The problem statement

Let $\tilde{R}(.) = \{\tilde{R}(t),\ t \in \mathbf{Z}, \tilde{R}(s) = R(s),\ |s| \leq T\}$ be admissible correlation function (it means that lag values $\tilde{R}(s)$, $|s| \leq T$, are known, $R(t) = R(-t), t \in \mathbf{Z}$, and $R(t)$ is a positive function). A function

$$\tilde{G}(\lambda) = \sigma^2 \sum_{t=-\infty}^{+\infty} \lambda^t \tilde{R}(t) \tag{2}$$

with an admissible correlation function $\tilde{R}(.)$ will be called the admissible spectral density.

The problem under consideration now may be formulated in the following manner: What an admissible spectral density is acceptable ?

4 Variational principles

One may pick out the function from the set of admissible spectral densities that satisfies some additional properties. For example, such function may be picked from an extremal condition. Consider such a case.

Introduce the following functionals,J_1, J_2, defined on admissible spectral densities :

$$J_1 = \frac{1}{2\pi i} \oint_{|\lambda|=1} \ln\{\tilde{G}(\lambda)\} \frac{d\lambda}{\lambda}, \tag{3}$$

$$J_2 = \frac{1}{2\pi i} \ln\{ \oint_{|\lambda|=1} \tilde{G}(\lambda)\} \frac{d\lambda}{\lambda}, \tag{4}$$

where $\oint_{|\lambda|=1} \frac{d\lambda}{\lambda}$ denotes the oriented circular integral. $J_1(\tilde{G})$ is known as the entropy associated with $\tilde{G}$ [1], $J_2(\tilde{G})$ may be called the negentropy. Evidently, $J_1(\tilde{G}) \leq J_2(\tilde{G})$ (if these values exist). Let

$$G_{opt}^{(1)} = \arg\min_{\tilde{G}} \ J_1(\tilde{G}), \tag{5}$$

$$G_{opt}^{(2)} = \arg\min_{\tilde{G}} \ J_2(\tilde{G}), \tag{6}$$

be extremal densities defined by these functionals.

Lemma. The relations (5), (6) may be written as

$$G_{opt}^{(1)}(\lambda) = \frac{\sigma^2}{a(\lambda)a(\lambda^{-1})}, \tag{7}$$

$$G_{opt}^{(2)} = \sigma^2 \sum_{t=-T}^{T} \lambda^t \tilde{R}(t) \tag{8}$$

where

$$a(\lambda) = 1 + \lambda a_1 + ... + \lambda^T a_T \tag{9}$$

is the polynomial with real coefficients $a_k, k = 1, ..., T$, defined from the linear system

$$\sum_{k=1}^{T} R(p-k+1)a_k = -R(p),\ p = 1, ..., T. \tag{10}$$

System (10) may be solved with the help of Levinson's method.

Remarks. Let the stationary time series $\tilde{\eta} = \{\tilde{\eta}_t,\ t \in \mathbf{Z}\}$ be determined by the equation

$$\tilde{\eta}_t + a_{t-1}\tilde{\eta}_t + ... + a_T\tilde{\eta}_{t-T} = \sigma e_t,\ t \in \mathbf{Z}, \tag{11}$$

where a_k - the above coefficients (see (9)) and $e = \{e_t,\ t \in \mathbf{Z}\}$, - a standard white noise ($Me_t = 0, Me_te_{t'} = \delta_{tt'}$). Then the relations

$$\tilde{R}(s) = M\tilde{\eta}_t\tilde{\eta}_{t-s} = M\eta_t\eta_{t-s} = R(s),\ |s| \le T, \tag{12}$$

are justified. (Here η is the stationary component of time series y.) It means that $G^{(1)}_{opt}$ is AR (Auto Regressive)- approximation of the spectral density $G = G(\lambda)$ [3].

Formula (8) corresponds to a periodogram approach. Indeed, let the stable polynomial $b(\lambda)$ ($b(\lambda) \neq 0,\ |\lambda| \neq 1$) be defined by factoring

$$G^{(2)}_{opt}(\lambda) = b(\lambda)\ b(\lambda^{-1})$$

and $e = \{e_t,\ t \in \mathbf{Z}\}$ be a standard white noise. For time series $\tilde{\eta} = \{\tilde{\eta}_t,\ t \in \mathbf{Z}\}$, $\tilde{\eta}_t = \sigma\{e_t + b_1e_{t-1} + ... + b_Te_{t-T}\}$, the relations (11) are fulfilled. It means that $G^{(2)}_{opt}(\lambda)$ is MA (Moving Average)-approximation of the spectral density $G(\lambda)$.

5 The general case

To consider the general case, we introduce for any natural $L, L \le T$, a quasi-polynomial

$$q^L(\lambda) = \lambda^{-L}q_{\ L} + ... + \lambda^{-1}q_{-1} + 1 + \lambda q_1 + ... + \lambda^L q_L \tag{13}$$

with real coefficients q_l which are determined from the following linear system of $2L$ equations:

$$\sum_{t=1}^{T} R(k-t)q_t = 0,\ T - L + 1 \le |k| \le T.$$

$p^{(M)}(\lambda),\ M = T - L$, be the quasi-polynomial,

$$p^{(M)}(\lambda) = \lambda^{-M}p_{-M} + ... + \lambda\ \ p_M,\ p_k = \sum_{t=-L}^{L} R(k-t)q_t,\ |k| \le M.$$

It is easy to see that the following relation is fulfilled

$$q^{(L)}(\lambda)G(\lambda) - p^{(M)}(\lambda) = \sum_{|k|>T} \lambda^k \left(\sum_{t=-T}^{T} R(k-t)q_t \right). \tag{14}$$

Definition. The rational function $\pi^{(L,M)}(\lambda)$,

$$\pi^{(L,M)}(\lambda) = \frac{p^{(M)}(\lambda)}{q^{(L)}(\lambda)}, \tag{15}$$

is called (L, M)-order Pade approximation of the quasi-polypolynomial $G(l)$. This definition is similar to Pade approximation of power series [4].

Remarks. It is evidently that $\pi^{(T,0)}(\lambda)$ coincides with AR-approximation and $\pi^{(0,T)}(\lambda)$ coincides with AR-approximation. To design $\pi^{(L,M)}(\lambda)$, $L + M = T$, it is necessary to know $R(s)$ for all $|s| \leq L + T$.

Theorem. Let us suppose that the spectral density $G(\lambda)$ expansion (1) absolutely and uniformly converges in the disk $D_\rho = \{\lambda : \rho^{-1} < |\lambda| < \rho\}$ for some real $\rho > 1$ (it means in particular that $G(\lambda)$ is analytic function in D_ρ).

Then the following assertions are satisfied:

1. The Pade approximation (14) of $G(\lambda)$ is symmetric, $q_{-t} = q_t$, $p_{-t} = p_t$.

2. If $G(\lambda) > 0$ on the unit circle $|\lambda| = 1$ then $q^{(L)}(\lambda) > 0$, $p^{(M)}(\lambda) > 0$ for all such λ and for sufficiently large $T = L + M$.

3. $\lim\limits_{L+M\to\infty} \frac{p^{(M)}(\lambda)}{q^{(L)}(\lambda)} = G(\lambda)$ for all $\lambda \in D_\rho$.

4. If $G(\lambda)$ is a rational function,

$$G(\lambda) = \frac{p(\lambda)}{q(\lambda)}, \tag{16}$$

where $p(\lambda), q(\lambda)$ are quasi-polynomials of some degrees l, m, then

$$G(\lambda) = \frac{p^{(M)}(\lambda)}{q^{(L)}(\lambda)} \tag{17}$$

for all $L > 2l$, $M > l + m$.

6 ARMA-approximation

Let $a(\lambda)$, $b(\lambda)$ be the result of spectral factoring of positive quasi-polynomials $q^{(L)}(\lambda), p^{(M)}(\lambda)$,

$$p^{(M)}(\lambda) = a(\lambda)\, a(\lambda^{-1}),\ q^{(L)}(\lambda) = b(\lambda)\, b(\lambda^{-1}),$$

$|a(\lambda)| + |b(\lambda)| \neq 0$ for all $|\lambda| \leq 1$, and let $\tilde{\eta} = \{\tilde{\eta},\ t \in \mathbf{Z}\}$ be the stationary time series defined by equation

$$\tilde{\eta}_t + a_1\tilde{\eta}_{t-1} + \ldots + a_L\tilde{\eta}_{t-L} = \sigma\,(b_0\epsilon_t + \ldots + b_M\epsilon_{t-M}). \tag{18}$$

Then the relations (11) are justified. It means that $\pi^{(L,M)}(\lambda)$ may be regarded as ARMA (Auto Regressive-Moving Average)-approximation of the spectral density $G(\lambda)$.

Looking over L, M, $L + M = T$, one may pick out the best Pade approximation (in different sense).

Appendix

Proof of the lemma. From (3) due to (2) we have

$$\delta J_1(\tilde{G}) = \frac{1}{2\pi i} \oint\limits_{|\lambda|=1} \frac{\delta\left\{\tilde{G}(\lambda)\right\}}{\tilde{G}(\lambda)} \frac{d\lambda}{\lambda}$$

$$= \frac{1}{2\pi i} \left[\sum_{t=-\infty}^{-T-1} + \sum_{t=T+1}^{\infty} \right] \oint\limits_{|\lambda|=1} \frac{\lambda^t \delta\left\{\tilde{R}(t)\right\}}{\tilde{G}(\lambda)} \frac{d\lambda}{\lambda}.$$

The equality $\delta J_1(\tilde{G}) = 0$ for arbitrary variation $\delta\{\tilde{R}(t)\}$, $|t| > T$, leads to relations for optimal $G_{opt}^{(1)}$:

$$\frac{1}{2\pi i} \oint\limits_{|\lambda|=1} \frac{\lambda^t}{\tilde{G}(\lambda)} \frac{d\lambda}{\lambda} = 0, \ |t| > T. \tag{A.1}$$

It means that $[\tilde{G}_{opt}(\lambda)]^{-1}$ is a quasi-polynomial,

$$[G(\lambda)]^{-1} = \sum_{t=-T}^{T} \lambda^t q_t \tag{A.2}$$

where

$$q_t = \frac{1}{2\pi i} \oint\limits_{|\lambda|=1} \frac{\lambda^t}{\tilde{G}_{opt}(\lambda)} \frac{d\lambda}{\lambda}, \ |t| \le T. \tag{A.3}$$

Because of positivity, the admissible density $\tilde{G}_{opt}(\lambda) = G_{opt}^{(1)}(\lambda)$ may be factorized, i.e.,

$$G_{opt}^{(1)}(\lambda) = \frac{\sigma^2}{a(\lambda)a(\lambda^{-1})}, \ \sigma^2 = \frac{1}{2\pi i} \oint\limits_{|\lambda|=1} G_{opt}^{(1)}(\lambda) \frac{d\lambda}{\lambda}, \tag{A.4}$$

where $a(\lambda)$ is a stable real polynomial (see (9)) with real coefficients. Let e be a standard white noise and $\tilde{\eta} = \{\tilde{\eta}_t, \ t \in \mathbf{Z}\}$ be the stationary time series defined by the equation (11). Due to polynomial $a(\lambda)$ stability we have from (11) the following relation

$$\tilde{R}(p) + a_1 \tilde{R}(p-1) + ... + a_T \tilde{R}(p-T) = 0$$

$$p = t-1, ..., t-T, \tag{A.5}$$

and due to the condition $\tilde{R}(k) = R(k)$, $k = 1, 2, ..., T$, the linear system (10) coincides with (A.5). Because of spectral density (A.4) positivity the system (A.5) is nondegenerate, so it exists the unique solution of this system. Relations (7)-(9) are established.

Formula (8) is almost evident. Indeed, with help of formulae (4), (2) we have

$$\delta J_2(\tilde{G}) = \frac{\sum\limits_{|t|>T} \frac{1}{2\pi i} \oint\limits_{|\lambda|=1} \lambda^t \frac{d\lambda}{\lambda} \delta\tilde{R}(t)}{\frac{1}{2\pi i} \oint\limits_{|\lambda|=1} \tilde{G}(\lambda) \frac{d\lambda}{\lambda}} = 0.$$

Because of the correlation function $R(.)$ admissibility it means $\tilde{R}(t) = 0$ for all $|t| > T$ and from (2) we have (8). Lemma is proved.

A brief proof of the theorem. The first assertion follows from symmetry of a correlation function and the definition of Pade approximation. The second assertion follows from the third one. To ground last assertion let us use Caushy formula

$$f(\lambda) = \frac{1}{2\pi i}\oint\limits_{\gamma_\rho}\frac{f(\mu)}{\lambda-\mu}d\mu - \frac{1}{2\pi i}\oint\limits_{\gamma_\rho{}^{-1}}\frac{f(\mu)}{\lambda-\mu}d\mu \tag{A.6}$$

that is just for all the analytical function

$$f(\lambda) = G(\lambda)q^l(\lambda), \tag{A.7}$$

$\gamma_{\rho^{-1}} = \{\lambda : |\lambda| = \rho^{-1}\}, \rho > 1$. Using then in (A.9) expansions

$$\frac{1}{\lambda-\mu} = \frac{1}{\mu}\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k \quad (\mu \in \gamma_\rho,\ \lambda \in D_\rho),$$

$$\frac{1}{\lambda-\mu} = \frac{1}{\lambda}\sum_{k=0}^{\infty}\left(\frac{\mu}{\lambda}\right)^k \quad (\mu \in \gamma_\rho^{-1},\ \lambda \in D_\rho),$$

we have

$$f(\lambda) = \frac{1}{2\pi i}\oint\limits_{\gamma_\rho} f(\mu)\sum_{k=0}^{\infty}\left(\frac{\mu}{\lambda}\right)^k\frac{d\mu}{\mu}$$

$$-\frac{1}{2\pi i}\oint\limits_{\gamma_\rho{}^{-1}}\frac{\mu}{\lambda}f(\mu)\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\frac{d\mu}{\mu}. \tag{A.8}$$

Introduce a quasi-polynomial

$$F^{(L+M)}(\lambda) = \frac{1}{2\pi i}\oint\limits_{\gamma_\rho} f(\mu)\sum_{k=0}^{L+M}\left(\frac{\lambda}{\mu}\right)^k\frac{d\mu}{\mu}$$

$$+\frac{1}{2\pi i}\oint\limits_{\gamma_\rho{}^{-1}}\frac{\mu}{\lambda}f(\mu)\sum_{k=0}^{L+M+1}\left(\frac{\mu}{\lambda}\right)^k\frac{d\mu}{\mu}$$

of degree $(L+M)$. Then

$$G(\lambda)q^{(L)}(\lambda) - F^{(L+M)}(\lambda) = \frac{1}{2\pi i}\oint\limits_{\gamma_\rho} f(\mu)\sum_{k=L+M+1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\frac{d\mu}{\mu}$$

$$-\frac{1}{2\pi i}\oint\limits_{\gamma_\rho{}^{-1}}\frac{\mu}{\lambda}f(\mu)\sum_{k=L+M}^{\infty}\left(\frac{\mu}{\lambda}\right)^k\frac{d\mu}{\mu}. \tag{A.9}$$

But in accordance with the definition of Pade approximation we have

$$G(\lambda)q^{(L)}(\lambda) = b^{(M)}(\lambda) + \sum_{t=-\infty}^{-L-M-1} q_t\lambda^t + \sum_{t=L+M}^{\infty} q_t\lambda^t \tag{A.10}$$

where

$$q_t = \frac{1}{2\pi i} \oint_{\gamma_1} \lambda^{-t} G(\lambda) q^{(L)}(\lambda) \frac{d\lambda}{\lambda}.$$

From (A.9) and (A.10) it follows that

$$F^{(L+M)}(\lambda) = \sum_{t=-\infty}^{-L-M-1} q_t \lambda^t + \frac{1}{2\pi i} \oint_{\gamma_\rho} f(\mu) \sum_{k=L+M+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{d\mu}{\mu}$$

$$- \left[\sum_{t=L+M}^{\infty} q_t \lambda^t + \frac{1}{2\pi i} \oint_{\gamma_{\rho^{-1}}} \frac{\mu}{\lambda} f(\mu) \sum_{k=L+M}^{\infty} \left(\frac{\mu}{\lambda}\right)^k \frac{d\mu}{\mu} \right]. \qquad (A.11)$$

The bracket function may have nonzero Fourier coefficients only for $|t| > L + M$. As $F^{(L+M)}(\lambda)$ is a quasi-polynomial of degree $(L + M)$, the relation (A.11) means that

$$F^{(L+M)}(\lambda) = b^{(M)}(\lambda)$$

for all $\lambda \in D_\rho$ and from (A.9) we have

$$G(\lambda) = \frac{p^{(M)}(\lambda}{q^{(L)}(\lambda} + \frac{1}{2\pi i} \oint_{\gamma_\rho} f(\mu) \sum_{k=L+M+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{d\mu}{\mu}$$

$$- \frac{1}{2\pi i} \oint_{\gamma_{\rho^{-1}}} \frac{\mu}{\lambda} f(\mu) \sum_{k=L+M}^{\infty} \left(\frac{\mu}{\lambda}\right)^k \frac{d\mu}{\mu}. \qquad (A.12)$$

Now the third assertion of the theorem is evident. Let $G(\lambda)$ be rational,

$$G(\lambda) = \frac{p(\lambda)}{q(\lambda)} \qquad (A.13)$$

where p, q are quasi-polynomials of some degrees m and l. From (A.12),(A.13) we have

$$p(\lambda) q^{(L)}(\lambda) - q(\lambda) p^{(M)}(\lambda)$$

$$= q(\lambda) \frac{1}{2\pi i} \oint_{\gamma_\rho} f(\mu) \sum_{k=L+M+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{d\mu}{\mu}$$

$$+ q(\lambda) \frac{1}{2\pi i} \oint_{\gamma_{\rho^{-1}}} \frac{\mu}{\lambda} f(\mu) \sum_{k=L+M}^{\infty} \left(\frac{\mu}{\lambda}\right)^k \frac{d\mu}{\mu}.$$

If $M > l + m$ and $L > 2l$ then the functions in the right part of this relation have zero Fourier coefficients for $|t| \leq \max(L + m, M + l)$, but in the left part we have a quasi-polynomial of degree $\max(L + m, M + l)$. It means that this quasi-polynomial identically is equal to zero. So

$$G(\lambda) = \frac{p(\lambda)}{q(\lambda)} = \frac{p^{(M)}(\lambda)}{q^{(L)}(\lambda)}$$

and the last assertion of the theorem is proved.

References

[1] Modern Spectrum Analysis / Edited by D.G.Childers. N.Y., IEEE Press, 1978.

[2] Burg, J.P. "Maximum entropy spectral analysis", Presented at the 37th Annual Meeting of Exploration Geophy sicists. Oklahoma City, Okla., 1967. (The paper is included in Modern Spectrum Analysis, see above, p.34-39).

[3] A.van den Bos. Alternative interpretation of Maximum Entropy spectral analysis // IEEE Trans. Inform. Theory, 1971. Vol. IT-17. P. 493-494.

[4] Baker, G.A.Jr. and Graves-Moris, P. Pade Approximant. London, Amsterdam,...: Adison-Wesley Publ. Co, 1981.

Statistical Safety Theory and Railway Applications

A.E. Kraskovsky

1 Introduction

The purpose of the work is the development of the methods and software for the statistical analysis including prognosis of the accidents in a technological process. The major application concern the safety increase in the railway transport.

The statistical safety theory deals with the following problems: (i) statistical analysis of accident data and search for the significant factors that mainly determine the accidents; (ii) the prognisis of possible accidents; (iii) on-line emergency indication; and (iv) mathematical modelling of the stochastic dynamic systems for safety control.

The time series theory is the ground of the statistical safety theory. Also, the factor, correlation and regression analysis are used in (i), the reliability theory is used in (ii), the parameter estimation and the change-point detection methodologies are key items of (iii), and the stochastic differential equations determine (iv).

The article presents the current state of art in the field, several particular mathematical models for the safety control in the railway transport and numerical results obtained with the help of statistical software and statistical simulation.

We shall mainly consider problems (ii) and (iii).

2 Patterns of the emergency and the accidents arising

There are several patterns of the accidents (Fig.1). The following situations may be considered as an example for these patterns. Pattern 1 is a sudden fault of the system that leads to an accident, for instance: faults of the equipment, wrong actions of a person, sudden influences of environment.

There is another situation for the pattern 2. There is emergency for the time interval $\tau_0 < t < \tau_1$, there is still not an accident. Examples of these situations are the following.

(i) The rail breaking by the train absence

(ii) The change of regime of fly-engine that may lead to the accident

(iii) The decrease of the isolation resistor of the electrical line that may lead to the fire.

Examples of situations for patterns 3 ... 5 are the following.

(i) The accumulation of injures within the construction elements

(ii) The increase of temperature by friction of construiction elements

(iii) The fatigue accumulation of a man.

For solving problems of current transport system indication parameters and diagnosing emergency mathematical methods of statistical estimators and making decisions about change values of observational data are used. High requirements to measurement accuracy and time reaction which produced systems of emergency indication make for to use some optimal or close to them observations processing algorithms. To synthesize such algorithms it is possible on the base of the theory detections of time series change-point.

3 Methods of the change-point detection of random processes

3.1

Let us give a definition of the change-point problem. The change -point is an abrupt (sometimes smooth) one of properties at a random process y_t, for instance, the probability distribution at the unknown moment of time, one or several parameters of the distribution.

Solving the change-point problem is testing two hypotheses H_0 and H_1 where the hypothesis H_0 is that the change-point is absent and the alternative hypothesis H_1 is the change happens. If we reject hypothesis H_0 in favour of alternative then we can state the estimation problem of the change-point τ.

Hypotheses for the stationary process are the following.

H_0: random values $y_1, y_2, \ldots$ have the dirtibution F_θ, where a parameter $\theta = \theta_0 \in \Theta_0$.

H_1: there is $\tau \geq 1$, that random values $y_1, y_2, \ldots, y_{\tau-1}$ have a distribution F_θ, $\theta = \theta_0$ and $y_\tau, y_{\tau+1}$ have the distribution F_θ, $\theta = \theta_1 \in \Theta_1$.

Sets of parameters Θ_0 and Θ_1 are subsets of R^m, $m \geq 1$, Θ_0 and Θ_1 have no intersections. There are abrupt and smooth, singular and multiply, with the limited and unlimited time change. We are largely interested in singular abrupt changes.

3.2

Consider the likelihood ratio test. This test for a posteriori change-point detection is in the following:

$$\max_{\theta_0 \in \Theta_0} \max_{\theta_1 \in \Theta_1} \max_{1 \leq \tau \leq N} L_{H_0/H_1} \underset{H_0}{\overset{H_1}{\gtrless}} h,$$

where h is the threshold, $L_{H_0/H_1} = L_N(y_1^N, \theta_0, \tau, \theta_1)$ is the likelihood ratio which is expressed by formula

$$L_{H_0/H_1} = \prod_{t=\tau}^{N} \frac{p_{\theta_1}(y_t|y_{t-1}, \ldots)}{p_{\theta_0}(y_t|y_{t-1}, \ldots)}$$

Here $p_{\theta_0(1)}(y_t|\ldots)$ is a conditional density of corresponding distributions $P_{\theta_0(1)}(y_t|\ldots)$ Equivalent test for likelihood ratio one is in the following

$$\max_{\tau} L_N(y_1^N, \widehat{\theta}_0(\tau), \widehat{\theta}_1(\tau)) \underset{H_0}{\overset{H_1}{\gtrless}} h,$$

where $\widehat{\theta}_0(\tau)$, $\widehat{\theta}_0(\tau)$ are estimators of maximal likelihood parameters $\widehat{\theta}_0$ and $\widehat{\theta}_1$ obtained provided that hypothesis H_0 is rejected if τ is fixed.

2.3. Consider algorithm of cumulative sums (CUSUM). In case of independent observations and known θ_0 and θ_1 this algorithm can be regarded as a recurrent form of the likelihood ratio test. If k is the current observation number in the recurrent algorithm then we have

$$S_i^k(y_t,\theta) = \ln L_{H_1/H_0} = \sum_{t=i}^{k} \ln\left(p_{\theta_1}(y_t)/p_{\theta_0}(y_t)\right), \tag{1}$$

where $S_i^k(y_t,\theta)$ is the value of the cumulative sum at the step k.

The algorithm (1) is often represented as CUSUM with the reflecting screen

$$g_t = \left(g_{t-1} + \ln\left(p_{\theta_1}(y_t)/p_{\theta_0}(y_t)\right)\right)^+,$$

where $g_0 = 0$, $a^+ = \max\{0, a\}$.

We can find the estimator $\widehat{\tau}$ by the maximal likelihood ratio method:

$$\widehat{\tau} = \underset{1<\tau<N}{\operatorname{argmax}} \; L_N(y_1^N, \theta_0, \tau, \theta_1) = \underset{1<\tau<N}{\operatorname{argmax}} \; S_\tau^N(y_t,\theta) \tag{2}$$

3.3

We can use the following statistics for the change-point testing

(i) Girshic-Rubin statistic

$$w_n = \frac{p_{\theta_1}(y_n)}{p_{\theta_0}(y_n)}[1 + w_{n-1}],$$

(ii) Exponential smoothing statistic

$$T_n = (1-r)T_{n-1} + ry_n,$$

(iii) Shyuhart's carts

$$G_n = \frac{1}{k}\sum_{i=n-k+1}^{n} y_i$$

The last two algorithms are nonparametric. In the case of the small unknown change CUSUM is optimal since it takes minimal time for the change-point detection if probability of the error is fixed. In the case of great change the exponential smoothing algorithm is more effective than CUSUM. For independent observations and known parameters of the random process before and after change-point CUSUM is optimal in sense of Neyman-Pearson criterion but CUSUM is not robust if there are anomal observation in the sample.

The likelihood ratio test often gives optimal or close to optimal approaches.

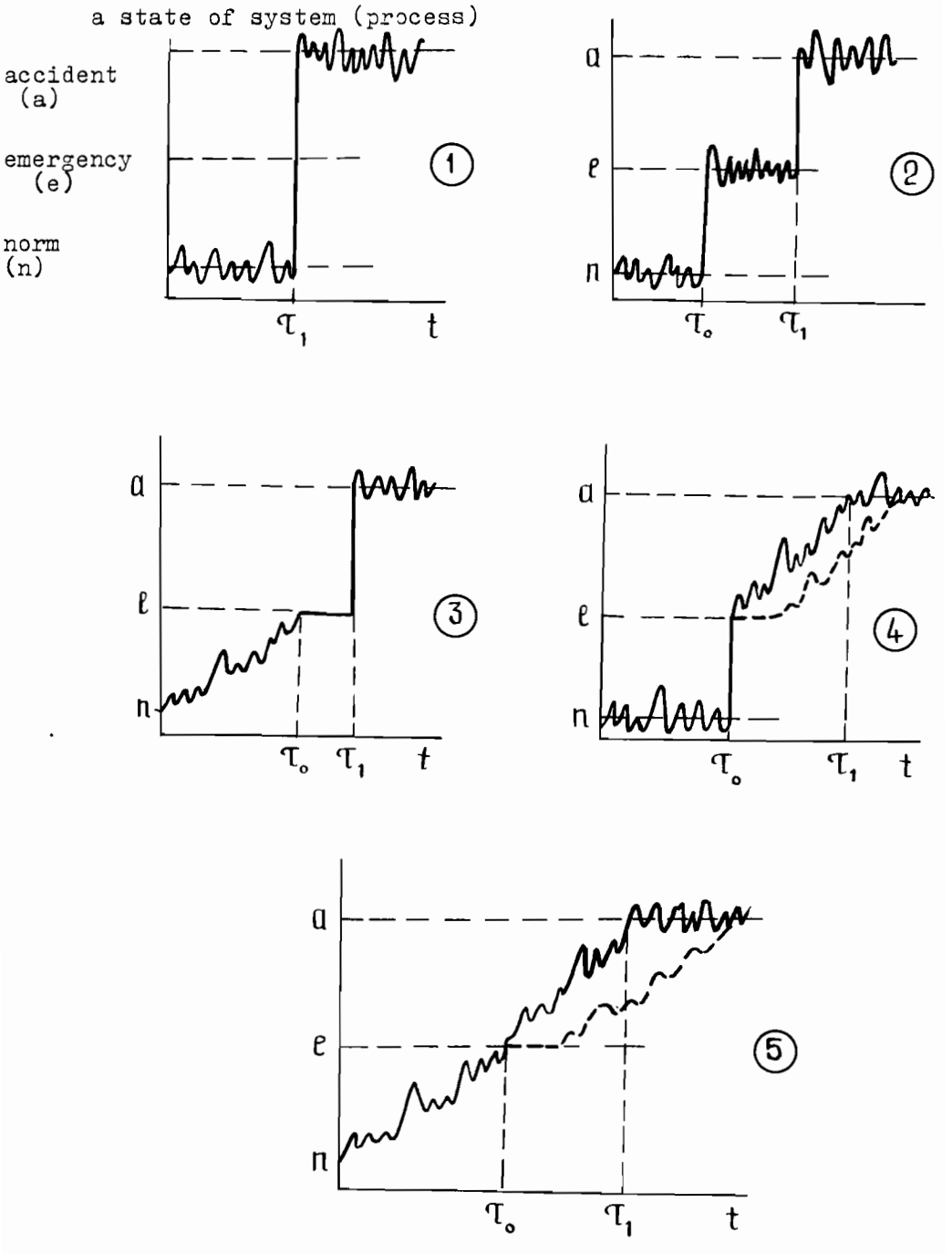


Figure 1: Patterns of the emergency and the accidents arising

3.4

There is a particular case which is important because of many practical problems reducing to the similar case. Consider the likelihood ratio change-point test for limited time. This problem often occurs in tasks of detection of emergency and signals with random appearance moment.

In this case a model is represented such that

$$y_t = \varphi(t-\tau) + \xi_t,$$

where $\varphi(t)$ is the function determining the form of the signal with known duration T, $t = 1, 2, \ldots, N$: τ is an unknown appearance moment; ξ_t is a random process with mean zero and variance σ^2.

We can reduce the time limited change-point detection problem to testing hypotheses H_0 and H_1 for the mean Ey_t on $[O, N]$. In this case the hypotheses are in the following.

$$H_0 : \; Ey_t = \theta\varphi(t), \; \theta = \theta_0 = 0$$

$$H_1 : \; \exists\tau, \; 0 \le \tau \le N_T \; : \; Ey_t = \theta\varphi(t-\tau), \theta = \theta_1 = 1$$

The likelihood ratio test for these hypotheses is as follows

$$\max_{\tau} \ln L_{H_1/H_0} = \max_{\tau} \ln \frac{L(y_{(\tau,\tau+T)}|\theta = \theta_1)}{L(y_{(\tau,\tau+T)}|\theta = \theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} h_0,$$

where $L(y_{(\tau,\tau+T)}|\theta = \theta_i)$ is a joint distribution density y_t for $t \in (\tau, \tau+T)$ providing that $\theta = \theta_i$, $i = 0, 1$.

In case of discrete time we have

$$\ln L_{H_1/H_0} = \sum_{t=\tau+1}^{\tau+T} \ln \frac{f_t(y_t - \varphi(t-\tau))}{f_t(y_t)} \tag{3}$$

where $f_t(z)$ is the distribution density of ξ_t at the moment t.

In a most interesting case when the density f_t is normal with a zero mean and variance σ^2 we have

$$\max_{\tau} \sum_{t=\tau+1}^{\tau+T} \left[y_t\varphi(t-\tau) - \frac{\varphi^2(t-\tau)}{2} \right] \underset{H_0}{\overset{H_1}{\gtrless}} h_0\sigma^2$$

To apply the change-point detection methods in practice it needs creating the mathematical sample model, selecting the change -point detection test and optimizing the test parameters by one of the criterion. For instance, to optimize a test we must choose the detection threshold if the probability of the error detection is given.

The decisive statistic for the change-point detection can be usually described by the random process. The change-point is detected when a random process reaches a boundary.

The second problem (ii) is the estimation and prognosis of the system reliability by the injure accumulation. The system fault interprets as the first crossing of the random process to a boundary too. Therefore the mathematical problem arising is the investigation of the first crossing probability for particular boundaries and random processes.

4 Calculating methods of the boundaries crossing probability by random processes

4.1

The general statement of the problem is in the following. Let $\xi(t)$ be a random process given on the segment $[O,T]$, $0 < T \leq \infty$ For $t = 0$ the value of$\xi(0)$ equal to x is fixed, $h(t)$ is a continuous function, $h(t)$ may be a direct linear, piece-wise linear or nonlinear function. The probability of the boundary crossing for random process is equal to:

$$P_\xi(T, h|x_0) = P\{\xi(t) \geq h(t) \text{ if only for one } t \in [0,T]|\xi(0) = x_0\} \tag{4}$$

If the value of $x_0 = \xi(0)$ is not fixed but its distribution density is given then the probability of boundary crossing is equal to

$$P_\xi(T, h) = \int_{-\infty}^{\infty} P_\xi(T, h|x_0)p(x_0)dx_0 \tag{5}$$

If T is a random value and its probability density $p(T)$ is known the probability $P_\xi(h)$ of the boundary crossing is equal to

$$P_\xi(h) = \int_0^{\infty} P_\xi(T, h)p(T)dT \tag{6}$$

The porobabilities (4) and (5) are defined by densities of the time of first boundary crossing h by a process $\xi(t)$: $g_\xi(t, h|x_0)$ and $g_\xi(t, h)$. They are relatively equal to

$$g_\xi(t, h|x_0) = \frac{d}{dt}P_\xi(t, h|x_0);\ g_\xi(t, h) = \frac{d}{dt}P_\xi(t, h) \tag{7}$$

and the probabilities (4) and (5) may be defined by the density (7). In this case the solution of the problem of boundary crossing via density or probability are equal.

For the class of processes with independent increments a useful result was obtained. If $\xi(t)$ is the process with independent increments then for $h(t) = h > 0$, $x_0 < h$ we have:

$$g_\xi(t, h|x_0) = \frac{h - x_0}{t}P[\xi(t) = h - x_0|\xi(0) = x_0]$$

where $P[\xi(t) = a|\xi(0) = x_0]$ is the probability density of the transition process from point x_0 to a point a during the time t.

For a more general class of random Markov processes there are some difficulties to obtain exact equations (4) and (5). The method based on the solution of the so-called Siegert regeneration equation fordirect boundary is proved to be effective. It is in the following

$$P[\xi(T) = x|\xi(0) = x_0] = \int_0^T g_\xi(h, t|x_0)p(\xi(T) = x|\xi(t) = h)dt \tag{8}$$

where $x_0 < h < x$, $p[\xi(t) = x|\xi(\tau) = z]$ is the process transition probability density $\xi(t)$ from a point z to a point x during the time from τ to t. The Eq. (8) means that before the transition to a point $\xi(T) = x > h$ the process first crosses the boundary h at some moment of time $t \in [0,T]$ and during the rest of time from t to T it will go from a point h to a point x.

4.2

Let us take as an example a standard Wiener process . For this process the regeneration Eq. (8) is readily solved and the time density of the first approach to the boundary h is equal to

$$q_w(t,h|x_0)=\frac{h-x_0}{\sqrt{2\pi t}}\exp\left[-(h-x_0)^2/(2t)\right] \tag{9}$$

Integrating the expression (9) in accordance with (7) we have

$$P_w(T,h|0)=2\Phi\left(-\frac{h}{\sqrt{T}}\right),$$

where Φ is the probability integral. The occasion of a linear boundary $h(t)=at+b$ is important. E.g., it occurs for Wiener processes when $Ew(t)\neq 0$. They may serve as models for a damage accumulation process in mechanical systems. The direct boundary is transformed into the linearboundary in transition from the occasion $Ew(t)\neq 0$ to $Ew(t)=0$.

For linear boundary there are two results [5]

(i) $P_w(\infty,at+b|0)=e^{-2ab}$,

(ii) $P\{w(t)\geq at+b \text{ if only for } t\in[t_1,t_2]|w(t_1)=x_1,w(t_2)=x_2\}=$
$=\exp\left\{-\frac{2}{t_2-t_1}(at_1+b-x_1)(at_2+b-x_2)\right\}$
where $t_2>t_1,\ x_1\leq at_1+b;\ x_2\leq at_2+b$

If a time interval $[0,T]$ occurs and $w(t=0)=0$ then the probability of the linear boundary $at+b$ crossing by the standard Wiener process $w(t)$ if possible may be obtained by means of the result (ii):

$$P_w(T,a+bt|0)=1-\Phi\left(\frac{bt+a}{\sqrt{T}}\right)+e^{-2ab}\Phi\left(\frac{bt-a}{\sqrt{T}}\right) \tag{10}$$

If $w(t=0)$ is a random value x_0 and the probability density $p_0(x_0)$ is given then the probability of crossing this boundary is equal to

$$P_w(T,a+bt)=\int_{-\infty}^{h}P_w(T,a+bt|x_0)p_0(x_0)dx_0, \tag{11}$$

where the probability $P_w(T,a+bt|x_0)=P_w(T,a+x_0+bt|0)$ is defined by formula (10).

The crossing of boundary h may occur at the time $t=0$ as well. Taking this fact into account we may find the total boundary crossing probability at least once during a given period of time $M=T/T_0$, where T_0 is a unit of time

$$P_w(M,a+b\tau,h)=P_w(M,a+b\tau)+\int_h^{\infty}P_0(x_0)dx_0, \tag{12}$$

where probability $P_w(M,a+b\tau)$ is calculated with the help of formula (11); $\tau=t/T_0$. After a number of transformations for the Exp. (12) for $M=1$ we will have

$$P_h(M=1)=1-\Phi^2(h)+\frac{1}{\sqrt{2\pi}}he^{-\frac{h^2}{2}}\Phi(h)+\frac{1}{\sqrt{2\pi}}e^{-h^2}$$

For $M>1$ there is an approximate formula [2]

$$P_h(M)\approx M(1-\lambda)$$

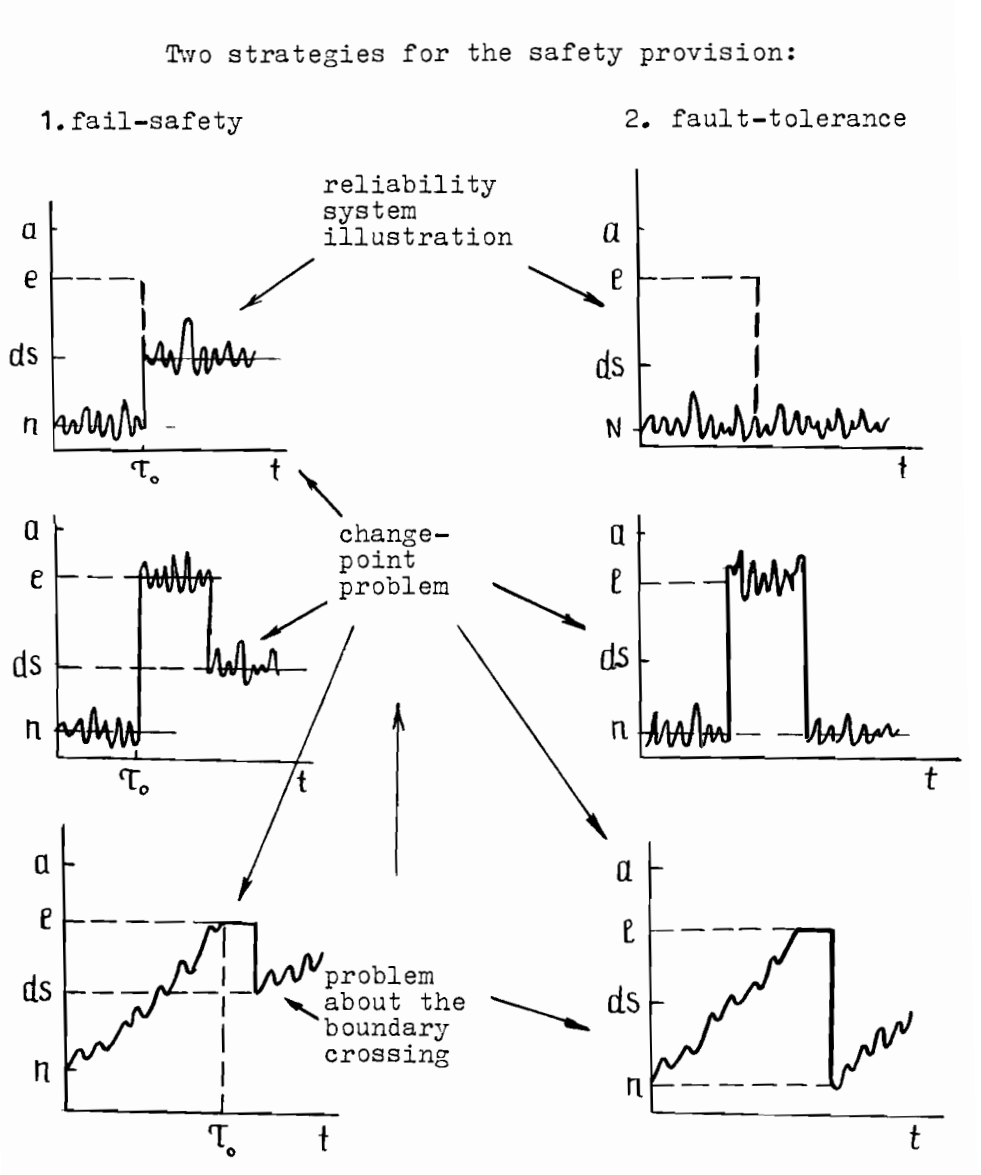


Figure 2: Applications of methods considered for the safety provision. States of system (process): norm (n) - emergency (e) - defense state (ds); accident (a).

where $\lambda = \Phi(h) + \frac{1}{4}(9\Phi^2(h) - 16(he^{-h^2/2}\Phi(h)/\sqrt{2\pi} + e^{-h^2}/(2\pi)))^{1/2}$

Calculating probabilities of the crossing piece-wise linear boundaries the monograph [2] is considered and the nonlinear boundaries case were studied in [8].

The mathematical methods considered may be used for the emergency diagnosis as well as for accidents prognosis. In Fig.2 one shows the role of methods discussed in realization of two safety provision strategy: 'fail-safety' and 'fault-tolerance'.

Therefore we have demonstrated some models and so mathematical results for estimation and prognosis of accidents occur on technological processes, in particular, on the railway transport.

Acknowledgement

The author expresses his satisfaction with the long-term fruitful scientific contacts with Professor of St.Petersburg University Anatoly Zhigljavsky.

References

1. Nikiforov I.W. The successive detection of the time series property change. M, 1983.
2. Zhigljavsky A.A., Kraskovsky A.E. The change-point detection of the random processes in the radiotechnique problem. L.,LGU, 1988.
3. Detection of abrupt changes in signals and dynamical systems.Ed. by M.Basseville and A.Benveniste. Springer-Verlag Berlin Heidelberg New York Tokyo.
4.Durbin I. The first passage density of a continuous Gaussian process to a general boundary // J.Appl.Probab. 1985, Vol N 22, N 1.
5. Anderson T.W. A modification of the sequential probability ratio test.// Ann.Math.Statist. 1960. Vol.31, N 1.
6. Siegmund D., Yuh Y.S. Brownian approximations to first passage probabilities. // Z.Wahrsch.verw. Gebiete. 1982, Vol. 59, N 2.
7. Fukuda H. Problems in railway accidents analysis and safety assesment indices QR of RIRI. 1990, Vol.31, N 1.
8. Woodroofe M. Nonlinear reneval theory in sequential analysis, society for industrial and applied mathematics. 1982. Philadelphia.

The Universal Scheme of Regulations in Biosystems for the Analysis of Neuron Junctions as an Example

A.G. Bart, N.P. Clochkova and V.M. Kozhanov

1 Introduction. Reflections principle

Any real statistical problem covers many different aspects each having its own influence on the result. The heterogeneity of their contributions determines a mathematical structure of observations. For example, in the standard most simple scheme $\zeta = \eta + \epsilon$ the independent (changeable) component ϵ (technical noise) is separated from the basis (steady) component η (signal). As a rule this latter component reflects the quality nature of the phenomenon under consideration.

In biological systems this quality aspects usualy are connected with the self-regulation mechanism. The symmetry of reflections underlies such a mechanism in a sense that ensures the stability of its work. Being guided by our long standing experience of studing different real biological systems [1,2,3], we have formulated the following principle, giving the mathematical formalization of the mentioned symmetry of reflections:

In the strictly self-regulating system the realisations of two processes one of the action on the system and the second of its counteraction are the interinverse functions.

The operation of function inversing is to be understanded in the generalized sense, so that one can take into account the function peculiarities such as inmonotony, discontinuity and so on.

It makes us introduce a sufficiently wide class of partly inverse functions. Such functions are investigated in the next section, where, in particular, it is shown why the realizations of stochastic processes are the natural domain of the function inversing operation. In the central third section the shift and scale modifications of the generalized binomial distributions are determined by the means of parametric representation of the set of partly inverse functions for Bernoulli scheme realizations. They are important for the description of biosystems regulation. The essential special case of scale modification is connected with Fibonacci series.

Finally we illustrate the application of the reflections principle by the example from neurophyziology. This section has an independent interest, since using the language of amplitude distributions of post-synaptic potentials the logical explanation of impulse transmission mechanizm is given here.

2 Partly inverse functions

Consider an into mapping $f : \mathcal{X} \to \mathcal{Y}^-$. An into mapping $f^- : \mathcal{Y}^- \to \mathcal{X}$ is a generalized inverse mapping for f if for every $y \in \mathcal{Y} = f(\mathcal{X})$ and every $x \in \mathcal{X}^- = f^-(\mathcal{Y}^-)$ one of the

following relations is valid

$$O_1 : ff^-(y) = y \qquad O_3 : f^-f(x) = x$$
$$O_2 : f^-ff^-(y) = f^-(y) \quad O_4 : ff^-f(x) = f(x).$$

Let $f_*^+ = f|_{\mathcal{X}^-}$ and $f_*^- = f|_{\mathcal{Y}}$. We denote the conditions of injection of functions f_*^+ and f_*^- by $\mathcal{M}^+$ and $\mathcal{M}^-$ respectively, and the conditions of their surjection by $\mathcal{E}^+$ and $\mathcal{E}^-$.

STATEMENT

The unconditional (an arrow) and conditional (an arrow with a letter) relations between $O_1 - O_4$ can be expressed by the following scheme

$$\begin{array}{ccc} & \mathcal{M}^- & \\ O_1 & \rightleftharpoons & O_2 \\ \mathcal{E}^+ \uparrow\downarrow & & \uparrow\downarrow \mathcal{E}^- \\ O_4 & \rightleftharpoons & O_3 \\ & \mathcal{M}^+ & \end{array}$$

Indeed, the unconditional relations are evident, since $ff^-(y) \subset \mathcal{Y}^-$ and $f^-f(x) \subset \mathcal{X}$ respectively, the same takes place with $\mathcal{M}^+$ and $\mathcal{M}^-$. The relations $O_2 \Rightarrow O_3$ and $O_4 \Rightarrow O_1$ we get by substitutions of $x = f^-(y)$ into O_2 and $y = f(x)$ into O_4 , respectively. The existence of suitable x and y is unsured by the conditions $\mathcal{E}^-$ and $\mathcal{E}^+$. ■

DEFINITION

Suppose that f_m is a restriction of mapping f to some domain of its injection, and f_m^- is a generalized inverse for f_m. If for f_m and f_m^- all the relations $O_1 - O_4$ are equivalent (i.e. the fulfilment of any of them leads to the fulfilment of all other), then f_m^- is called as a patly inverse mapping for f.

A class of partly inverse mappings is wider a rule than the class of generalized inverse mappings. The analogy with the classes of partial and general recursive functions from mathematical logic is quite appropriate here. Further we shall consider only functions, defined on a set $\mathcal{X} \subset R^1$. For single-valued construction of partly inverse functions it is necessary to point a method of choosing the injection domains in the set $\mathcal{X}$ and the principle of definition of f^- on $\mathcal{Y}^- \setminus \mathcal{Y}$. The conditions of the $O_1 - O_4$ equvalence are necessary for the domain of the restriction f_m to be equal to injection domain of f.

The general parametrical method of description of partly inverse functions for a measurable function $f : R^1 \to R^1$ is pointed out in [1]. It will be used in the next section for the Bernoulli scheme realizations. Now we shall dwell upon the important particular case of the function inversing operation.

For every y we define the oriented pre-images: left $\overleftarrow{\mathcal{X}}_y = \{x \in \mathcal{X}; y \le f(x)\}$, right $\overrightarrow{\mathcal{X}}_y = \{x \in \mathcal{X}; y \ge f(x)\}$ and real $\dot{\mathcal{X}}_y = \{x \in \mathcal{X}; y = f(x)\}$. We shall give the operation of extreme inversions by the means of matrix notation of its possible versions

$$If = \begin{array}{c} \downarrow \\ \uparrow \end{array} \begin{array}{cc} \overleftarrow{f}^- & \overrightarrow{f}^- \\ f_-^- & f_+^- \end{array} = \begin{array}{cc}] & [\\ ] & [\end{array} = \begin{bmatrix} {}_0f_0^- & {}_1f_0^- \\ {}_0f_1^- & {}_1f_1^- \end{bmatrix}.$$

Here the horizontal pointers show the space orientation of preimages and the vertical ones show that of extremums (let's call it temporal). In designation ${}_if_j$ the right and left indexes

$(i,j = 0,1)$ correspond to these orientations. For example

$$f_{\rceil}^{-}(y) =_O f_O^{-}(y) = inf\ \overleftarrow{\mathcal{X}}_y;\ f_{\lfloor}^{-}(y) =_O f_1^{-}(y) = \sup\ \overleftarrow{\mathcal{X}}_y\ .$$

The left (f_O^-) and right (f_1^-) inverse functions are defined with extremums along the rows of matrix If. Minus one and plus one powers of the extreme inverse operation are set by

$$\overset{-1}{\rceil} f(y) = f_O^{-}(y) = \max\left\{f_{\rceil}^{-}; f_{\lceil}^{-}\right\};\ \overset{+1}{\rceil} f(y) = f_1^{-}(y) = \min\left\{f_{\rfloor}^{-}; f_{\lfloor}^{-}\right\}.$$

The notation $\rceil f$ is considered as a unified simbol of the extreme inverse function. Index of power (± 1) shows the orientation. The symbol of negation $(\rceil)$ can be treated from the semantic point of view here.

THEOREM 1

In the matrix If:

1) the elements of main diagonal do not decrease, those of secondary diagonal do not increase;

2) the elements of odd columns are continuons from the left, those of even columns are continuons from the right;

3) the left and the right extreme inverse functions preserve the monotony.

PROOF

1) The first assertion is evident: for every $\delta > 0$, $\overleftarrow{\mathcal{X}}_{y+\delta} \subset \overleftarrow{\mathcal{X}}_y$ and $\overrightarrow{\mathcal{X}}_y \subset \overrightarrow{\mathcal{X}}_{y+\delta}$.

2) We shall show,for example, continuity from the left for ${}_O f_O^-$. We fix y and suppose for every $\delta > 0$

$$x_\delta = {}_O f_O^{-}(y-\delta),\ x_O = \lim_{\delta \to O} x_\delta.$$

From the monotony of ${}_O f_O^-(y)$ it follows, that $x_\delta \le x_O \le {}_O f_O^-(y)$.

Let $x_O < {}_O f_O^-(y)$ then by the equality $\overleftarrow{\mathcal{X}}_{y-\delta} = \overleftarrow{\mathcal{X}}_y + \{y - \delta \le f(x) < y\}$ we have $x_\delta = \inf \overleftarrow{\mathcal{X}}_{y-\delta} = \min\left\{{}_O f_O^-(y);\ \inf\{y > f(x) \ge y - \delta\}\right\} = \inf\{y > f(x) \ge y - \delta\} \in \{y \ge f(x) \ge y - \delta\}$. If δ tends to zero, then $x_O \in \bar{\mathcal{X}}_y$ ($\bar{\mathcal{X}}$ is the closure of $\mathcal{X}$). Hence $x_O \ge \inf \mathcal{X}_y \ge \inf \overleftarrow{\mathcal{X}}_y = {}_O f_O^-(y)$ that contradicts the assumption.

3) We shall show that for fixed y and j $f_j^-(y) = {}_j f_j^-(y)$ if f does not decrease in a neighbourhood of point $f_j^-(y)$ $(j = 0;1)$ and $f_j^-(y) = {}_i f_j^-(y)$, $i \ne j$ if f does not increase.

At the point y f_j is determined simultaneously with two functions ${}_i f_j^-((i = 0,1)$. Only these points y are under consideration.

Let f for example be non-descreasing in a neighbourhood of $f_O^-(y)$ and $x_O = {}_O f_O^-(y) < x_1 = {}_1 f_O^-(y)$. Then $x \ne \inf \mathcal{X}$ and there exist such $\delta_1 \ge 0$ and $\delta_0 > 0$ that $x_1 - \delta_O \notin \overrightarrow{\mathcal{X}}_y$ (otherwise $x_1 \ne \inf \overrightarrow{\mathcal{X}}_y$) and $x_1 + \delta_1 \in \overrightarrow{\mathcal{X}}_y$. So $f_O^-(y) = x_1 = \inf \overrightarrow{\mathcal{X}}_y$ and $f(x_1 - \delta_O) > y \ge f(x_1 + \delta_1)$, that contradicts non-decreasing property of f. ■

It follows from the theorem, that, firstly, the continuity orientation in discontinuity points is determined by the operation itself, and consequently, the restoration of initial function f

in these points by the repeated inversion cannot be guaranteed. This is of no importance in probability schemes (moments of up-crossing in stochastic processes, fiducial distributions, etc), for the null sets can be usually ignored there.

Secondly, it follows from the theorem, that the operation repetition leads to a sharp increase of possible variants of inverse functions, but most of them degenerate into constants. Only two kinds of functions keep shape: either in a trivial case of singularities absence, where all inverse functions variants are equal, or such functions, which posesses but singularities.

We exemplificate it by a repeated inversions of a monotone function. The singularities possible for it are either discontinities of the first kind or intervals of constancy, turning one to other while inversing.

Suppose that f is (for example) non-descreasing. Then

$$I = If = \begin{bmatrix} f_O^- & f_N \\ f_\nu & f_1^- \end{bmatrix}; I^2 = \begin{bmatrix} f_{OO}^- & f_{ON}^- & f_{NO}^- & f_{NN} \\ f_{O\nu}^- & f_{O1}^- & f_{N\nu}^- & f_{N1} \\ f_{\nu O}^- & f_{\nu N}^- & f_{1O}^- & f_{1N} \\ f_{\nu\nu}^- & f_{1\nu}^- & f_{1\nu}^- & f_{11} \end{bmatrix};$$

where constants $\mathrm{Sup}\mathcal{X}$ and $\mathrm{Inf}\mathcal{X}$ are appointed with the degenerate upper f_ν^- and lower f_N^- extreme inversions. The operation repetition leads to tensor multiplication of operator matrices: $I^2 = I \otimes I$ (operator matrix elements multiplication is understanded as a result of their consequent use). The result of theorem 1 for $I^r = I \otimes \ldots \otimes I$ remains with evident changes. That is why only the diagonal elements of $I^r f$ (of principal one, f being non-decreasing and of secondary one, f being non-increasing) will be nondegenerate.

We see the repeated inversions result to depend on the sequence of repetitions. We define the r-th power of repeated inversion operation by a repetition of $|r|$ left inverses, when $r < 0$, and r right ones, when $r > 0$. For the double inversion operator we choose a symbol of double negation $\overset{r}{\Vert} f = \overset{2r}{\rceil} f$.

At last we consider functions, which structure is determined by the type of singularities under consideration (discontinuities and constantness intervals).

Suppose $\mathcal{X}$ be the set of integers and $f(x) = x$. The domain of inverse functions we expand to $\mathcal{Y}^- = R^1$. Then $\overset{+1}{\rceil} x = \lfloor x \rfloor$ is entire of x (the left entier), $\overset{-1}{\rceil} x = \lceil x \rceil$ the nearest integer to x from the right (the right entier). The repeated inversion turns them into each other:

$$\overset{+1}{\rceil} \lceil \; \rceil(x) = \lfloor x \rfloor, \overset{-1}{\rceil} \lfloor \; \rfloor(x) = \lceil x \rceil$$

Suppose, that $r \geq 0$, then $\overset{+r}{\Vert} \lfloor \; \rfloor(x) = \lfloor x + r \rfloor - r$, $\overset{-r}{\Vert} \lceil \; \rceil(x) = \lceil x - r \rceil + r$. It is usefull to consider all notions introduced in the section on this example.

3 The generalized binomial distribution

Let's consider the number of successes (k) in Bernoulli scheme as a function of the number of trials (n), elementary event (ω) being fixed: $k = \xi(n|\omega) = \xi(n)$. Following [1] we describe the partly inverse functions parametrically, by considering the convex of extreme inverses:

$\xi_\alpha^-(k) = \lfloor \alpha\xi_1^-(k) + (1-\alpha)\xi_O(k) \rfloor$, $0 \le a \le 1$. We define $\xi_\alpha^-(0) := \lfloor \alpha\xi_1^-(0) \rfloor$ and $\xi_\alpha^+(n) := \overset{-1}{\rceil} \xi_\alpha^-(n)$. In a probability space of elementary events the equality (in law)

$$\xi_\alpha^-(k) = \xi_\alpha^-(0) + \xi_O^-(k), \qquad (*)$$

is evidently true and the constructions mentioned determine the generalized binomial distributions (exactly, their shift modifications) .[1]

THEOREM 2.

In the Bernoulli scheme:

1. *the distribution of results of repeated inversions doesn't depend on their order;*
2. *the following equalities are valid in law:*

$$\overset{r}{\rceil\!\rceil} \xi_\alpha^+(n) = \xi_\alpha^+(n+r) - r \qquad (r \ge -n)$$

$$\overset{r}{\rceil\!\rceil} \xi_\alpha^-(k) = \xi_\alpha^-(k+r) - r \qquad (r \ge -k)$$

3. $\beta_-^*(n|1,p,\alpha) := P\{\xi_\alpha^-(0) = n\} = q^{\lceil n/\alpha \rceil} - q^{\lceil (n+1)/\alpha \rceil}$.

PROOF.

Firstly we remark, that by the equality (*) $P\{\xi_\alpha^-(k) = n\}$ $= \sum_{t=0}^{n-k} P\{\xi_\alpha^-(0) = t\}\ P\{\xi_O^-(k) = n-t\}$ and thus the first propositions are necessary to be proved only for extreme inverses. Then, as $\xi_O^-(k)$ and $\xi_1^-(k)$ are equal to the minimal and maximal expectation times of the k-th success, so $\xi_1^-(k) = \xi_O^-(k+1) - 1$.

1. By the definition, $\overset{+O}{\rceil\!\rceil} \xi(n) = \overset{+1-1}{\rceil}\ \rceil \xi(n) = \max\{k : \xi_O^-(k) \le n\}$.
$P\{\overset{+O}{\rceil\!\rceil} \xi(n) = k\} = P\{\xi_O^-(k) \le n; \xi_O^-(k+1) \ge n+1\} =$
$\sum_{t=k}^{n} \beta_-(t|k,p) P\{\xi_O^-(1) \ge n-t+1\} = p^k q^{n-k} \sum_{t=1}^{n} C_{t-1}^{k-1} = \beta_+(k|n,p)$.

On the other hand $\overset{-O}{\rceil\!\rceil} \xi(n) = \overset{-1+1}{\rceil}\ \rceil \xi(n) = \min\{k : \xi_1^-(k) \ge n\}$ and $P\{\overset{-O}{\rceil\!\rceil} \xi(n) = k\}$
$= P\{\xi_1^-(k) \ge n;\ \xi_1^-(k-1) \le n-1\} = P\{\xi_O^-(k) \le n;\ \xi_O^-(k+1) \ge n+1\}$
$= \beta_+(k|n,p)$.

2. Let us use the induction wiht respect to r. The case $r = 0$ is already checked up. Suppose that $\overset{r}{\rceil\!\rceil} \xi(n) = \xi(n+r) - r$, then $\overset{r+1}{\rceil\!\rceil} \xi(n) = \max\{k \max\{n : \overset{r}{\rceil\!\rceil} \xi(n) \le k\} \le n\} =$
$\max\{k; \max\{n+r; \xi(n+r) \le k+r\} \le n+r\} = \overset{O}{\rceil\!\rceil} \xi(k+r+1) - (r+1)$.

3. $P\{\lfloor \alpha\xi_1^-(0) \rfloor = n\} = P\{\frac{n}{\alpha} + 1 \le \xi_1^-(0) \le \frac{n+1}{\alpha} + 1\} = q^{\lceil \frac{n}{\alpha} \rceil} - q^{\lceil \frac{n+1}{\alpha} \rceil}$, since $\beta_-^<(x|1,p) =$
$P\{\xi_1^-(0) < x\} = \sum_{t=1}^{\lceil x \rceil - 1} pq^{t-1} = 1 - q^{\lceil x \rceil - 1}$. ■

Another (scale), modifications, marked by asterisk, $\beta_+^*(k|n,p,\alpha)$ and $\beta_-^*(n|k,p,\alpha)$ are defined as the distributions of random variables: $\overset{*}{\xi}_\alpha^-(k) = \xi_\alpha^-(0) + \ldots + \xi_\alpha^-(0)$ - the sum of k independent terms, and $\overset{*}{\xi}_\alpha^+(n) = \overset{-1}{\rceil}\ \overset{*}{\xi}_\alpha^-(n)$.

[1] Further a vertical line divides argument of a function from parameters. The probabilities of usual binomial distributions (positive and negative) are denited by $\beta_+(k|n,p)$ and $\beta_-(n|k,p)$.

The type of these distributions essentially depends on the structure of partiality parameter α. Thus $\alpha = s/m$ being rational, $\xi^{-}_{\alpha}(k)$ describes a time of expectation of k-successes in a scheme, where the trials are aggregated into homogeneous groups of size m, each group is divided into s heterogeneous parts. As a result we have division of time $n = ts + i$ into homogeneous outer t and inner i ($i = 0, ..., s-1$) and thus we deal with the problem of coordinating rhythms. That is why α is interpreted in biosystems as a measure of tolerance of contradicting processes in a system. These considerations lie in the bases of notions of threshold and scale of immunity to invasion (see [1],[2]). The inner time mainly determines the kind of distribution. We shall illustrate it by generalized binomial distribution $\beta^{*}_{-}(n|1,p,\alpha)$ as an example. At first, we introduce the disignations, which allows to present it in a form analogous to a usual geometric distribution.

Suppose $\alpha_n = \lceil n/\alpha \rceil$, $q_n = q^{\alpha_{n+1}-\alpha_n}$ and $p_n = 1 - q_n$, then $q_0 = q^{\alpha_1}, q^{\alpha_{n+1}} = q^{\alpha_n} q_n = q_0 q_1 ... q_n$ and $\beta^{*}_{-}(n|1,p,\alpha) = p_n q^{\alpha_n}$. If $\alpha = s/m$, where $m = sr + l$ and $n = ts + i$; $l, i = 0, ...s-1$, then $\alpha_{n+1} - \alpha_n = \alpha_{i+1} - \alpha_i$, and $q_n = q_i$, i.e. the kind of distribution is mainly determined by the inner time: $\beta^{*}_{-}(ts + i|1,p,\frac{s}{rs+l}) = (q^m)^t q^{\alpha_i} p_i$. In particular, when $s = 1$, $\beta^{*}_{-}(n|k,p,1/m) = \beta_{-}(n|k,1-q^m)$ that means no heterogeneity and scale modifications turn into usual binomial distributions.

The most simple heterogenity corresponds to $s = 2$ ($\alpha = \frac{2}{2h+1}$). In [8] the description of distribution $\beta^{*}_{+}(k|n,p,\alpha)$ is given in this case in terms of random walk. We shall show this distribution to be connected with Fibonacci series.

It can be show, that its generating function is of a kind $h_n(\nu) = q_0 d_n + q_1 d_{n-1}$, where $d_n = \sum_{i=0}^{\lfloor n/2 \rfloor} C^i_{n-i} H_0^{n-2i} H_1^i$ and

$$H_i = \left\{ \begin{array}{ll} (q_i - q_{i+1})\nu, & i < s-1, \\ (q_i - q_{i+1})\nu + q_{i+1}, & i = s-1, \end{array} \right\}$$

When $\nu = 1$ we obtain an expansion

$$1 = \sum_{i=0}^{\lfloor n/2 \rfloor} C^i_{n-i} p_1^{n-2i} q_1^i + q_1 \sum_{i=0}^{\lfloor (n-1)/2 \rfloor} C^i_{n-i-1} p_1^{n-2i} q_1^i.$$

If we choose $p_1 = \Phi_{-} = 0.618...$- the golden section, then $q_1 = 1 - p_1 = 1/\Phi^2_{+} = \Phi^2_{-}$ and the identity is $\Phi^n_{+} = \varphi_n + \Phi_{-}\varphi_{n-1}$, where φ_n is Fibonacci series (1,1,2,3,5,8,13,21,34,55...). Thus, when $s = 2$, $\beta^{*}_{+}(k|n,p,\frac{2}{2h+1})$ can be called a distribution of Fibonacci type.

Finally we remark the explicit form of the generalized binomial distributions to be rather complicated. The recurrent relations, used in practice, are indicated in [4]. If $\alpha = 1$, then all generalized distributions turn into usual ones.

4 The analysis of postsynaptic potentials (PSP) amplitudes distributions

A part of scheme of morfological reconstruction of the interneuron junctions for two neurons (A and B) in an experiment on the isolated spinal hord of a frog ([5],p.40 pic.1) is shown at pic.1. The electrical impulse of cell A irritation passes to the cell B. A potential is being

registrated by an electrode introduced into the cell B. The amplitude η of this potential is the basis characteristics to be investigated. The black points on the scheme marks the places of effective neuron contacts (synapses). The problem is to make clear the logical nature of impulse transmission mechanizm in the interneuron junctions.

In accordance with the well-known quantum theory of neuron impulse transmission [6] our observation has a form $\zeta = Q\eta + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$ is the normal instrument noise and Q is the quantum size. The noise-taking-away techique is discribed in [7]. We are interested in the structure of main signal η. That makes us have a more detail look at a separate synapse. A result of a signale synapse activity can be realy observed, for example, in a neuromuscular junction. A principle scheme of the chemical transmission in synapse is shown at pic.2. A mediator quantum is released from the presynaptic membrane. While mediator is in the contact with the post-synaptic membrane, the chanals of the latter are open for the passage of ions Na^+ and K^+ , which concentrations are different inside and outside of the cell. The redistribution of these ions generates a membrane potential.

The synaps activity in this scheme is determined by two independent Poisson processes: the inner one ($\xi_1(t)$) of mediator accumulation in pre-synaptic region and preporation for its release (being controlled by ion Mg^{2+}), and the outer one ($\xi_2(t)$) of mediator spontaneous release (being controlled by Ca^{2+}). Let λ_1 and λ_2 be Poisson parameters.

The nerve impulse absent, the first process dominate over the second one in a sense that in pre-synaptic region the mediator will be ever enough for its spontaneous release. So $n = \xi_1(t) + \xi_2(t)$ can be considered as fixed. In this case the distribution η corresponds to a positive binom, where $p = \lambda_1/(\lambda_1 + \lambda_2)$ is the mediator quantum release probability $p + q = 1$:

$$P\{\eta = k\} = P\{\xi_2(t) = k | \xi_1(t) + \xi_2(t) = n\} = C_n^k p^k q^{n-k} = \beta_+(k|n,p),$$

Nerve impulse forthcoming to synaps, a mass mediator release occures and the second process begins to dominate sharply over the first one. The self-regulation takes place between the mediator accumulation (n) and its release (k). If this regulation is discribed in terms of mentioned reflections principle, we conserve of binomial law (or that of its Poisson limit) with possible appearance of it modifications discribed. Indeed, after r reflections the number of realized mediator portions $\eta = \overset{r}{\coprod} \xi(t)$ has by theorem 2 a distribution $\beta_+(k+r|n+r,p)$, and the number of portions ready to be released $\eta^- = \overset{r}{\coprod} \xi_O^-(k)$ has a negative binomial $\beta_-(n+r|k+r,p)$ distribution. We note, that the observed mediator release by homogeneous groups is explained by the scale modification $\beta_+^*(k|n,p,\alpha)$ of distribution η, where $\alpha = 1/m$ ($s = 1, m$ - is the group size).

The applications of the generalized binom distributions to the discription of neuron junctions appears to be more important. In that case, as it is seen in pic.1, the registrated total PSP is a sum of heterogeneous contributions from different synaptic groups. This leads to the additional (generation) noises in observations, which the binomial distributions modifications take into account.

Some results in data processing and the morphology observations in the experiment discribed in [8] (including data from pic.1) are shown at the Table 1 for illustration. The general conclusion is that the data concordance with $\beta_+^*(k|n,p,s/m)$ is better, than it is with usual binom and not worse, than with more general scheme of binomial laws mixture, the latter being bad in interpretation. And which is more important, we see that the estimate of s

is close to number of synapses groups observed, that corresponds to the parameter the α structure considered above.

Table 1

Binomial parameters of elementary PSP and
the interneuron junction morphological characteristics

N	Binom			Convolution of two binoms				
	n	p	P	n_1	p_1	n_2	p_2	P
1	20	0.73	0.00	27	0.18	10	0.98	0.37
2	18	0.65	0.00	35	0.12	8	0.95	0.25
3	9	0.51	0.00	21	0.08	4	0.76	0.01
4	16	0.63	0.23	20	0.25	5	0.99	0.95
5	6	0.54	0.02	12	0.13	2	0.84	0.38
6	14	0.55	0.10	30	0.12	4	0.96	0.17
7	7	0.43	0.21	9	0.24	1	0.87	0.26
8	8	0.59	0.02	54	0.08	5	0.79	0.11
9	16	0.65	0.71	1	0.39	15	0.66	0.45
10	6	0.66	0.11	2	0.46	4	0.75	0.00

N	Generalized binom					Morphology
	n	p	r	α	P	*groups of synapses* / synapses
1	23	0.39	-8	$24/28_c$	0.38	23/72
2	32	0.16	-7	$9/39_c$	0.19	9/18
3	12	0.15	-2	$10/16_M$	0.04	11/16
4	24	0.25	-5	$12/13_c$	0.92	11/40
5	7	0.28	-1	$4/5_M$	0.53	4/6
6	21	0.04	-4	$5/40_M$	0.18	5/19
7	9	0.16	-1	$10/16_M$	0.25	11/20
8	8	0.15	-3	$14/36_M$	0.08	14/43
9	16	0.65	+5	$15/60_c$	0.03	15/42
10	6	0.39	-2	$9/21_c$	0.02	8/26

Note. c-shiftal, M-scalous modification of a generalized binom, P- significance of χ^2 - test.

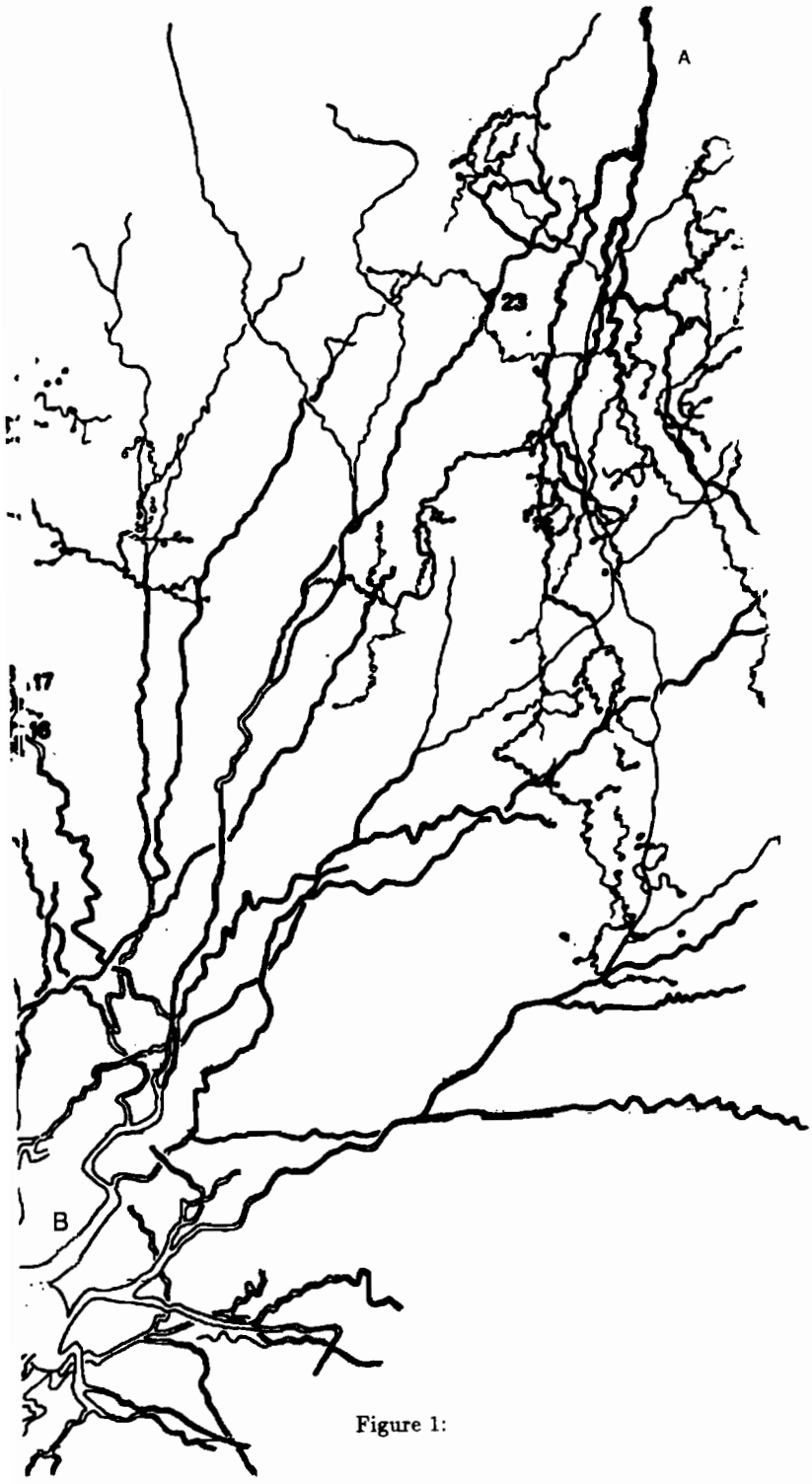


Figure 1:

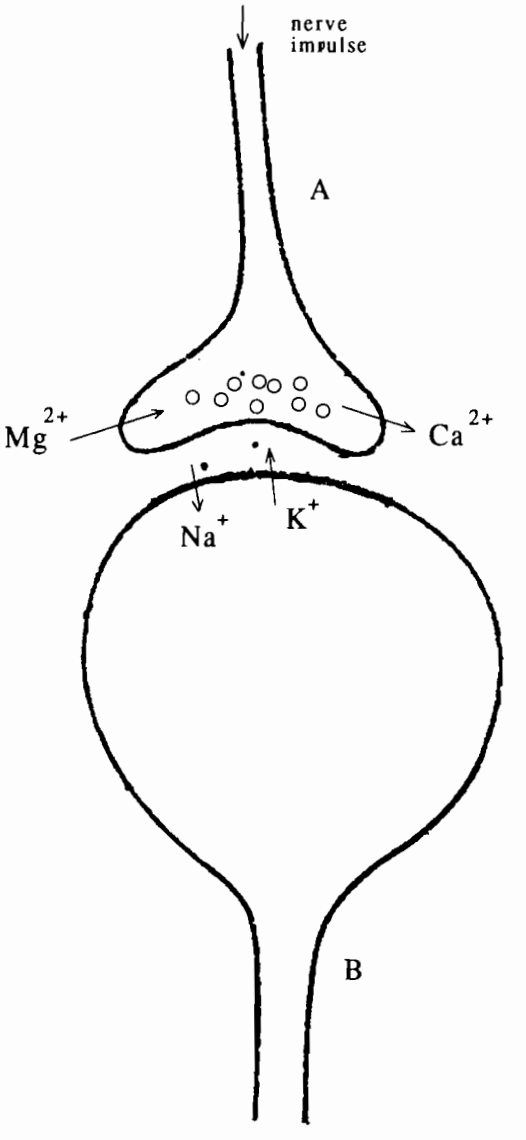


Figure 2: Scheme of the chemical transmission in synapse (o,-mediator)

References

[1] Bart,A.G.(1987).Integerity and the control over the biological systems. Biometrical aspects of the organizm integrity study. M., 141-151.(in Russian)

[2] Bart,A.G., Minar,J. (1984). Basic regulatory parameters of the host-parasite system for warble flies of arm animals using Hypoderma bovis as an example. Folia parazitologica (Praha), 31, 277-287.

[3] Bart,A.G., Dzhafarova,O.A., Zhiharev,S.S., Mineev,V.N. (1991). Probability models of the regulatory mechanizm of membrane-receptory complex im bronchial asthma. Actual problems of pulmonology. St.- Petersburg, 20-25. (in Russian)

[4] Bart,A.G., Dzhafarova,O.A.(1989). Generalized binomial distributions . Vestnik LGU, 1 , 2(8),87-89 (in Russian).

[5] Shapovalov,A.I., Shiriaev,B.I.(1987). The signal transmission in interneurone synapses. Leningrad, Nauka,173.(in Russian)

[6] Katz,B.(1966). Nerve, muscule and synapse. London,220.

[7] Bart,A.G.,Dityatev,A.E.,Kozhanov,V.M.(1988). Quantum analysis of the postsynaptic potentials in the interneuron junctions.Neuroph., 20,4, 479-487(in Russian).

[8] Bart,A.G.,Dityatev,A.E.,Kozhanov,V.M.(1989). Signals transmission analysis in the interneuron synapses based on a reflections principle DAN USSR, 306,61503-1507 (in Russian).

Sign Statistical Methods Software

G.I.Simonova and Yu.N.Tyurin

Theoretical sign statistical results derived by Yu. N. Tyurin are given. New statistical software "SIGN" is based on such results. Problems that can be solved by such software are proposed. Numerical illustrations and comparisons between sign and minimum least squares methods for contaminated samples with outliers are given.

1 Introduction

New statistical sign methods for the solution of the various applied problems are proposed by Yu.N.Tyurin [1]. The assumptions about the statistical properties of the samples are modest. The disturbance terms are restricted only to have independence and zero median. To the contrary classical methods based on a number assumptions whose failure to apply to the data. Sign methods are especially of use to applied statistical data analysis with outliers because they are robust methods. We developed a useful way of computing such methods and set it to software "SIGN". "SIGN" is unique PC software including widespread statistical problems that are solved by the advanced sign statistical methods. Using software "SIGN" you can solve the following statistical problems:

1. Estimate parameters in the linear model.
2. Test any hypothesis in the regression model.
3. Examinate a few samples for gomogeneous.
4. Estimate parameters in one- and two-factor analysis.

Software is based on the algorithms that can be used to the problems when classical methods, such as minimum least squares method, are crucial. These algorithms are obtained by minimizing the some functional wich includes not residuals themselves (as in minimum least squares method) but them signs only. Some simulation results and comparisons between sign and minimum least squares methods for various statistical problems are made. Software is organized as a integrating environment. "SIGN" provides access to graphics from statistical procedures such as estimation of parameters in linear regression, one- and two- factor analysis, testing linear hypotheses. Software is intended for use in a broad range of data processing. Software works with the computer IBM PC with EGA or VGA. The language used is TURBO C. The purpose of this study is to develop a useful way of computing sign statistical procedures such as hypotheses testing and parameters estimating and to demonstrate it by simulation. The paper is organized as follows. Section 2 considers some sign statistical results proposed by Tyurin. The problems to wich such mathematical results applied and their numerical illustrations are given in Section 3. Conclusions are drawn in Section 4.

2 Sign statistical methods

In paper of Yu.N.Tyurin [1] linear model is considered. It can be written in the form

$$x_i = \sum_{j=1}^{r} c_{ij}\theta_j + \varepsilon_i \qquad (1)$$

where $c_{ij}(i = 1, \ldots, n;\ j = 1, \ldots, r)$ - are fixed known constants that form design matrix $C = \|c_{ij}\|$, $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)$ - is vector of the unknown regression coefficients. Error terms ε_i are independent identically distributed random variables with distribution function $F(x)$ that is assumed to be continuous in zero, to have zero median $F(0) = 0.5$, to have first and second derivatives such that $F'(0) > 0$, $F''(0) = 0$. Such assumptions are performed e.g. for symmetric disributions. Results can be generalized for not identically distributed errors with distribution functions that satisfy the mentioned assumptions. If a number of such assumptions are made the problems of deriving statistic for zero hypothesis testing $\{ H_0 : \theta = 0 \}$ contrary to alternatives $\{ H_1 : \theta \neq 0 \}$ are discussed. Optimal local unbiased sign test is derived. The power function of such test has zero derivative in zero (local unbased test) and has maximum average curvature in zero amongst all local unbiased tests. Optimal local unbiased test given by $\{T(X) >$ const $\}$ where

$$T(X) = \sum_{j=1}^{r}(\sum_{i=1}^{n} c_{ij}\text{sign } x_i)^2 \qquad (2)$$

Under zero hypothesis H_0 and above-mentioned regular assumptions distribution of this statistic $T(X)$ is free from distribution law of random errors. It is universal and can be derived for each matrix C. Therefore zero hypothesis test that based on statistic $T(X)$ is nonparametric. This general result gives methods of parameter estimation in model (1). Confidence set for unknown vector Θ given by

$$\{\ \theta :\ \sum_{j=1}^{r}[\sum_{i=1}^{n} c_{ij}\ \text{sign}\ (x_i - \sum_{k=1}^{r} c_{ik}\theta_k)]^2 <\ q_{1-\varepsilon}\ \}$$

where $q_{1-\varepsilon}$ - is quantil of level $1 - \varepsilon$ for random variable

$$q(\xi) = \sum_{j=1}^{r}(\sum_{i=1}^{n} c_{ij}\xi_i)^2$$

and ξ_i is the sequence of mutual independent identically distributed random variables with Bernoulli probability law that is 1 or -1 with equal probability. Point estimate of vector parameter θ is solution of the following extremal problem

$$\theta = argmin \sum_{j=1}^{r}[\sum_{i=1}^{n} c_{ij}\ \text{sign}\ (x_i - \sum_{k=1}^{r} c_{ik}\theta_k)]^2 \qquad (3)$$

3 Problems that can be solved by means of software "SIGN" and numerical examples

We consider the following linear problems: estimation of unknown regression parameters in model (1), hypothesis testing about that expectation of the observation vector lies in the

given subspace, parameter estimation in one- and two-factor analysis. Sign method for solution such problems is used. Feature of this method is that expressions for statistics and minimiziring functions include not residuals of observations from model expectations themselves but them signs only. Therefore we have step functions of vector variables. We used iteration procedures for solving extremal problems. For zero hypothesis testing in not large samples we used Monte-Karlo method. In large samples we used another form of test statistic that has asymptotically X-squared distribution.

Problem 1. Estimation of unknown regression coefficients in the linear model. We can solve extremal problem: find values θ_j $(\theta = 1, \ldots, n)$ wich are solution (3). We proposed the following algorithm using the iterations.

1. Start with an feasible parameter vector $\vec{\theta}(0) = (\theta_1(0), \theta_2(0), \ldots, \theta_r(0))$.
2. k-iteration step consist in calculating vector $\vec{\theta}(k)$ by means of values in k-1 step $\vec{\theta}(k-1)$ with following expression:

$\vec{\theta}(k) =$ Med $\{ \frac{x_i - \sum_{i \neq j}^{r} c_{il}\theta_i(k-1)}{c_{ij}}$, $p_i^j = \frac{|c_{ij}|}{\sum_{i=1}^{n} |c_{ij}|}$, $i = 1, \ldots, n \}$

for $j = 1, \ldots, r$.

3. Exit iteration process if values $\vec{\theta}(.)$ are stable or statistic (3) equals zero.

Here symbol Med$\{ u_i, p_i, i = 1, \ldots, n\}$ denote median of distribution probabilities $p_1, \ldots, p_n$ which is concentrated in the points of the real line $u_1, \ldots, u_n$ accordingly. This definition can be made more precisely.

Let be $F(x) = \sum_{i:u_i \leq x} p_i$, $x \in R$ - distribution function of above-mentioned probability law. Median is solution of equation $F(x) = 0.5$. If a number of values (semiinterval) satisfies to equation $F(x) = 0.5$ we shall choose center of such semiinterval as median.

Example.

Regression model (1) with the known design matrix is created by simulation. It given by $y_i = \theta_0 + \theta_1 x_i + \theta_2 \sin(\frac{2\pi x_i}{25}) + \varepsilon_i$ with $x_i = 0.8i, i = 0, \ldots, 64$ (size of sample equals 65). Errors ε_i are independent identically distributed random variables with distribution function $F(x) = (1-\delta)F_0(x) + \delta\ H(x)$ where F_0 is standard normal distribution and H is distribution with heavy tails. It is sample with outliers ($\delta = 0.2$) besides the disribution of outliers is not symmetrical. We compare two methods of estimation - sign and minimum least squares ones. The following table gives model and estimated parameters.

Parameter	Model	Minimum Least Squares	SIGN-method
θ_0	2.50	439.72	2.47
θ_1	-5.00	-10.78	-5.01
θ_2	100.00	-164.62	99.78

Minimum least squares method is crucial in such example but sign method gives suitable parameter estimates. We considered examples when the data have outliers. But if we have sample with unknown error's distribution such that the data do not meet the classical assumptions we recommend to make use of sign method which is nonparametric one and can be used for anywhere distribution. It is especially true if we want to estimate the confidence intervals.

Problem 2. Linear hypotheses testing.

Let linear model (1) is given. We test zero hypothesis about some regression coefficients θ_j equals zero. Such problem appears e.g. in estimating of useful signal in model (1) if the researcher can represent nonrandom part of signal by means of known basis function as follows

from physical sense. Question is to choose significal with statistical point of view number of basis functions. If we rename coefficients in model (1) vector $\theta = (\theta_1, \ldots, \theta_r)$ can be offered by the two subvectors $\sigma = (\sigma_1, \ldots, \sigma_a)$ and $\tau = (\tau_1, \ldots, \tau_b)$, $a + b = r$, and matrix $C = \|c_{ij}\|$ ($i = 1, \ldots, n; j = 1, \ldots, r$) subdivided in (1) into two submatrixes A and B with orthogonal columns. Therefore model (1) is
$x_i = \sum_{j=1}^{a} a_{ij}\sigma_j + \sum_{k=1}^{b} b_{ik}\tau_k + \varepsilon_i$, $i = 1, \ldots, n$.
Zero hypothesis H_0 now is expressed as $\{ H_0 : \sigma = 0 \}$. Algorithm for testing such hypothesis is:
Step 1. Implying that $\sigma = 0$ we define estimate τ by sign method (problem 1);
Step 2. Let define residuals $\hat{x}_i = x_i - \sum_{k=1}^{b} b_{ik}\hat{\tau}_k$.
Step 3. For random values $\hat{x}_i$ $(i = 1, \ldots, n)$ we apply sign test for zero hypothesis $\{ H_0 : \sigma = 0 \}$: test (2) where we must replace coefficients c_{ij} by a_{ij} and x_i by $\hat{x}_i$.

Problem 3. Estimation in one factor analysis.
1. Data. We have m independent samples of sizes $n_1, \ldots, n_m$ accordingly. Values in ith sample are $x_{1i}, x_{2i}, \ldots, x_{n_i i}$, $i = 1, \ldots, m$.
2.Model is $x_{ij} = \theta_i + \varepsilon_{ij}$, $i = 1, \ldots, m; j = 1, \ldots, n_i$ where $\theta_1, \ldots, \theta_m$ are the unknown parameters and ε_{ij} are mutual independent identically distributed random values.
3. Test for homogeneous.
In order to employ sign analysis to such problem we can introduce another parametric system $\theta_i = \mu + \alpha_i/\sqrt{n_i}$, $i = 1 \ldots, m$ with with constraint

$$\sum_{i=1}^{m} \alpha_i = 0 \tag{4}$$

Hypothesis $\{ H : \alpha_1 = \alpha_2 = \ldots, \alpha_m = 0\}$ is linear. It is equivalent to hypothesis $\{\theta_1 = \theta_2 = \ldots = \theta_m\}$. For zero hypothesis testing we used sign method. We can perform
1) Consider that H is true we can estimate parameter μ. Its sign estimate is $\hat{\mu} = \mathrm{med}(x_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i)$.
2) Make up the residuals $\hat{x}_{ij} = x_{ij} - \hat{\mu}$.
3) Make up the statistic

$$T^2 = \sum_{i=1}^{m}(n_i^{-1}\sum_{j=1}^{n_i}\mathrm{sign}\hat{x}_{ij})^2 - m^{-1}(\sum_{i=1}^{m} n_i^{-1}(\sum_{j=1}^{n_i}\mathrm{sign}\ \hat{x}_{ij})^2 \tag{5}$$

4) Compare sampling statistic T^2 with quantil $t^2_{1-\gamma}$ of level $1 - \gamma$ for random variable
$t^2 = \sum_{i=1}^{m}(n_i^{-1}\sum_{j=1}^{n_i} s_{ij})^2 - m^{-1}(\sum_{i=1}^{m} n_i^{-1}\sum_{j=1}^{n_i} s_{ij})^2$
where s_{ij} $(i = 1, \ldots, m; j = 1, \ldots, n_i)$ are mutual independent identically distributed random variables whose values are 1 or -1 with probabilities 0.5.
4. Estimation μ and α $(i = 1, \ldots, m)$. Estimation follows from the test statistic T^2. We can solve extremal problem
$(\mu, \alpha_i) = argmin\ T^2$ under constraint (4) where $\hat{x}_{ij}$ in (5) replaced by
$x_{ij} - \mu - \alpha_i/\sqrt{n_i}$.

Problem 4. Estimation in the two factor analysis.
We consider this problem in the assumption that the sizes of variables are equal. Problem can be generalized into case of unequal sizes of variables.

1. Data are $N = rtm$ independent values. They are recieved when two factors say A and B have unequal values. Factor A have r levels $A_1, A_2, \ldots, A_r$ and factor B have t levels $B_1, B_2, \ldots, B_t$. For each combination A_iB_j (where $i = 1, \ldots, r; j = 1, \ldots, t$) m repeated observations are made. They are x_{ijk}, $k = 1, \ldots, m$. Often one of the factors (say B) is general. It is called the treatment. Second factor A can be considered as interfered. Its influence divide all observations into blocks.
2. Model is additive for the factors and given by $x_{ijk} = \theta_{ij} + \varepsilon_{ijk}$, $i = 1, \ldots, r; j = 1, \ldots, t; k = 1, \ldots, m$ where $\theta_{11}, \ldots \theta_{rt}$ are unknown parameters, ε_{ijk} are mutual independent identically distributed random errors. The errors are assumed to have $P(\varepsilon_{ijk} > 0) = P(\varepsilon_{ijk} < 0) = 0.5$. Parameters θ_{ij} can be represented in the form $\theta_{ij} = \mu + \alpha_i + \beta_j$ where μ is general level (population mean) from which residuals by A and B factor's actions are made;
α_i is influence of the block A_i , $i = 1, \ldots, r$.
β_j is influence of the treatment B_j , $j = 1, \ldots, t$.
Therefore two factor model can be written in the form $x_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ for $i = 1, \ldots, r; j = 1, \ldots, t; k = 1, \ldots, m$ with constraints $\sum_{i=1}^{r} \alpha_i = 0$, $\sum_{j=1}^{t} \beta_j = 0$.
3. In order to estimate parameters in the two factor model we can used sign test for the following hypothesis testing $\{ H : \mu = 0, \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0, \beta_1 = \beta_2 = \ldots = \beta_t = 0 \}$. Statistic of the optimal local unbiased sign test is $S^2 = \sum_{i=1}^{r} \sum_{j=1}^{t} (Z_{i.} + Z_{.j} - Z_{..})^2$ where $Z_{ij} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \text{sign } x_{ijk}$, $Z_{i.} = \frac{1}{t} \sum_{j=1}^{t} Z_{ij}$, $Z_{.j} = \frac{1}{r} \sum_{i=1}^{r} Z_{ij}$, $Z_{..} = \frac{1}{rt} \sum_{i=1}^{r} \sum_{j=1}^{t} Z_{ij}$. Therefore sign estimates are solition of the following extremal problem $\min T(\mu, \alpha, \beta) = \min \sum_{i=1}^{r} \sum_{j=1}^{t} (\tilde{Z}_{i.} + \tilde{Z}_{.j} - \tilde{Z}_{..})^2$ under constraints $\sum_{i=1}^{r} \alpha_i = 0$, $\sum_{j=1}^{t} \beta_j = 0$ where $\tilde{Z}_{ij} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \text{sign}(x_{ijk} - \mu - \alpha_i - \beta_j)$ and $\tilde{Z}_{i.}$ and $\tilde{Z}_{.j}$ are the average values of Z_{ij} through j and i accordingly and $\tilde{Z}_{..}$ is the average value of $\tilde{Z}_{ij}$ through i and j. Estimate of vector parameter $\theta = (\mu, \alpha, \beta)$ is argmin of the function $T(\mu, \alpha, \beta)$. Such function can be written in the form $T(\mu, \alpha, \beta) = t \sum_{i=1}^{r} \tilde{Z}_{i.}^2 + r \sum_{j=1}^{t} \tilde{Z}_{.j}^2 - rt\tilde{Z}_{..}^2$. Parameter estimates are given by the following iteration algorithm
1. Let fixed β and find parameters μ and α as $(\mu, \alpha) = \text{argmin} \sum_{i=1}^{r} \tilde{Z}_{i.}^2$ under constraint $\sum_{i=1}^{r} \alpha_i = 0$.
2. Let place such α-values into functions $\tilde{Z}_{.j}^2$ and find estimates as $(\mu, \beta) = \text{argmin} \sum_{j=1}^{t} \tilde{Z}_{.j}^2$ under constraint $\sum_{j=1}^{t} \beta_j = 0$.
3. If distance in Euclidean norm between estimates in the two neighbouring iterations greater then little number eps that go to step 1 otherwise iteration process is stopped.

Example for estimating parameters in the two factor analysis and comparison between sign and minimum least squares methods in the contaminating samples are given. In the following table values of the simulated parameters and their estimates derived by sign and minimum least squares methods are given.

Number of values	Parameters	Model	MLS	SIGN
84	α_1	0.5000	1.5042	0.4559
84	α_2	2.0000	2.2441	2.2118
84	α_3	7.8000	4.1048	7.3907
84	α_4	-5.3000	-5.6409	-5.2891
84	α_5	-5.0000	-2.2122	-4.7693
60	β_1	-6.4000	-4.0641	-6.2083
60	β_2	7.3000	4.3219	7.0507
60	β_3	13.5000	11.7733	13.6179
60	β_4	0.0000	5.9295	0.1074
60	β_5	5.4000	-0.1857	5.0072
60	β_6	-2.6000	-0.6527	-2.3961
60	β_7	-17.2000	-17.1224	-17.1788
	μ	1.0000	6.4765	1.1970

Table shows that sign estimates near to the true parameters than minimum least squares estimates.

4 Conclusions

In this paper we presented some problems for hypotheses testing and parameters estimating which was included in software "SIGN". We used new methods and algorithms for such problems. That are sign methods. Software "SIGN" based on such algorithms. Our sampling experiments show that sign method performs better than minimum least squares method if distribution of error terms is unknown or the sample have the outliers.

References

[1] Yu.N.Tyurin "Sign statistical linear analysis". In book: Data analysis, estimation of parameters and social choice in system studies. Issue 14, M.: All-union inst. of system stud. 1986.

A Brief Survey on the Linear Methods in Variance–Covariance Components Models

Júlia Volaufová

1 Introduction

The well known form of a linear variance-covariance components model is often given as

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is an n-dimensional vector, $\beta \in R^k$ is a k-dimensional vector of fixed unknown parameters. The $n \times k$-matrix X is known. The assumptions on the random vector of errors are as follows

$$E(\varepsilon) = 0 \quad E(\varepsilon\varepsilon') = \sum_{i=1}^{p} \vartheta_i V_i.$$

The matrices V_i, $i = 1, \ldots, p$ are given, and the vector $\vartheta = (\vartheta_1, \ldots, \vartheta_p)' \in \Theta$ is considered as unknown parameter. The parameter space is the product $R^k \times \Theta$ and $\Theta \subseteq R^p$ contains a nonempty open set.

Moreover we assume that there exist finite third and fourth moments of the vector ε and the matrices of them are denoted as

$$E(\varepsilon \otimes \varepsilon\varepsilon') = \Phi \quad E(\varepsilon\varepsilon' \otimes \varepsilon\varepsilon') = \Psi.$$

There are many authors interested in the problems of estimating a linear function of the parameter ϑ, say $f'\vartheta$. The basis of their consderations is to construct a quadratic function of the vector Y, say $Y'AY$, which fulfils the required assumptions of optimality. The problem is then restricted to finding the proper matrix A which satisfies the conditions which are connected with e.g. unbiasedness, invariance, etc.. Let us mention some of them: Rao 5, 6, Kleffe 1, 2, Kubáček 3, 4, and many others.

The present contribution is based on the ideas presented by e.g. Seely 9, 10, Verdooren in 11, 12, who suggested to transform the original model to the form of the linear model, where the unknown parameter ϑ represents the vector-parameter of the expectation. This approach enables to use the methods of linear models for estimation of the function $f'\vartheta$ as it was done by Volaufová and Witkovský in 14, and for special structures by Volaufová in 13.

2 Preliminaries to the linear approach

Consider the model (1). For the expectation of the vector $Y \otimes Y$ $(= \operatorname{vec} YY')$ the following equality holds

$$E(\operatorname{vec} YY') = X\beta \otimes X\beta + Q\vartheta,$$

where the matrix $Q = (\operatorname{vec} V_1, \ldots, \operatorname{vec} V_p)$. "$\otimes$" denotes the Kronecker product of matrices, "$\operatorname{vec} A$" denotes the column-vector formed by the columns of the matrix A.

In general the covariance matrix of the vector $\operatorname{vec} YY'$ depends on first four moments of Y and we shall use the notation $\operatorname{var}_\vartheta(\operatorname{vec} YY') = \Sigma(\vartheta)$.

If we have to estimate the parameter function $f'\vartheta$ it is natural to take into account estimators which do not depend on the translation in the mean.

Definition 1 *The statistic $T(Y)$ is invariant under the group of translations in the mean in model (1) if $T(Y) = T(Y - X\alpha)$ for all $\alpha \in R^k$.*

It can be shown that the maximal invariant in model (1) under the group of translations in the mean is the statistic $U = MY$, where M is the orthogonal projection onto the orthogonal complement to the column space of the matrix X, $M = I - XX^+$. The symbol "$^+$" is used for the Moore-Penrose inverse of the matrix.

In the next we shall consider a more general situation, the estimation of a linear function of the type $p'\beta + f'\vartheta$. Consider the vector $(Y', (\operatorname{vec} UU')')'$.

Lemma 1 *The expectation and covariance matrix of the vector $(Y', (\operatorname{vec} UU')')'$ is expressed as follows*

$$E\begin{pmatrix} Y \\ \operatorname{vec} UU' \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & Q_1 \end{pmatrix}\begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \operatorname{cov}\begin{pmatrix} Y \\ \operatorname{vec} UU' \end{pmatrix} = \begin{pmatrix} V(\vartheta) & \Phi_1' \\ \Phi_1 & \Sigma_1(\vartheta) \end{pmatrix},$$

where the matrix $Q_1 = (\operatorname{vec} MV_1M, \ldots, \operatorname{vec} MV_pM)$, $\Phi_1 = (M \otimes M)\Phi$, and the covariance matrix $\Sigma_1(\vartheta)$ of the $\operatorname{vec} UU'$ is in general given by

$$\Sigma_1(\vartheta) = (M \otimes M)\Psi(M \otimes M) - (\operatorname{vec} MV(\vartheta)M)(\operatorname{vec} MV(\vartheta)M)'.$$

PROOF. The proof is straightforward by using the explicite form of the covariance matrix $\Sigma(\vartheta)$ of the vector $\operatorname{vec} YY'$

$$\begin{aligned}\Sigma(\vartheta) &= \Psi - \operatorname{vec} V(\vartheta)(\operatorname{vec} V(\vartheta))' + (X\beta \otimes \Phi')(I + F_{nn}) \\ &\quad + (I + F_{nn})(\beta'X' \otimes \Phi) + (I + F_{nn})(X\beta\beta'X' \otimes V(\vartheta))(I + F_{nn}),\end{aligned}$$

where the matrix F_{nn} is uniquely given by the relation $F_{nn}\operatorname{vec} A = \operatorname{vec} A'$ for each $n \times n$-matrix A. See also 8. □

3 Linear models in parameters β and ϑ

It is easy to see from Lemma 1 that both the expectation and the covariance matrix of the vector $(Y', (\operatorname{vec} UU')')$ depend on unknown vector parameter ϑ.

In special case under the normality of the vector Y it may be possible to consider maximum likelihood estimators of β and ϑ as well, what can lead to complicated nonlinear system of equations.

Another possible approach to the estimating problem is to create a linear model in all unknown parameters of the model and to use a commonly known linear methods.

For that let ϑ_0 be an arbitrary but fixed vector from the parameter space Θ. Let Φ_0 and Ψ_0 be suitable fixed matrices of the third and fourth moments. Let's consider the covariance matrix of the vector $(Y', (\mathrm{vec}\, UU')')$ at ϑ_0, Φ_0, and Ψ_0. It is obvious now that the covariance matrix does not depend on the parameters of expectation. We get the linear model

$$\left(\begin{pmatrix} Y \\ \mathrm{vec}\, UU' \end{pmatrix}, \begin{pmatrix} X & 0 \\ 0 & Q_1 \end{pmatrix} \begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \begin{pmatrix} V(\vartheta_0) & \Phi_{10}' \\ \Phi_{10} & \Sigma_{10}(\vartheta_0) \end{pmatrix} \right). \tag{2}$$

As it will be shown later it is more convenient to fix the parameter ϑ_0 in the original covariance matrix of the vector Y. Let V_0 denotes the matrix $V(\vartheta_0)$. Let us denote $T = V_0 + XX'$. There exists a matrix $T^{+1/2}$ for which the equality $T^+ = T^{+1/2'}T^{+1/2}$ is valid, and moreover the matrix $T^{+1/2}$ is of full rank. The transformed model by the matrix $T^{+1/2}$ is as follows

$$\left(T^{+1/2}Y, T^{+1/2}X\beta, T^{+1/2}V(\vartheta)T^{+1/2'}\right). \tag{3}$$

The maximal invariant under the group of translations in the mean in model (3) is the statistic $Z = M_0T^{+1/2}Y$, where $M_0 = I - T^{+1/2}X(X'T^+X)^-X'T^{+1/2'}$. For the vector $(Y', (\mathrm{vec}\, ZZ')')$ the following lemma is valid.

Lemma 2 *The expectation and covariance matrix of the vector* $(Y', (\mathrm{vec}\, ZZ')')$ *is expressed as follows*

$$E\begin{pmatrix} Y \\ \mathrm{vec}\, ZZ' \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \ \mathrm{cov} \begin{pmatrix} Y \\ \mathrm{vec}\, ZZ' \end{pmatrix} = \begin{pmatrix} V(\vartheta) & \Phi_2' \\ \Phi_2 & \Sigma_2(\vartheta) \end{pmatrix},$$

where $Q_2 = \left(\mathrm{vec}\, M_0T^{+1/2}V_1T^{+1/2'}M_0, \ldots, \mathrm{vec}\, M_0T^{+1/2}V_pT^{+1/2'}M_0\right)$. *The matrix* $\Phi_2 = \left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)\Phi$ *and the covariance matrix* $\Sigma_2(\vartheta)$ *of the* $\mathrm{vec}\, ZZ'$ *is in general given by*

$$\begin{aligned} \Sigma_2(\vartheta) &= \left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)\Psi\left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)' \\ &\quad - \left(\mathrm{vec}\, M_0T^{+1/2}V(\vartheta)T^{+1/2'}M_0\right)\left(\mathrm{vec}\, M_0T^{+1/2}V(\vartheta)T^{+1/2'}M_0\right)'. \end{aligned}$$

The previous lemma implies the next linear model at fixed ϑ_0, Φ_0, and Ψ_0.

$$\left(\begin{pmatrix} Y \\ \mathrm{vec}\, ZZ' \end{pmatrix}, \begin{pmatrix} X & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \begin{pmatrix} V(\vartheta_0) & \Phi_{20}' \\ \Phi_{20} & \Sigma_{20}(\vartheta_0) \end{pmatrix} \right), \tag{4}$$

where $\Phi_{20} = \left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)\Phi_0$.

4 Unbiased and invariant estimability

The previous section and the linear theory implie directly the next two theorems.

Theorem 1 *The linear function* $p'\beta + f'\vartheta$ *is unbiasedly and invariantly estimable*

1. *in the model (2) iff the vectors* $p \in \mathcal{R}(X'X)$ *and* $f \in \mathcal{R}(Q_1'Q_1)$, *where the matrix* $Q_1'Q_1$ *is given by its entries* $\{Q_1'Q_1\}_{i,j} = \mathrm{tr}\,(MV_iMV_j)$,

2. *in the model (4) iff* $p \in \mathcal{R}(X'X)$ *and* $f \in \mathcal{R}(Q_2'Q_2)$, *with* $\{Q_2'Q_2\}_{i,j} = \mathrm{tr}\,(MV_0M)^+V_i(MV_0M)^+V_j$.

Note that due to the inclusion $\mathcal{R}(X) \subseteq \mathcal{R}(T)$ and to the equality $T^{+1/2'}M_0T^{+1/2} = T^+ - T^+X(X'T^+X)^-X'T^+$ the equality $T^{+1/2'}M_0T^{+1/2} = (MV_0M)^+$ holds.

Remark 1 The necessary and sufficient condition of the estimability stated in Theorem 1 *1.* coincides with the well known unbiased estimability of the function $p'\beta$ and unbiased and invariant estimability of the function $f'\vartheta$, see e.g. 8. The *2.* states the unbiased and invariant estimability of the function $f'\vartheta$ which coincides with so called MINQE(U,I) -estimability. It is obvious that the MINQE(U,I)-estimability implies the unbiased invariant estimability.

5 Locally best estimators

Theorem 2 *Let us consider the models (2) and (4), respectively. The locally best linear unbiased estimator at the point* $\vartheta_0 \in \Theta$, Φ_0, *and* Ψ_0 *(LBLUE) of estimable function* $p'\beta + f'\vartheta$ *is given*

1. *in model (2) as*

$$\widehat{p'\beta + f'\vartheta} = (p', f')\left[\begin{pmatrix} X' & 0 \\ 0 & Q_1' \end{pmatrix}\begin{pmatrix} T & \Phi_{10}' \\ \Phi_{10} & T_1 \end{pmatrix}^-\begin{pmatrix} X & 0 \\ 0 & Q_1 \end{pmatrix}\right]^- \times \begin{pmatrix} X' & 0 \\ 0 & Q_1' \end{pmatrix}\begin{pmatrix} T & \Phi_{10}' \\ \Phi_{10} & T_1 \end{pmatrix}^-\begin{pmatrix} Y \\ \mathrm{vec}\,UU' \end{pmatrix}, \tag{5}$$

where $\Phi_{10} = (M \otimes M)\Phi_0$, $T = V_0 + XX'$, *and* $T_1 = \Sigma_{10}(\vartheta_0) + Q_1Q_1'$,

2. *and in model (4) as*

$$\widehat{\widehat{p'\beta + f'\vartheta}} = (p', f')\left[\begin{pmatrix} X' & 0 \\ 0 & Q_2' \end{pmatrix}\begin{pmatrix} T & \Phi_0^{*\prime} \\ \Phi_0^* & T_2 \end{pmatrix}^-\begin{pmatrix} X & 0 \\ 0 & Q_2 \end{pmatrix}\right]^- \times \begin{pmatrix} X' & 0 \\ 0 & Q_2' \end{pmatrix}\begin{pmatrix} T & \Phi_0^{*\prime} \\ \Phi_0^* & T_2 \end{pmatrix}^-\begin{pmatrix} Y \\ \mathrm{vec}\,ZZ' \end{pmatrix}, \tag{6}$$

where $T_2 = \Sigma_{20}(\vartheta_0) + Q_2Q_2' = \left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)T_1\left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)'$.

Remark 2 The words "linear estimator" in the Theorem 2 means linear in the vector $(Y', (\mathrm{vec}\,UU')')$ in the model (2) and in the vector $(Y', (\mathrm{vec}\,ZZ')')$ in the model (4), respectively. In fact the estimators are linear plus quadratic in the original vector Y.

Remark 3 In case that the distribution of the vector ϵ is symmetric around zero the locally best linear estimators of the estimable function $p'\beta$ in both models coincide and have the well known form of ϑ_0-LBLUE

$$\widehat{p'\beta} = p'(X'T^-X)^-X'T^-Y. \tag{7}$$

Under the same condition of symmetry the LBLUE of estimable $f'\vartheta$ does not depend on linear term and it results to a quadratic form in the vector Y. For comparison see 1 and 4.

5.1 Normality of the vector Y

Under the normality condition the covariance matrix of the vector $\operatorname{vec} YY'$ is

$$\Sigma(\vartheta) = (I + F_{nn})(V(\vartheta) \otimes V(\vartheta)) + (I + F_{nn})(X\beta\beta'X' \otimes V(\vartheta))(I + F_{nn}),$$

(see 8) what means that it depends only on the parameter ϑ. Consequently the matrices $\Sigma_1(\vartheta)$ and $\Sigma_2(\vartheta)$ are of the form

$$\Sigma_1(\vartheta) = (M \otimes M)(I + F_{nn})(V(\vartheta)M \otimes V(\vartheta)M)$$

$$\Sigma_2(\vartheta) = \left(M_0T^{+1/2} \otimes M_0T^{+1/2}\right)(I + F_{nn})\left(V(\vartheta)T^{+1/2'}M_0 \otimes V(\vartheta)T^{+1/2'}M_0\right).$$

From that we get the theorem

Theorem 3 *The locally best linear unbiased estimator of estimable function $p'\beta + f'\vartheta$ is under normality of Y given*

1. *in model (2) as*

$$\begin{aligned} \widehat{p'\beta + f'\vartheta} &= \widehat{p'\beta} + \widehat{f'\vartheta} \\ &= p'(X'T^-X)^-X'T^-Y + f'(Q_1'T_1^-Q_1)^-Q_1'T_1^-\operatorname{vec} UU'; \end{aligned} \tag{8}$$

2. *in model (4) as*

$$\begin{aligned} \widehat{\widehat{p'\beta + f'\vartheta}} &= \widehat{\widehat{p'\beta}} + \widehat{\widehat{f'\vartheta}} \\ &= p'(X'T^-X)^-X'T^-Y + f'(Q_2'Q_2)^-Q_2'\operatorname{vec} ZZ' \\ &= \widehat{p'\beta} + \sum_{i=1}^{p} \lambda_i q_i, \end{aligned} \tag{9}$$

where $q_i = Y'(MV_0M)^+V_i(MV_0M)^+Y$ and the vector λ is the solution of the system $(Q_2'Q_2)\lambda = f$.

PROOF. According to the normality condition the estimator (5) directly turns to the form (8).

The estimator (6) can be expressed as

$$\widehat{p'\beta} + f'(Q_2'T_2^-Q_2)^-Q_2'T_2^-\operatorname{vec} ZZ'.$$

The second term occurs to be the locally best "linear" estimator of the function $f'\vartheta$ in the model

$$(\operatorname{vec} ZZ', Q_2\vartheta, \Sigma_{20}(\vartheta_0)). \tag{10}$$

We show that the model (10) fulfils the necessary and sufficient condition for the ordinary least squares estimator to be the locally best linear unbiased estimator, i. e. $\mathcal{R}(\Sigma_{20}(\vartheta_0)Q_2) \subseteq \mathcal{R}(Q_2)$. It is enough to show that

$$\Sigma_{20}(\vartheta_0)\left(\operatorname{vec} M_0T^{+1/2}V_iT^{+1/2'}M_0\right) \in \mathcal{R}(Q_2).$$

$$\Sigma_{20}(\vartheta_0)\left(\operatorname{vec} M_0T^{+1/2}V_iT^{+1/2'}M_0\right) =$$
$$\left(M_0T^{+1/2}\otimes M_0T^{+1/2}\right)(I+F_{nn})\left(\operatorname{vec} V_0(MTM)^+V_i(MTM)^+V_0\right) =$$
$$2\left(\operatorname{vec} M_0T^{+1/2}T(MTM)^+V_i(MTM)^+TT^{+1/2'}M_0\right) =$$
$$2\left(\operatorname{vec} M_0T^{+1/2}V_iT^{+1/2'}M_0\right).$$

From that

$$f'(Q_2'Q_2)^-Q_2'\operatorname{vec} ZZ' = f'(Q_2'T_2^-Q_2)^-Q_2'T_2^-\operatorname{vec} ZZ'.$$

Finally we show that $Q_2'\operatorname{vec} ZZ' = q$, where $q = (q_1,\ldots,q_p)'$ and q_i are given in the statement of the theorem.

$$\begin{aligned} q_i &= (\{Q_2\}_{.i})'\operatorname{vec} ZZ' = \left(\operatorname{vec} M_0T^{+1/2}V_iT^{+1/2'}M_0\right)'\operatorname{vec} ZZ' \\ &= \operatorname{tr}(MTM)^+V_i(MTM)^+YY' = Y'(MV_0M)^+V_i(MV_0M)^+Y, \end{aligned}$$

what completes the proof. □

Remark 4 It is necessary to mention that both the estimator (8) and (9) of the function $p'\beta + f'\vartheta$ are the locally best unbiased linear – (invariant) quadratic in Y, and that implies that they coincide for $f \in \mathcal{R}(Q_2'Q_2)$ (MINQE(U,I) – estimability). That is the reason that we shall concentrate in the next to the model (4).

Remark 5 From the equality $MX = 0$ we can equivalently calculate q_i as

$$q_i = Y'(MTM)^+V_i(MTM)^+Y \quad \text{for } i = 1,\ldots,p. \tag{11}$$

More generally, the matrix $(MTM)^+$ can be replaced by T^-M_T in (11) where $M_T = I - P_T$, and $P_T = X(X'T^-X)^-X'T^-$. For that it is enough to consider the transformation $M_0T^{-1/2}Y$ instead of $M_0T^{+1/2}Y$ in the model (4), where the matrix $T^{-1/2}$ is defined by the relation $T^- = T^{-1/2'}T^{-1/2}$ for arbitrary but fixed g-inverse of the matrix T. That follows that the estimator $\widehat{\widehat{f'\vartheta}}$ is based on the residual vector $M_TY = Y - \widehat{X\beta}$. For more details see 8 page 96.

6 MINQE(U,I) of the $f'\vartheta$

In this section we concentrate on the functions of the form $f'\vartheta$, i.e. $p \equiv 0$. The well known method of estimating functions of that type is the MINQE-theory based by Rao (see e.g.6, 7, and 8). However, the MINQE principle is based on the idea to find a quadratic form, say $Y'AY$, where A is symmetric and minimizes a suitable Euclidean norm, we restrict ourselves, moreover, on the forms which are unbiased and invariant estimators of $f'\vartheta$ (i.e. MINQE(U,I)). In that case the Euclidean norm of the matrix A is of the form $\operatorname{tr} AGAG$, for a suitably chosen matrix G. In case of normality of the vector Y the variance of the statistic $Y'AY$ which is invariant under the group of translations ($Y'AY = (Y - X\alpha)'A(Y - X\alpha)$ for all $\alpha \in R^k$) at a given point ϑ_0 is $2\operatorname{tr} AV_0AV_0$. One of the reasons why to choose the matrix V_0 for G is given e.g. in 8. Nevertheless, we propose a modified definition of the MINQE(U,I).

Definition 2 *The MINQE(U,I) of the estimable function $f'\vartheta$ is the linear-quadratic statistic $T(Y)$ of the form $T(Y) = Y'AY + b'Y + d$ which is unbiased for $f'\vartheta$, invariant under the group of translations $Y \mapsto Y + X\beta$ and minimizes the variance $\mathrm{var}_{\vartheta_0} T(Y)$ under the normality of the vector Y.*

The direct consequence of the Theorem 3 and the Definition 2 is the following proposition.

Proposition 1 *The ordinary least squares estimator in the model (4) is the MINQE(U,I) of the estimable function $f'\vartheta$.*

Corollary 1 *The MINQE(U,I) of the estimable function $f'\vartheta$ is $\widehat{\widehat{f'\vartheta}}$ given in (9).*

7 Linear restrictions on parameters β and ϑ

Let us consider the model, where linear restrictions on the parameters β and ϑ are given, say in the form

$$H\beta = h \quad R\vartheta = c, \tag{12}$$

where the $r_1 \times k$-matrix H and the $r_2 \times p$-matrix R are given full-rank matrices, i.e. $r(H) = r_1$ and $r(R) = r_2$. We use the linear methods to derive the locally best estimators of the function $p'\beta + f'\vartheta$ in model (4) with linear restrictions (12). The linear model is of the form

$$(Y, X\beta \,|H\beta = h, V(\vartheta)\,|R\vartheta = c) \tag{13}$$

The restrictions on β As the first step we shall consider only the linear restrictions on the parameter β, i.e. the model

$$(Y, X\beta \,|H\beta = h, V(\vartheta)) \tag{14}$$

This model can be reformulated as follows

$$\left(\begin{pmatrix} Y \\ h \end{pmatrix}, \begin{pmatrix} X \\ H \end{pmatrix} \beta, \begin{pmatrix} V(\vartheta) & 0 \\ 0 & 0 \end{pmatrix} \right). \tag{15}$$

We shall proceed analogously as in the section 3. Let us denote by

$$W(\vartheta) = \sum_{i=1}^{p} \vartheta_i W_i = \begin{pmatrix} V(\vartheta) & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{where } W_i = \begin{pmatrix} V_i & 0 \\ 0 & 0 \end{pmatrix}; \tag{16}$$

$$T_* = \begin{pmatrix} V(\vartheta_0) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} X \\ H \end{pmatrix} (X', H') = \begin{pmatrix} V(\vartheta_0) + XX' & XH' \\ H'X & H'H \end{pmatrix} \tag{17}$$

being an $(n + r_1) \times (n + r_1)$-matrix. As before we transform the model (14) by the matrix $T_*^{+1/2}$. The maximal invariant in the transformed model under the group of translations is then $Z_* = M_* T_*^{+1/2} (Y', h')'$ with an $(n + r_1) \times (n + r_1)$-matrix

$$M_* = I - T_*^{+1/2'} \begin{pmatrix} X \\ H \end{pmatrix} \left((X', H') T_*^{+} \begin{pmatrix} X \\ H \end{pmatrix} \right)^{-} (X', H') T_*^{+1/2}.$$

The linear model in β, ϑ is then

$$\left(\begin{pmatrix} Y \\ h \\ \operatorname{vec} Z_* Z_*' \end{pmatrix}, \begin{pmatrix} X & 0 \\ H & 0 \\ 0 & Q_* \end{pmatrix} \begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \begin{pmatrix} V(\vartheta_0) & 0 & \Phi_{*0}' \\ 0 & 0 & 0 \\ \Phi_{*0} & 0 & \Sigma_{*0}(\vartheta_0) \end{pmatrix}\right), \tag{18}$$

in which

$$E(\operatorname{vec} Z_* Z_*') = Q_* \vartheta = \left(\operatorname{vec} M_* T_*^{+1/2} W_1 T_*^{+1/2'} M_*, \ldots, \operatorname{vec} M_* T_*^{+1/2} W_p T_*^{+1/2'} M_*\right) \vartheta;$$

the covariance matrix of $\operatorname{vec} Z_* Z_*'$ at ϑ_0 and Ψ_0 is

$$\begin{aligned} \Sigma_{*0}(\vartheta_0) = &\left(M_* T_*^{+1/2} \otimes M_* T_*^{+1/2}\right) \Psi_{*0} \left(M_* T_*^{+1/2} \otimes M_* T_*^{+1/2}\right)' \\ &- \left(\operatorname{vec} M_* T_*^{+1/2} W(\vartheta) T_*^{+1/2'} M_*\right) \left(\operatorname{vec} M_* T_*^{+1/2} W(\vartheta) T_*^{+1/2'} M_*\right)'. \end{aligned}$$

where $\Psi_{*0} = \begin{pmatrix} \Psi_0 & 0 \\ 0 & 0 \end{pmatrix}$ is the $(n+r_1)^2 \times (n+r_1)^2$-matrix and the $(n+r_1)^2 \times (n+r_1)$-matrix Φ_{*0} is $\Phi_{*0} = \left(M_* T_*^{+1/2} \otimes M_* T_*^{+1/2}\right) \begin{pmatrix} \Phi_0 & 0 \\ 0 & 0 \end{pmatrix}$.

The following assertions are the direct consequense of this set up.

Theorem 4 *The linear function $p'\beta + f'\vartheta$ is unbiasedly estimable in model*

$$p \in \mathcal{R}(X'X + H'H), \text{ and } f \in \mathcal{R}(Q_*' Q_*). \tag{19}$$

Theorem 5 *The locally best linear estimator at ϑ_0, Φ_0, and Ψ_0 of the estimable function $p'\beta + f'\vartheta$ in the model (18) is in general*

$$\begin{aligned} &\widetilde{\widetilde{p'\beta + f'\vartheta}} = \\ &(p', f') \left[\begin{pmatrix} X' & H' & 0 \\ 0 & 0 & Q_*' \end{pmatrix} \begin{pmatrix} T & XH' & \Phi_{*0}' \\ HX' & HH' & 0 \\ \Phi_{*0} & 0 & T_{*1} \end{pmatrix}^- \begin{pmatrix} X & 0 \\ H & 0 \\ 0 & Q_* \end{pmatrix}\right]^- \\ &\times \begin{pmatrix} X' & H' & 0 \\ 0 & 0 & Q_*' \end{pmatrix} \begin{pmatrix} T & XH' & \Phi_{*0}' \\ HX' & HH' & 0 \\ \Phi_{*0} & 0 & T_{*1} \end{pmatrix}^- \begin{pmatrix} Y \\ h \\ \operatorname{vec} Z_* Z_*' \end{pmatrix}, \end{aligned} \tag{20}$$

where $T = V_0 + XX'$ and $T_{*1} = \Sigma_{*0}(\vartheta_0) + Q_* Q_*'$.

Remark 6 Under the condition of symmetry the estimator $\widetilde{\widetilde{p'\beta + f'\vartheta}} = \widetilde{\widetilde{p'\beta}} + \widetilde{\widetilde{f'\vartheta}}$. It is easy to verify that under the additional condition $\mathcal{R}(H') \subseteq \mathcal{R}(X')$ the following holds.

$$\widetilde{\widetilde{p'\beta}} = \widehat{p'\beta} - p' \left(X'T^- X\right)^- H' \left(H \left(X'T^- X\right)^- H'\right)^{-1} \left(\widehat{H\beta} - h\right).$$

Theorem 6 *The MINQE(U,I) of the estimable function $f'\vartheta$ in the model (14) is*

$$\widehat{\widehat{f'\vartheta}} = \sum_{i=1}^{p} \lambda_i q_{*i} = \sum_{i=1}^{p} \lambda_i \left((Y', h')(M_H W_0 M_H)^+ W_i (M_H W_0 M_H)^+ \begin{pmatrix} Y \\ h \end{pmatrix} \right), \qquad (21)$$

where $M_H = I - \begin{pmatrix} X \\ H \end{pmatrix} (X'X + H'H)^- (X', H')$ *and vector* λ *is the solution of the system* $(Q_*' Q_*)\lambda = f$.

PROOF. According to the Definition 2 it is enough to show that the estimator (21) is invariant ϑ_0-LBLUE of the function $f'\vartheta$ under the normality of the vector Y. Due to the Remark 6 and the fact $p \equiv 0$ the LBLUE of the function $f'\vartheta$ from the Theorem 5 is

$$\widehat{\widehat{f'\vartheta}} = f' \left(Q_*' T_{*1}^- Q_* \right)^- Q_*' T_{*1}^- \mathrm{vec}\, Z_* Z_*',$$

under the normality the matrix $\Sigma_{*0}(\vartheta_0)$ has the form

$$\Sigma_{*0}(\vartheta_0) = \left(M_* T_*^{+1/2} \otimes M_* T_*^{+1/2} \right) (I + F_{(n+r_1)(n+r_1)})(W_0 \otimes W_0) \left(T_*^{+1/2'} M_* \otimes T_*^{+1/2'} M_* \right).$$

By the analogous argumentation as in the Theorem 3 it is possible to show that $\Sigma_{*0}(\vartheta_0) Q_* = 2Q_*$ (i.e. $\mathcal{R}(\Sigma_{*0}(\vartheta_0) Q_*) \subseteq \mathcal{R}(Q_*)$). From that

$$f' (Q_*' Q_*)^- Q_*' \mathrm{vec}\, Z_* Z_*' = f' \left(Q_*' T_{*1}^- Q_* \right)^- Q_*' T_{*1}^- \mathrm{vec}\, Z_* Z_*'.$$

To end the proof it is enough to denote the vector $f' (Q_*' Q_*)^-$ by λ and the vector $Q_*' \mathrm{vec}\, Z_* Z_*'$ by q_*. □

Remark 7 As a matter of fact, the MINQE(U,I) of a function $f'\vartheta$ in model with linear restrictions on the parameter β is not a purely quadratic form. If we denote a block-matrix $(M_H W_0 M_H)^+$ as follows

$$(M_H W_0 M_H)^+ = \begin{pmatrix} C_1 & C_2 \\ C_2' & C_3 \end{pmatrix}$$

then the MINQE(U,I) may be expressed as

$$\widehat{\widehat{f'\vartheta}} = Y' C_1 \left(\sum_{i=1}^{p} \lambda_i V_i \right) C_1 Y + 2h' C_2' \left(\sum_{i=1}^{p} \lambda_i V_i \right) C_1 Y + h' C_2' \left(\sum_{i=1}^{p} \lambda_i V_i \right) C_2 h.$$

If $\mathcal{R}(H') \subseteq \mathcal{R}(X')$ the explicite form for C_1, C_2, and C_3 can be found in e.g. 8.

The restrictions on β and ϑ simultaneously In this paragraph we consider the most general case presented in the model (13). The linear model only with restrictions on ϑ was investigated in 14. We establish our considerations on the model (18) with additional restrictions on the vector ϑ, i.e. $R\vartheta = c$. ¿From that we get the model

$$\left(\begin{pmatrix} Y \\ h \\ \mathrm{vec}\, Z_* Z_*' \\ c \end{pmatrix}, \begin{pmatrix} X & 0 \\ H & 0 \\ 0 & Q_* \\ 0 & R \end{pmatrix} \begin{pmatrix} \beta \\ \vartheta \end{pmatrix}, \begin{pmatrix} V(\vartheta_0) & 0 & \Phi_{*0}' & 0 \\ 0 & 0 & 0 & 0 \\ \Phi_{*0} & 0 & \Sigma_{*0}(\vartheta_0) & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right). \qquad (22)$$

Simple conclusions are the sequel results.

Theorem 7 *The linear function $p'\beta + f'\vartheta$ is unbiasedly estimable in model (22) if and only if*

$$p \in \mathcal{R}(X'X + H'H) \text{ and } f \in \mathcal{R}(Q_*'Q_* + R'R). \tag{23}$$

Theorem 8 *The locally best linear estimator at ϑ_0, Φ_0, and Ψ_0 of the estimable function $p'\beta + f'\vartheta$ in the model (22) is in general*

$$\begin{aligned}\overline{p'\beta + f'\vartheta} = {} & \\ & (p', f') \left[\begin{pmatrix} X' & H' & 0 & 0 \\ 0 & 0 & Q_*' & R' \end{pmatrix} \begin{pmatrix} T & XH' & \Phi_{*0}' & 0 \\ HX' & HH' & 0 & 0 \\ \Phi_{*0} & 0 & T_{*1} & Q_*R' \\ 0 & 0 & RQ_*' & RR' \end{pmatrix}^- \begin{pmatrix} X & 0 \\ H & 0 \\ 0 & Q_* \\ 0 & R \end{pmatrix} \right]^- \\ & \times \begin{pmatrix} X' & H' & 0 & 0 \\ 0 & 0 & Q_*' & R' \end{pmatrix} \begin{pmatrix} T & XH' & \Phi_{*0}' & 0 \\ HX' & HH' & 0 & 0 \\ \Phi_{*0} & 0 & T_{*1} & Q_*R' \\ 0 & 0 & RQ_*' & RR' \end{pmatrix}^- \begin{pmatrix} Y \\ h \\ \operatorname{vec} Z_*Z_*' \\ c \end{pmatrix},\end{aligned} \tag{24}$$

where $T = V_0 + XX'$ and $T_{*1} = \Sigma_{*0}(\vartheta_0) + Q_*Q_*'$.

Corollary 2 *Under the condition of symmetry of the distribution of the vector ε and further under $\mathcal{R}(H') \subseteq \mathcal{R}(X')$ and $\mathcal{R}(R') \subseteq \mathcal{R}(Q_*')$ the locally best unbiased linear-quadratic estimator at ϑ_0 and Ψ_0 of the estimable function $p'\beta + f'\vartheta$ has the form*

$$\begin{aligned}\overline{p'\beta + f'\vartheta} &= \overline{p'\beta} + \overline{f'\vartheta} = \widehat{p'\beta} - p'\left(X'T^-X\right)^- H' \left(H\left(X'T^-X\right)^- H'\right)^{-1} \left(\widehat{H\beta} - h\right) \\ &\quad + \widetilde{\widetilde{f'\vartheta}} - f'\left(Q_*'T_{*1}^-Q_*\right)^- R' \left(R\left(Q_*'T_{*1}^-Q_*\right)^- R'\right)^{-1} \left(\widetilde{\widetilde{R\vartheta}} - c\right).\end{aligned} \tag{25}$$

The estimator $\widehat{p'\beta}$ is defined by (7) and the $\widetilde{\widetilde{f'\vartheta}}$ is the locally best estimator of $f'\vartheta$ in model (18), see also Remark 6.

Theorem 9 *The MINQE(U,I) of the estimable function $f'\vartheta$ in the model (13) is*

$$\overline{f'\vartheta} = \sum_{i=1}^{p} \kappa_i q_i^\circ, \tag{26}$$

where the vector

$$q^\circ \equiv (Q_*', R') \begin{pmatrix} I + Q_*Q_*' & Q_*R' \\ RQ_*' & RR' \end{pmatrix}^- \begin{pmatrix} \operatorname{vec} Z_*Z_*' \\ c \end{pmatrix}$$

and κ is the solution of the system

$$\left[(Q_*', R') \begin{pmatrix} I + Q_*Q_*' & Q_*R' \\ RQ_*' & RR' \end{pmatrix}^- \begin{pmatrix} Q_* \\ R \end{pmatrix}\right] \kappa = f.$$

PROOF. Let us consider the locally best linear estimator (24) of the function $p'\beta + f'\vartheta$ in the model (22). In our case $p \equiv 0$. Under the normality of Y $\Phi_* = 0$ and $\Sigma_{*0}(\vartheta_0)Q_* = 2Q_*$ what implies that the estimator (24) of $f'\vartheta$ turns to (26). Due to the Definition 2 it is MINQE(U,I) of $f'\vartheta$ what completes the proof. □

Remark 8 Another possible way to get MINQE(U,I) of the estimable function $f'\vartheta$ in the model (13) is to realize that the MINQE(U,I) is $f'\overline{\vartheta}$, where $\overline{\vartheta}$ is the solution of the system

$$\begin{aligned} Q'_*Q_*\vartheta + R'\nu &= q_* \\ R\nu &= c, \end{aligned}$$

where $q_* = (q_{*1}, \ldots, q_{*p})'$, $q_{*i} = \left((Y', h')(M_H W_0 M_H)^+ W_i (M_H W_0 M_H)^+ \begin{pmatrix} Y \\ h \end{pmatrix}\right)$, and ν is the vector of Lagrangian multipliers.

Corollary 3 *Under the condition $\mathcal{R}(R') \subseteq \mathcal{R}(Q'_*)$ the MINQE(U,I) of $f'\vartheta$ is given by*

$$\widetilde{\overline{f'\vartheta}} = \widetilde{f'\vartheta} - f' (Q'_*Q_*)^- R' \left(R (Q'_*Q_*)^- R'\right)^{-1} (R(Q'_*Q_*)^- q_* - c) .$$

Under the normality of Y the variance of MINQE(U,I) at ϑ_0 is

$$\operatorname{var}_{\vartheta_0}(\overline{f'\vartheta}) = 2f' \left[(Q'_*Q_*)^- - (Q'_*Q_*)^- R' \left(R (Q'_*Q_*)^- R'\right)^{-1} R(Q'_*Q_*)^-\right] f.$$

For more details the reader is referred to 15.

References

[1] J.Kleffe. Simultaneous estimation of expectation and covariance matrix in linear models. *Math. Operationsforsch. Statist., Series Statistics*, 9(3):443–478, 1978.

[2] J.Kleffe. On recent progress of MINQUE theory — nonnegative estimation, consistency,asymptotic normality and explicite formulae. *Math. Operationsforsch. Statist., Series Statistics*, 11(4):563–588, 1980.

[3] L.Kubáček. Comment on C.R. Rao's MINQUE for replicated observations. *Mathematica Slovaca*, 35(2):131–136, 1985.

[4] L.Kubáček. Locally best quadratic estimators. *Mathematica Slovaca*, 35(4):393–408, 1985.

[5] C.R. Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical AssociationTheory and Methods Section*, 329(63):161–172, 1970.

[6] C.R. Rao. Estimation of variance and covariance components — MINQUE theory. *Journal of Multivariate Analysis*, 1:257–275, 1971.

[7] C.R. Rao. Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association, Theory and Methods Section*, 67(337):112–115, 1972.

[8] C.R. Rao and J.Kleffe. *Estimation of Variance Components and Applications*, volumeß3 of *Statistics and probability*. North-Holland, Amsterdam, New York, Oxford, Tokyo, first edition, 1988.

[9] J.Seely. Linear spaces and unbiased estimation. *Ann. Math. Stat.*, 41:1725–1734, 1970.

[10] J.Seely. Linear spaces and unbiased estimation – Application to a mixed linear model. *Ann. Math. Stat.*, 41:1735–1745, 1970.

[11] L.R. Verdooren. Practical aspects of variance component estimation. Invited lecture for the 4th International Summer School on Problems of Model Choice and Parameter Estimation in Regression Analysis. Mülhausen, GDR., May 1979.

[12] L.R. Verdooren. Least squares estimators and non-negative estimators of variance components. *Commun. Statist.-Theory Meth.*, 17(4):1027–1051, 1988.

[13] J.Volaufová. MINQUE of variance components in replicated and multivariate linear model with linear restrictions. *QÜESTIIÓ*, Submitted for publication.

[14] J.Volaufová and V.Witkovský. Estimation of variance components in mixed linear models. *Applications of Mathematics*, 37(2):139–148, 1992.

[15] V.Witkovský. Testing linear hypotheses in mixed linear model (In Slovak). Minimová práca, Ústav merania SAV, Bratislava, November 1991.

PART III. STOCHASTIC OPTIMIZATION

Chaotic Behaviour of Search Algorithms: Introduction

Henry P. Wynn and Anatoly A. Zhigljavsky

1 Introduction

Certain search and optimization algorithms exhibit chaotic local behavior which is often masked by the simple requirement of convergence. If the algorithm is considered as a dynamic system in its original form convergence means that there is a single attractor and the behaviour might seem uninteresting. The device used in this paper is to rescale, or renormalize, the search region at each iteration and to observe the behaviour of the (unknown) target in this renormalized region. Thus the original dynamic process is converted into a new one in which the starting point is the target itself and the trajectory is the trajectory of the target in the normalized region (rather than the trajectory of the trial values).

This is most simply seen in the case of elementary bifurcation search. Thus let the base-2 expansion of a real number $x^* \in [0,1]$ be

$$x^* = \sum_{i=1}^{\infty} \frac{a_i}{2^i} \qquad (\text{ where } a_i = 0,1\ )$$

Bifurcation search lays down at the n-th iteration the interval

$$x_n^- = \sum_{i=1}^{n-1} \frac{a_i}{2^i} \le x^* \le \sum_{i=1}^{n-1} \frac{a_i}{2^i} + \frac{1}{2^n} = x_n^+$$

and tests whether

$$x_n^- \le x^* \le x_n^- + \frac{1}{2^{n+1}}$$

or

$$x_n^- + \frac{1}{2^{n+1}} < x^* \le x_n^+$$

This is clearly equivalent to exhibiting the value of a_{n+1} as 0 or 1. Rescaling the interval $[x_n^-, x_n^+]$ back to $[0,1]$ induces the transformation

$$x^* \to \frac{x^* - x_n^-}{x_n^+ - x_n^-} = x_n^*.$$

And x_n^* is computed as

$$x_n^* = 2^n \sum_{i=n}^{\infty} \frac{a_i}{2^i}.$$

Taking $x_0^* = x^*$ we can investigate the behaviour of the sequence x_n^* in $[0,1]$. (Of course, this is a well known and classical problem.)

Consider first the transformation

$$h: \; x_n^* \to x_{n+1}^*.$$

It can be written as

$$h(u) = \begin{cases} 2u & \text{if } 0 \le u \le \frac{1}{2} \\ 2u - 1 & \text{if } \frac{1}{2} < u \le 1 \end{cases}$$

This mapping is the simple Bernoulli shift, see Figure 1.

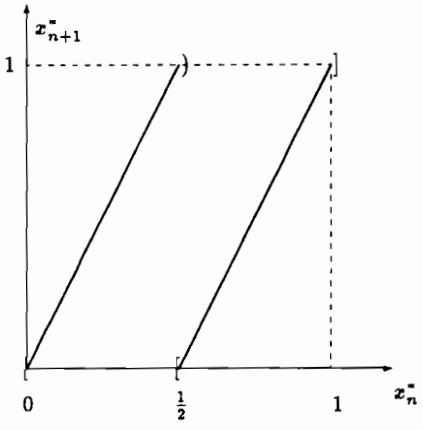


Figure 1. Bernoulli shift.

Considered as a dynamic process the iterative scheme

$$x_{n+1}^* = h(x_n^*)$$

has the invariant measure that is uniform in $[0,1]$. The Ljapunov exponent of the scheme is $\log 2 \simeq 0.693 > 0$ and the sequence x_n^* can be described as chaotic. Another issue is to count the proportion of 0's and 1's in the sequence a_n. Classical results are that the set of numbers (normal numbers) x^* for which the proportion is $\frac{1}{2}$ has the Lebesgue measure 1 and the set of numbers for which the proportion is $\alpha \neq \frac{1}{2}$ has the Hausdorff dimension

$$-\frac{1}{\log 2}(\alpha \log \alpha + (1-\alpha)\log(1-\alpha)).$$

If the original scheme is considered as a search procedure, for example arising from finding the root x^* of a monotonic function $f(x)$ on $[0,1]$, the procedure is well-known to be minimax in the following sense. Let $[x_n^-, x_n^+)$ now be a general interval in which x^* is known to lie, test whether

$$x_n^- \le x^* < x_n^- + \epsilon_n$$

or

$$x_n^- + \epsilon_n \leq x^* < x_n^+$$

and create the new interval as

$$\begin{aligned}[x_{n+1}^-, x_{n+1}^+) &= [x_n^-, x_n^- - \epsilon) \\ &= [x_n^- - \epsilon, x_n^+)\end{aligned}$$

accordingly. Then

$$\min_{\{\epsilon_i\}_1^n} \max_{x^*} (x_n^+ - x_n^-)$$

is achieved for $\epsilon_i = \frac{1}{2^i}$. That is, the constant rate algorithm with rate $\frac{1}{2}$ is the best.

The paper is an introduction to the consequences of the renormalization idea. Our starting point for a more detailed investigation revealing many of the main ideas in the classical Golden Section algorithm for searching for the maximum of a unimodal function.

2 The Golden Section algorithm

We shall consider the problem of finding the minimum of a uni-extremal function $f(x)$ on an interval $[a, b]$. The golden section algorithm is "second order" in the sense that two points x' and x'' with function values $f(x')$ and $f(x'')$ are used at each iteration to eliminate a part of the interval, see Kiefer (1953,1957). It belongs to a family which may be called symmetric optimization algorithms. Let $[a, b]$ be the current interval. At each iteration points x' and x'' are symmetrically placed between a and b:

$$a \leq x' < x'' \leq b, \qquad x' + x'' = a + b \tag{1}$$

If

$$f(x') < f(x'')$$

then $[a, b]$ is reduced to $[x', b]$ because there is no minimizer in $[a, x']$ and conversly if

$$f(x') \geq f(x'')$$

then $[a, b]$ is reduced to $[a, x'']$, see Figure 2. (For a moment we ignore the complications arising in the case $f(x') = f(x'')$. Where we can delete either or both ends from the interval $[a, b]$, we shall delete the right one.)

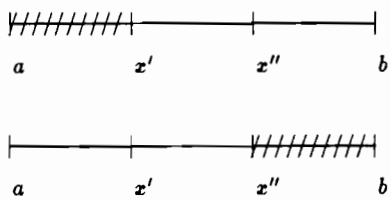


Figure 2. The deletion in the symmetric algorithms.

The golden section algorithm selects x', x'' so that the ratio of the length of the whole interval $[a, b]$ to the bigger subinterval $[a, x'']$ is the same as the length ratio of the bigger to the smaller subintervals:

$$\frac{b-a}{x''-a} = \frac{x''-a}{x'-a}$$

This together with the symmetry condition (1) easily leads to

$$\frac{x''-a}{b-a} = \frac{b-x''}{b-x'} = \lambda = \frac{\sqrt{5}-1}{2} \simeq 0.61804\ldots,$$

$$\frac{x'-a}{b-a} = \frac{b-x''}{b-a} = 1-\lambda \simeq 0.38196\ldots$$

Here λ is derived as the positive root of the equation $\lambda^2 + \lambda - 1 = 0$ and is known as the Golden Section. (The number $1 + \lambda$ sometimes also has the same name.)

The above formulae give a constant rate of reduction of the size of the interval (the analogue of $\frac{1}{2}$ for the bifurcation algorithm). After n iterations the initial interval is reduced by a ratio λ^{n-1}, noting that the algorithm starts with the initial points.

We now produce an iterative scheme under renormalization for the golden section algorithm. Suppose that x^* in $[0, 1]$ is the (unknown) minimizer of $f(x)$. Assume also that $f(x)$ is uniextremal and symmetric about x^*. Thus for any $\delta > 0$ such that $0 < x_* - \delta < x^* + \delta \leq 1$ we suppose that

$$f(x^* - \delta) = f(x^* + \delta)$$

and

$$f(x') > f(x'') \quad \text{if} \quad x^* \leq x' < x'' \leq 1$$
$$f(x') < f(x'') \quad \text{if} \quad 0 \leq x' < x'' \leq x^*$$

Let $[a_n, b_n]$ be the undeleted interval after the n-th iteration. Renormalize $[a_n, b_n]$ back to $[0, 1]$, using the transformation

$$x \to \frac{x - a_n}{b_n - a_n}$$

Then define

$$x_n^* = \frac{x^* - a_n}{b_n - a_n}$$

as the location of the unknown minimizer in the renormalized interval. This defines an iterative scheme

$$x_{n+1}^* = h(x_n^*) \tag{2}$$

where

$$h(u) = \begin{cases} u(1+\lambda) & \text{if } u \leq \frac{1}{2} \\ u(1+\lambda) - \lambda & \text{if } u > \frac{1}{2} \end{cases} \tag{3}$$

The constant rate λ and the symmetry of f implies that $h(.)$ does not depend otherwise on f.

Figure 3 shows this transformation.

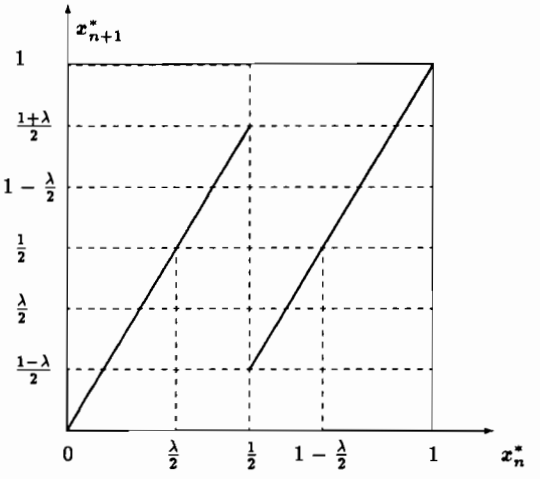


Figure 3. The golden section iteration.

The sequence (2) is ergodic. The Ljapunov exponent of the mapping $h(.)$ is $\log(\lambda+1) \simeq 0.4812 > 0$, thus the behaviour of the sequence (2) can be described as chaotic. Particularly, it has an invariant measure , say ν, such that

$$\lim_{k\mapsto\infty} \frac{1}{k} \sum_{i=1}^{k} g(x_i) = \int_0^1 g(x)\nu(dx)$$

for any continuous function g on $[0,1]$ and ν-almost all starting points $x_0 \in [0,1]$. The invariant measure ν will be computed below in Theorem 1. First, we shall discuss which initial points x_0 generate cycles in the sequence $\{x_n\}$

If $x_k = \lambda$ or $1-\lambda$ for some k then the points x_n for $n \geq k$ perform a cycle in the two-point set $\{\lambda, 1-\lambda\}$, see Figure 4. This case happens when at some moment k the trial point (i.e. the point where we evaluate the objective function) coincides with the minimizer x^*. It is easy to understand that this occurs if and only if for some $k \geq 0$ the minimizer x^* can be represented as

$$x^* = \sum_{i=1}^{k} a_i \lambda^i, \qquad a_i = 0, 1. \tag{4}$$

Another case, mentioned above, is when

$$f(a_k) = f(b_k) \tag{5}$$

for some k, that is to say the function f is symmetric about $x^* = (a_k + b_k)/2$. We have two choices: (i) delete both ends of the interval and restart the algorithm with two additional points and continue like this, (ii) use some convention to delete the left or rigth side of the

interval; for example always delete the right side. Defining the mapping (3) we have used the second choice with the convention to delete the right side of the intervals in the case (5).

This phenomena happens for the countable set of points x^* in $[0,1]$ with the expansion

$$x^* = \sum_{i=1}^{k} a_i \lambda^i + \frac{1}{2}\lambda^{k+1}, \quad a_i = 0,1, \tag{6}$$

for some k. It is easy to see that the points (6) force the sequence $\{x_n\}$ to cycle on the three-point set $\{\frac{1}{2}, \frac{1}{2}-\frac{\lambda}{2}, \frac{\lambda}{2}\}$ for $n > k$. Note that if we change the convention in (ii) above always deleting the left sides of the intervals in cases (5) we obtain the three point set $\{\frac{1}{2}, \frac{1}{2}+\frac{\lambda}{2}, 1-\frac{\lambda}{2}\}$ as the attractor of the sequence $\{x_n\}$.

As is usual, all the points of any n-point attractor set can be obtained as the solutions of the equation $x = h^n(x)$ where $h^n(.)$ is the n-th iterate of the mapping h. The expressions for the starting points leading to these attractors can be derived analogously with (4) and (6). It is worth noting that only a countable number of starting points lead to cycling in the sequence $\{x_n^*\}$.

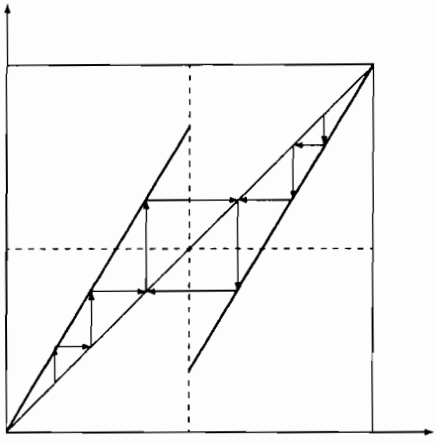


Figure 4. Cyclic attractor for Golden Section.

It is clear from observing the behavior outside the interval $D = [\frac{1-\lambda}{2}, \frac{1+\lambda}{2}]$ that the sequence $\{x_n\}$ attracts to this interval rather than $[0,1]$ and that therefore the invariant measure can not be supported outside D. The following theorem presents the explicit form of the invariant measure of the chaotic sequence (2).

Theorem 1. The normalized invariant measure ν of the Golden section renormalized iterative scheme (2) has the density (with respect to the Lebesgue measure on $[0,1]$) $p(x)$

given by

$$p(x) = \begin{cases} 0 & \text{if } 0 \le x < \frac{1-\lambda}{2} \\ \frac{\lambda}{3-4\lambda} & \text{if } \frac{1-\lambda}{2} \le x < \frac{\lambda}{2}, \\ \frac{1}{3-4\lambda} & \text{if } \frac{\lambda}{2} \le x < 1 - \frac{\lambda}{2} \\ \frac{\lambda}{3-4\lambda} & \text{if } 1 - \frac{\lambda}{2} \le x < \frac{1+\lambda}{2} \\ 0 & \text{if } \frac{1+\lambda}{2} \le x \le 1 \end{cases} \tag{7}$$

Proof. The uniqueness of the normalized invariant measure ν of the sequence x_n absolutely continuous with respect to the Lebesgue measure μ follows from standard ergodic theory and the fact that the mapping $h(.)$ is measure preserving with respect to μ in the domain of attraction D. The condition for invariance is that

$$\nu(h^{-1}(A)) = \nu(A) \tag{8}$$

for any measurable set $A \subseteq D$ with $\mu(A) > 0$.

Let

$$h^{-1}(y) = \{x_i \mid h(x_i) = y\}$$

which in this case has cardinality 1 or 2. Then the invariance condition (8) can be rewritten as the following condition with respect to the density $p(.)$ of the measure ν:

$$p(y) = \sum_{x_i \in h^{-1}(y)} \frac{p(x_i)}{\frac{dh}{dx}\big|_{x=x_i}}.$$

This leads to the functional equations

$$p(y) = \begin{cases} 0 & \text{if } 0 < y < (1-\lambda)/2, \\ \lambda p(\lambda y + 1 - \lambda) & \text{if } (1-\lambda)/2 < y < \lambda/2, \\ \lambda(p(\lambda y) + p(\lambda y + 1 - \lambda)) & \text{if } \lambda/2 < y < 1 - \lambda/2, \\ \lambda p(\lambda y) & \text{if } 1 - \lambda/2 < y < b(1+\lambda), \\ 0 & \text{if } (1+\lambda)/2 < y < 1 \end{cases}$$

which have the solution (7). □

Figure 5 gives the graph of the invariant density.

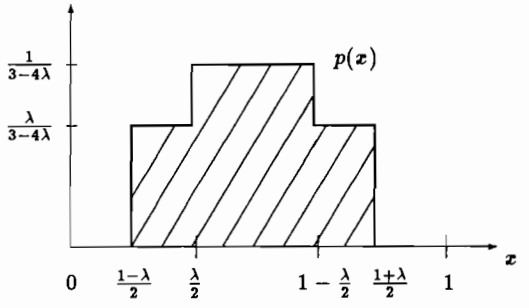


Figure 5. The graph of the invariant density (2).

3 A Bayesian interpretation

It is pleasing to give the invariant measure a Bayesian interpretation. Assume, then that x_n^* is exactly distributed according to $p(x)$. The information gained from the new observation in $[a_n, b_n]$ is of the form

$$x^* \leq \frac{a_n + b_n}{2}$$

if the right subinterval is deleted and

$$x^* \geq \frac{a_n + bn}{2}$$

if the left one is. We are assuming the information that $f(x)$ is symmetric with respect to x^* is used. This information (after renormalization) is conveyed to x_{n+1}^* in the form of the conditional distribution which is defined as follows

$$p(x_{n+1}^* \mid x_n^* \leq \frac{1}{2}) = \begin{cases} 0 & \text{if } 0 \leq x_{n+1}^* < \frac{\lambda}{2} \\ \frac{2\lambda}{3-4\lambda} & \text{if } \frac{\lambda}{2} \leq x_{n+1}^* < \frac{1}{2}, \\ \frac{2}{3-4\lambda} & \text{if } \frac{1}{2} \leq x_{n+1}^* < \frac{1+\lambda}{2}, \\ 0 & \text{if } \frac{1+\lambda}{2} \leq x_{n+1}^* \leq 1 \end{cases}$$

Similarly

$$p(x_{n+1}^* \mid x_n^* > \frac{1}{2}) = \begin{cases} 0 & \text{if } 0 \leq x_{n+1}^* < \frac{1-\lambda}{2} \\ \frac{2}{3-4\lambda} & \text{if } \frac{1-\lambda}{2} \leq x_{n+1}^* < \frac{1}{2}, \\ \frac{2\lambda}{3-4\lambda} & \text{if } \frac{1}{2} \leq x_{n+1}^* < 1 - \frac{\lambda}{2}, \\ 0 & \text{if } 1 - \frac{\lambda}{2} \leq x_{n+1}^* \leq 1. \end{cases}$$

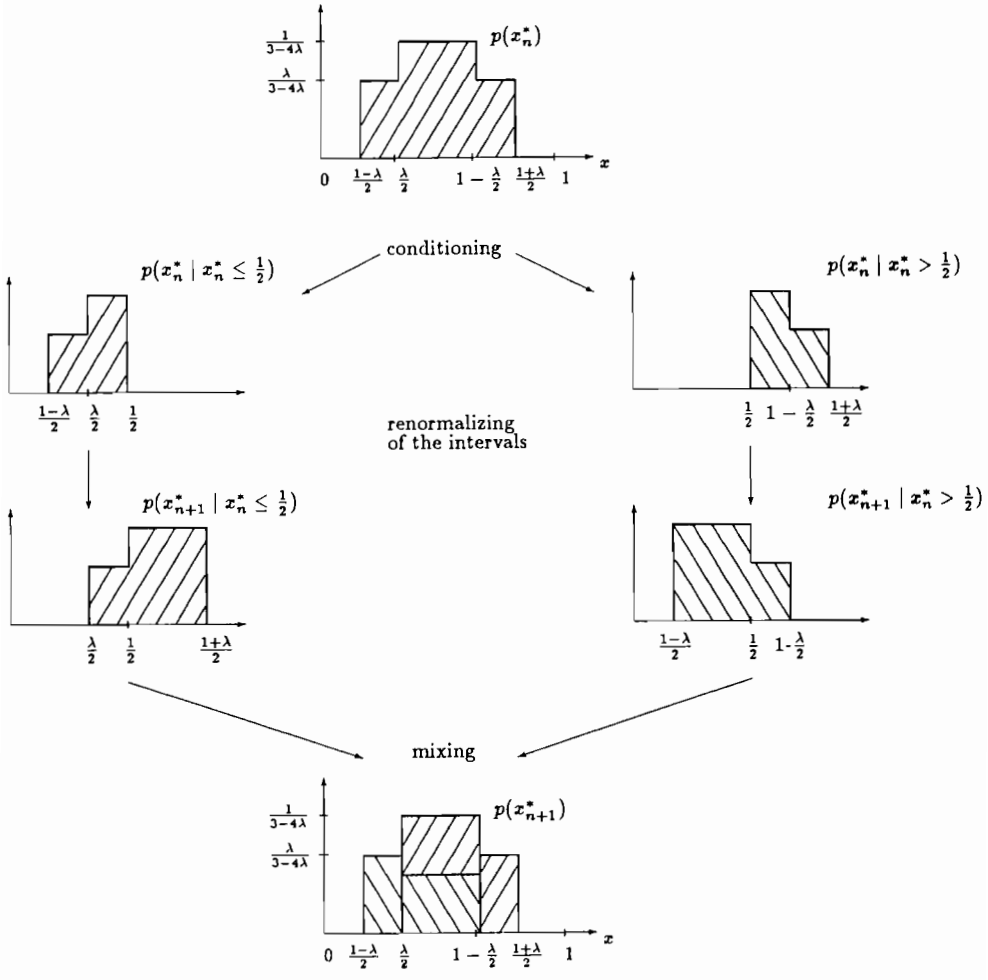


Figure 6. Bayes interpretation.

Note also that

$$Pr(x_n^* \leq \frac{1}{2}) = Pr(x_n^* > \frac{1}{2}) = \frac{1}{2}$$

Now $p(x)$ is invariant in the sense that the marginal distribution for x_{n+1}^* under this updating rule is the same as for x_n^* namely $p(x)$ itself:

$$p(x_{n+1}^*) = p(x_{n+1}^* \mid x_n^* < \frac{1}{2})Pr(x_n^* < \frac{1}{2}) + p((x_{n+1}^* \mid x_n^* > \frac{1}{2})Pr(x_n^* > \frac{1}{2})$$

Figure 6 demonstrates the Bayesian interpretation.

4 The Golden Section algorithm for nonsymmetric functions

The situation for nonsymmetric functions is more complex. In general we loose the dependence of the iteration function $h(.)$ only on the location of the minimum within the renormalized interval, namely x_n^*. One example in which this dependence is preserved is a function of the form

$$\begin{aligned} f(x) &= x^* - c(x - x^*) \quad 0 \leq x < x^* \\ &= x^* + d(x - x^*) \quad x^* \leq x \leq 1 \end{aligned}$$

where $c, d > 0$. In this case the iteration is

$$x_{n+1}^* = h(x_n^*)$$

where

$$h(u) = \begin{cases} u(1+\lambda) & \text{if } 0 < u < b \\ u(1+\lambda) - \lambda & \text{if } b < x < 1 \end{cases}$$

and

$$b = \frac{d}{c+d}.$$

The algorithm is non degenerate when

$$1 - \lambda \leq b \leq \lambda.$$

The Lyapunov exponent is again $\log(\lambda + 1)$ and the conditions for the invariant measure are

$$p(y) = \begin{cases} 0 & \text{if } 0 < y < b(1+\lambda) - \lambda, \\ \lambda p(\lambda y + 1 - \lambda) & \text{if } b(1+\lambda) - \lambda < y < b(\lambda+2) - 1, \\ \lambda(p(\lambda y) + p(\lambda y + 1 - \lambda)) & \text{if } b(\lambda+2) - 1 < y < b(\lambda+2) - \lambda, \\ \lambda p(\lambda y) & \text{if } b(\lambda+2) - \lambda < y < b(1+\lambda), \\ 0 & \text{if } b(1+\lambda) < y < 1 \end{cases}$$

which has the solution

$$p(y) = \begin{cases} 0 & \text{if } 0 < y < b(1+\lambda) - \lambda, \\ \frac{\lambda}{3-4\lambda} & \text{if } b(1+\lambda) - \lambda < y < b(\lambda+2) - 1, \\ \frac{1}{3-4\lambda} & \text{if } b(\lambda+2) - 1 < y < b(\lambda+2) - \lambda, \\ \frac{\lambda}{3-4\lambda} & \text{if } b(\lambda+2) - \lambda < y < b(1+\lambda), \\ 0 & \text{if } b(1+\lambda) < y < 1 \end{cases}$$

Figures 7 and 8 give the function h and the invariant density respectively.

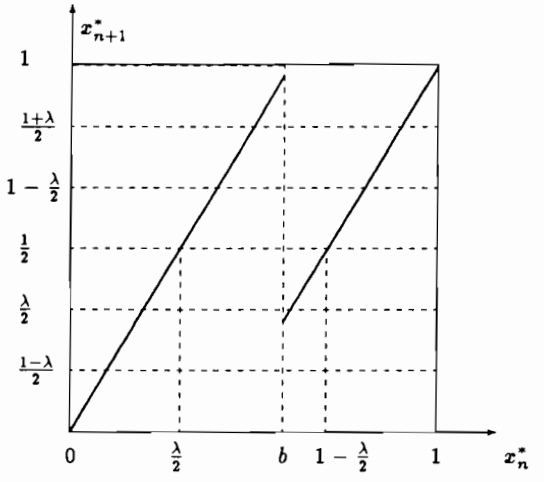


Figure 7. The Golden section iteration in the nonsymmetric case.

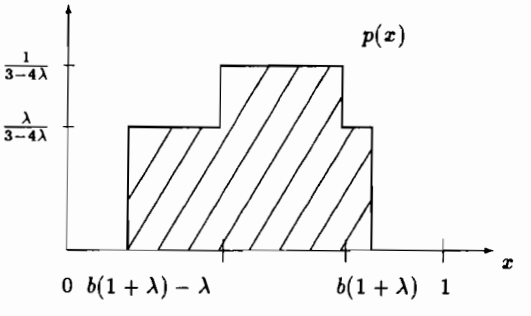


Figure 8. Invariant density for the Golden section algorithm in the nonsymmetric case.

Again we have the $\{\lambda, 1-\lambda\}$ cycle and the point b plays the same role as $\frac{1}{2}$ in the symmetric case.

Analogously to the symmetric case we exclude x^* for which

$$x^* = \sum_{i=1}^{k} \frac{a_i}{\lambda^i}$$

$$= \sum_{i=1}^{k} \frac{a_i}{\lambda^i} + \frac{b}{\lambda_i^{k+1}}$$

A Bayesian interpretation is also possible in this case.

In a sense the symmetric and nonsymmetric examples above are canonical cases. It is possible to show that if $f(x)$ is unimodal, differentiable and

$$f(x - x^*) = f(x^*) + | x - x^* |^2 f''(x^*) + O(| x - x^* |^{2+\delta})$$

that the approximate symmetry:

$$f(x - x^*) \simeq f(x^*) + | x - x^* | f'(x^*)$$

forces an invariant measure on the sequence x_n^* except regarding the h function will change also with n

$$x_{n+1}^* = h_n(x_n^*)$$

For differential nonsymmetric functions which are unimodal and possess a uniform quadratic approximation around x^* it is possible to show that the invariant measure is indentical to the measure which appears in the symmetric case. The proof, which is omitted here, follows using standard generalizations of the ergodic theorems. There is also an interesting "self-correcting" property in this case. For the nonsymmetric case when x_n^* is close to $(a_n + b_n)/2$ the sequence departs from the behaviour for the symmetric case however (i) it returns to the symmetric behaviour (ii) the departure from the symmetric case happens for an asymptotically negligible number of cases.

5 General class of algorithms

Let x_n' and x_n'' be the two observation points at iteration n in the normalized interval $[0,1]$, $x_n' < x_n''$. Let $f(x)$ be unimodal and locally symmetric and we delete $[0, x_n')$ or $(x_n'', 1]$ in the usual way. Let x_n^* be the location of the minimum in the renormalized interval. Define a general algorithm by

$$x_{n+1}^* = \begin{cases} \frac{x_n^*}{x_n''} & \text{if } x_n < \overline{x}_n \\ \frac{x_n^* - x_n'}{1 - x_n'} & \text{if } x_n > \overline{x}_n \end{cases}$$

where $\overline{x}_n = \frac{x'_n + x''_n}{2}$ and

$$z_{n+1} = \begin{cases} \frac{x'_n}{x''_n} & \text{if } x_n < \overline{x}_n \\ \frac{x''_n - x'_n}{1 - x'_n} & \text{if } x_n > \overline{x}_n \end{cases}$$

$$x'_{n+1} = \min\{e_{n+1}, z_{n+1}\},$$
$$x''_{n+1} = \max\{e_{n+1}, z_{n+1}\}$$

Here $e_{n+1}, 0 < e_{n+1} < 1$ represents the placement of a new point in the next renormalized interval.

This general class of algorithms includes the major part of interesting one-dimensional algorithms. Let us define certain more narrow classes.

1. Symmetric algorithms: $x'_n = 1 - x''_n$ for all n.
2. Fixed algorithms: $x'_n = x'$ and $x''_n = x''$ for all n.
3. Semi-fixed algorithms: $e_n = e$ for all n.
4. Fixed width algorithms: $w_n = x''_n - x'_n$ is a constant for all n

There is a rich theory for these classes of algorithms which the authors will present in a longer work (Wynn and Zhigljavsky (1992)). One of the most intriguing developments is that there are examples of algorithms in the classes (3) and (4) which have a faster asymptotic rate of convergence than the Golden Section algorithm despite the fact that the convergence rate for the latter is constant at λ. (In particular, the average asymptotic rate for (3) above with $e_n = 0.5$ and (4) above with $w_n = 0.125$ are approximately 0.6023 and 0.5713 respectively compared with the larger value of 0.6180... for the Golden Section algorithm.) The seeming contradiction is resolved because the Golden Section algorithm concentrates essentially on the worst case which form only a countable number of cases (as explained). These other algorithms which beat the Golden Section pay attention to the set of x^* which generate the invariant measure and which has the Lebesgue measure 1 over [0,1]. The authors believe that this qualitative conclusion may have important implications for the algorithmic theory in continuous spaces.

References

Kiefer J.(1953) Sequential minimax search for a maximum. *Proc. Amer. Mathem. Soc.*, vol. 4, N 3, 502-506.

Kiefer J.(1957) Optimum sequential search and approximation methods under minimum regularity assumptions. *J.Soc.Indust.Appl.Math.*, vol. 5, N 3, 105-136.

Wynn H.P. and Zhigljavsky A.A. (1992) Chaotic behaviour of search algorithms (Submitted to *Acta Applicandae Mathematicae*).

On a Class of Stochastic Optimization Algorithms with Applications to Manufacturing Models

G. Yin , H.M. Yan and S.X.C. Lou

A class of stochastic optimization algorithms is developed in this work. The algorithms have recursive form, and use averaging in the updating. By virtue of the weak convergence methods, it is shown that a sequence of continuous time interpolation converges to a process which satisfies an ordinary differential equation. Order of magnitude estimates on the error is derived and a suitably scaled sequence is then shown to converge to a solution of stochastic differential equation. The scaling together with the asymptotic covariance can be used as a measure of rate of convergence. Applications to manufacturing models are also considered.

1 Introduction

Let X, $\xi \in \mathbb{R}^r$, $b(\cdot)$ be an $\mathbb{R}^r$-valued function, and $\varepsilon > 0$, $T_\varepsilon > 0$. Consider the following stochastic optimization algorithm:

$$X_{n+1} = X_n + \varepsilon \frac{1}{T_\varepsilon} \int_{nT_\varepsilon}^{nT_\varepsilon + T_\varepsilon} b(X_n, \xi(t))dt \tag{1.1}$$

or

$$X_{n+1} = X_n + \varepsilon \frac{1}{T_\varepsilon} \sum_{j=nT_\varepsilon}^{nT_\varepsilon + T_\varepsilon - 1} b(X_n, \xi_j), \tag{1.2}$$

where ε is known as a constant step size or a gain parameter and T_ε is chosen in such a way that $T_\varepsilon \to \infty$ as $\varepsilon \to 0$. In (1.2), T_ε is understood to be an integer. Our goal is to develop various asymptotic properties of the above algorithms.

Algorithms of the form (1.1) and (1.2) arise from various stochastic optimization problems, in which one wishes to minimize a functional $J(\cdot)$. In order to carry out the indicated task, gradient estimates of $J(\cdot)$ are needed, and simulations are conducted. Suppose that $\xi(\cdot)$ is a strictly stationary process satisfying certain regularity conditions. For each x, the gradient estimate of $J(\cdot)$ is of the form

$$\bar{\nabla} J_T(x) = -\frac{1}{T} \int_0^T b(x, \xi(t))dt$$

if the simulation can be done at continuous time, or

$$\bar{\nabla} J_T(x) = -\frac{1}{T} \sum_{j=0}^{T-1} b(x, \xi_j)$$

if the simulation is conducted at discrete time. To approximate the optimal vector x^*, i.e., $J(x^*) = \min_x J(x)$, a stochastic approximation type of algorithm is employed. This algorithm takes the form (1.1) or (1.2) in accordance with the ways that the gradient estimates are obtained.

Here we are mainly interested in the asymptotic properties of the above stochastic procedures and applications to data analysis of stochastic manufacturing models. (1.1) and/or (1.2) are not standard stochastic approximation algorithms owing to the fact averaging is used in the scheme. Sampling controlled stochastic approximation algorithms with averaging of observations was also considered in 4. The entire past was incorporated in the averaging in 4, whereas the averages are taken when x is fixed in our approach. Moreover, another class of averaging algorithms were also proposed by 12 and developed further in 14. In these papers, the main concerns are the asymptotic optimality and related matters. In contrast to the aforementioned references, continuous time random processes are also dealt with here. It seems that the algorithms considered here can also be used in conjunction with parallel processing methods (cf. 13 and the references therein.)

The remainder of the paper is arranged as follows. Convergence of the algorithms is considered in the next section. Section 3 presents applications of using the algorithms solving problems in manufacturing systems. Section 4 deals with bounds of the estimation errors. A local limit theorem is obtained via a suitably scaled sequence.

To proceed, a word about the notations is in order. K will be used to denote a generic positive constant; x' will denote the transpose of x; $f_x(\cdot)$ will stand for the x-derivative of $f(\cdot)$.

2 Convergence

This section is devoted to investigating the convergence of the proposed algorithms. We shall mainly work with the algorithm (1.1). (1.2) can be handled similarly. To proceed, the following assumptions are needed.

(A1) $\xi(\cdot)$ is stationary. There is a continuous function $\bar{b}(x)$ such that for each x

$$\frac{1}{T}\int_0^T b(x,\xi(t))dt \xrightarrow{T} \bar{b}(x) \text{ in probability.}$$

(A2) For each $T < \infty$, $t \in [0,T]$,

$$\lim_{\delta \to 0} E \sup_{|x-y|<\delta} |b(x,\xi(t)) - b(y,\xi(t))| = 0.$$

(A3) For each $N < \infty$, the set

$$\{\sup_{|x|\le N} |b(x,\xi(t))|\} \text{ is uniformly integrable.}$$

Remark: These assumptions are motivated by the particular applications of manufacturing models (cf. Section 3). (A1) is an ergodic condition in the sense of convergence in probability. It is a basic averaging condition. If $\xi(\cdot)$ is a ϕ-mixing process with $E|\xi(t)| < \infty$ then it is a strongly ergodic process and hence (A1) holds. In fact, in this case, the convergence is in the sense of with probability one (w.p.1.)

(A2) indicates that the function $b(\cdot, \xi)$ may not be continuous, but its expectation is continuous such as the case that $b(\cdot, \xi)$ is an indicator function or combination of indicator functions.

In various applications, the function $b(x, \xi)$ is often bounded. In such a case, (A3) is verified. Nevertheless, (A3) can deal with more complex situation, for example, if

$$|b(x, \xi)| \leq h_0(x)g_1(\xi) + g_2(\xi)$$

where $h_0(x)$ is a continuous function, $E|g_i(\xi)|^{1+\alpha} < \infty$ for some $\alpha > 0$, the condition (A3) is also satisfied.

For Algorithm (1.2) (the case of discrete time averaging), the averaging in (A1) will be replaced by:

$$\frac{1}{T}\sum_{j=0}^{T-1} b(x, \xi_j) \to \bar{b}(x) \text{ in probability as } T \to \infty. \tag{2.1}$$

To proceed, we work with continuous time interpolated processes. Let $x^\epsilon(\cdot)$ be defined by $x^\epsilon(t) = X_n$ for $t \in [n\epsilon, n\epsilon + \epsilon)$. Under the framework of weak convergence (cf. 7), it will be shown that the following limit theorem holds.

Theorem 2.1. *Suppose that (A1)-(A3) are satisfied and the differential equation*

$$\dot{x} = \bar{b}(x) \tag{2.2}$$

has a unique solution for each initial condition. If $x^\epsilon(0) \Rightarrow x(0)$ then $\{x^\epsilon(t)\}$ is tight in $D^r[0, \infty)$. Every convergent subsequence has limit $x(\cdot)$ which satisfies the differential equation (2.2).

Remark: $D^r[0, \infty)$ denotes the space of $\mathbf{R}^r$-valued functions which are right continuous and have left-hand limits, endowed with the Skorohod topology. For various notations and terms in weak convergence theory such as Skorohod topology, Skorohod representation etc. and many others, see 5, and the references therein.

Proof: To avoid the problem with possible unboundedness, a truncation device will be used (cf. 7 Chapter 3). For each $N < \infty$, let $x^{\epsilon,N}(\cdot)$ be the N-truncation of $x^\epsilon(\cdot)$ such that $x^{\epsilon,N}(t) = x^\epsilon(t)$ up until the first exit from the N-sphere $S_N = \{x; |x| \leq N\}$.

Owing to the definition of the interpolation, (without loss of generality, assume that t/ϵ and $(t+s)/\epsilon$ are integers), and choosing a sequence of integers $\{m_\epsilon\}$ such that $m_\epsilon \to \infty$ as

$\varepsilon \to 0$ and $\varepsilon m_\varepsilon = \Delta_\varepsilon \to 0$, we have

$$\begin{aligned} x^{\varepsilon,N}(t) &= x^{\varepsilon,N}(0) + \varepsilon \sum_{j=0}^{t/\varepsilon-1} \frac{1}{T_\varepsilon}\int_{jT_\varepsilon}^{jT_\varepsilon+T_\varepsilon} b(x_j^{\varepsilon,N}, \xi(u))du \\ &= x^{\varepsilon,N}(0) + \sum_{0\le l\Delta_\varepsilon \le t} \Delta_\varepsilon \frac{1}{m_\varepsilon} \sum_{lm_\varepsilon \le j \le lm_\varepsilon+m_\varepsilon-1} \frac{1}{T_\varepsilon}\int_{jT_\varepsilon}^{jT_\varepsilon+T_\varepsilon} b(x_j^{\varepsilon,N}, \xi(u))du \qquad (2.3) \\ &= x^{\varepsilon,N}(0) + \int_0^t B^\varepsilon(\tau)d\tau \end{aligned}$$

where

$$B^\varepsilon(t) = \frac{1}{m_\varepsilon} \sum_{lm_\varepsilon \le j \le lm_\varepsilon+m_\varepsilon-1} \frac{1}{T_\varepsilon}\int_{jT_\varepsilon}^{jT_\varepsilon+T_\varepsilon} b(x_j^{\varepsilon,N}, \xi(u))du \text{ on } t \in [l\Delta_\varepsilon, l\Delta_\varepsilon + \Delta_\varepsilon). \qquad (2.4)$$

It follows from (2.3) that $\dot{x}^{\varepsilon,N}(t) = B^\varepsilon(t)$.

Due to the fact that $x^{\varepsilon,N}(0) \Rightarrow x^N(0)$, $\{x^{\varepsilon,N}(0)\}$ is tight. By virtue of the Markov inequality, for any $\eta > 0$, $\gamma > 0$, there exists a $\delta = \gamma\eta$ such that for some $\tilde{t} \in [t, t+s]$,

$$\begin{aligned} & P\left(\sup_{|s|\le\delta} |x^{\varepsilon,N}(t+s) - x^{\varepsilon,N}(t)| \ge \gamma\right) \\ \le\ & \frac{1}{\gamma} E \sup_{|s|\le\delta} |x^{\varepsilon,N}(t+s) - x^{\varepsilon,N}(t)| \\ =\ & \frac{1}{\gamma} E \sup_{|s|\le\delta} (|B^\varepsilon(\tilde{t})|\,|s|) \\ \le\ & \frac{1}{\gamma} K\delta \le K\eta \end{aligned}$$

for some $K > 0$. Then by virtue of 3 Theorem 8.2, $\{x^{\varepsilon,N}(\cdot)\}$ is tight and the limit of any convergent subsequence has continuous paths with probability one (w.p.1.)

Pick out an arbitrary convergent subsequence and denote the limit by $x^N(\cdot)$. By Skorokhod imbedding and without changing notations, we assume that $x^{\varepsilon,N}(\cdot) \to x^N(\cdot)$ w.p.1 and the convergence is uniform on any finite time interval.

Define

$$M^N(t) = x^N(t) - x^N(0) - \int_0^t \bar{b}(x^N(u))du. \qquad (2.5)$$

If we can show that $M^N(t)$ is a continuous martingale, the limit theorem will hold for the truncated process. Since $M^N(0) = 0$ and $M^N(t)$ is Lipschitz continuous if it is a martingale it must be $M^N(t) \equiv 0$. Therefore, only the martingale property needs to be verified.

To this end, let $g(\cdot)$ be any bounded and continuous function, ν be any positive integer, $t_i < t < t+s$ for $i \le \nu$. In view of the weak convergence and Skorokhod imbedding,

$$\begin{aligned} & \lim_\varepsilon Eg(x^{\varepsilon,N}(t_i), i \le \nu)\left(x^{\varepsilon,N}(t+s) - x^{\varepsilon,N}(t)\right) \\ &= Eg(x^N(t_i), i \le \nu)\left(x^N(t+s) - x^N(t)\right). \end{aligned} \qquad (2.6)$$

On the other hand,

$$\begin{aligned}
&\lim_{\epsilon} Eg(x^{\epsilon,N}(t_i), i \le \nu)\left(x^{\epsilon,N}(t+s) - x^{\epsilon,N}(t)\right) \\
&= \lim_{\epsilon} Eg(x^{\epsilon,N}(t_i), i \le \nu)[\sum_{l\Delta_\epsilon = t}^{t+s} \Delta_\epsilon \frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x_j^{\epsilon,N}, \xi(u))du] \\
&= \lim_{\epsilon} Eg(x^{\epsilon,N}(t_i), i \le \nu)[\sum_{l\Delta_\epsilon = t}^{t+s} \Delta_\epsilon \frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x^{\epsilon,N}(\tau), \xi(u))du].
\end{aligned} \tag{2.7}$$

The last equality above follows from the weak convergence, the Skorokhod imbedding, assumption (A2) and $\epsilon j \to \tau$ for $lm_\epsilon \le j \le lm_\epsilon + m_\epsilon$ as $\epsilon \to 0$.

Using a basic result of analysis, for any $\eta > 0$, there exists a function $x^\eta(\cdot)$ that takes only finitely many values (say $x_1, \cdots, x_q$), such that

$$|x(\tau) - x^\eta(\tau)| < \eta.$$

Consequently, by applying (A2), the limit in (2.7) is the same as that of

$$\lim_{\epsilon} Eg(x^{\epsilon,N}(t_i), i \le \nu)\left(\sum_{l\Delta_\epsilon = t}^{t+s} \Delta_\epsilon \frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x^\eta(\tau), \xi(u))du\right).$$

Since

$$\begin{aligned}
&\frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x^\eta(\tau), \xi(u))du \\
&= \sum_{i=1}^{q} \frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x_i, \xi(u))du\, I_{\{x^\eta(\tau) = x_i\}} \\
&\to \sum_{i=1}^{q} \bar{b}(x_i) I_{\{x^\eta(\tau) = x_i\}} \text{ in probability} \\
&= \bar{b}(x^\eta(\tau)),
\end{aligned}$$

and $\eta > 0$ is arbitrary, we have

$$\frac{1}{m_\epsilon} \sum_{lm_\epsilon \le j \le lm_\epsilon + m_\epsilon - 1} \frac{1}{T_\epsilon}\int_{jT_\epsilon}^{jT_\epsilon + T_\epsilon} b(x^\eta(\tau), \xi(u))du \xrightarrow{\epsilon} \bar{b}(x^N(\tau)) \text{ in probability.}$$

Substituting this into (2.7),

$$\begin{aligned}
&\lim_{\epsilon} Eg(x^{\epsilon,N}(t_i), i \le \nu)\left(x^{\epsilon,N}(t+s) - x^{\epsilon,N}(t)\right) \\
&= Eg(x^N(t_i), i \le \nu)\left(x^N(t+s) - x^N(t) - \int_t^{t+s} \bar{b}(\tau)d\tau\right).
\end{aligned} \tag{2.8}$$

Combining (2.6) and (2.8), we arrive at

$$Eg(x^N(t_i), i \le \nu)\left(x^N(t+s) - x^N(t) - \int_t^{t+s} \bar{b}(\tau)d\tau\right) = 0.$$

Hence $M^N(t)$ is a martingale.

Finally, use the idea of 7, Theorem 2.2 (and the corollary) to finish the proof. Let $P_{x(0)}(\cdot)$ (the subscript $x(0)$ signifies the dependence on the initial data) and $P^N(\cdot)$ be the measures induced by $x(\cdot)$ and $x^N(\cdot)$, respectively, on $\mathcal{B}$, where $\mathcal{B}$ denote the σ-algebra of Borel subsets of $D^r[0,\infty)$. $P_{x(0)}(\cdot)$ is unique since there is a unique solution to the ordinary differential equation for the initial condition value $x(0)$. Thus, for each $T < \infty$,

$$P_{x(0)}(x(\cdot) \in A) = P^N(x^N(\cdot) \in A)$$

for each $A \in \mathcal{B}$ such that $x(t)$ takes values in S_N (the N-sphere). As a result,

$$\lim_{N\to\infty} P_{x(0)}\left(\sup_{t\le T}|x(t)| \le N\right) = 1.$$

This together with the weak convergence of $x^{\epsilon,N}(\cdot)$ implies that $x^\epsilon(\cdot) \Rightarrow x(\cdot)$. Since the limit is unique, it does not depend on the chosen subsequence. The proof of the theorem is completed. □

Consider Algorithm (1.2). Suppose that the conditions of Theorem 2.1 are satisfied with (A1) replaced by (2.1). Define $x^\epsilon(\cdot)$ as in the previous theorem. Then the result of Theorem 2.1 still holds. The proof is similar to that of Theorem 2.1.

Theorem 2.1 is similar to the law of large numbers. It gives information on the location and/or distribution of $x^\epsilon(\cdot)$ for small ϵ and for large but bounded t. It can be seen that there is a natural connection between the recursive procedure and the corresponding ordinary differential equation. The optimal threshold value we are seeking for in fact, is a stable point of the differential equation (2.2).

Theorem 2.2. *Assume that the conditions of Theorem 2.1 hold, and (i) the ODE (2.2) has a unique asymptotically stable point x^* (in the sense of Liapunov stability). (ii) The set*

$$\{X_n;\ n < \infty, \epsilon > 0\} \tag{2.9}$$

is bounded in probability, i.e., for each $\eta > 0$, there is a $\kappa_\eta > 0$ such that for all $\epsilon > 0$, and all n, $P(|X_n| \ge \kappa_\eta) \le \eta$. Let $t_\epsilon \to \infty$ as $\epsilon \to 0$. Then $x^\epsilon(t_\epsilon + \cdot)$ is tight in $D^r[0,\infty)$ and any weak limit is equal to x^*. □

(2.9) can be established by using a perturbed Liapunov function methods (cf. 9 and 11). The proof of this Theorem can be obtained analogously as in 9 Theorem 5.1.

3 Applications to manufacturing models

Applications of the stochastic optimization algorithms to manufacturing models are dealt with in this section. Data analysis and numerical methods for manufacturing models under threshold controls policies are considered.

In various applications, threshold controls are widely utilized since the idea is very appealing and the principle is easy to apply. Once a threshold value is determined, a controller or an

operator can ignore detailed variations and concentrate only on adjusting controls according to the threshold criteria.

Apparently, first and foremost important task is to locate the optimal threshold values. For one machine models, an explicit solution was found by means of stochastic optimal control techniques in 1 for a discounted cost function, and in 2 for an average cost per unit time problem.

It should be pointed out that the solutions (the explicit form of the threshold expression) in 1 and 2 have complicated forms. In addition, for multi-machine models, the problems become very hard to handle. Although some attempt has been made and optimal solutions were shown to be of threshold type 10, no 'closed' form solution has been found up to date. It is thus sensible to look at possible alternatives–numerical solutions.

Multi-machine manufacturing models will be considered. In lieu of solving the dynamic programing equation as in 2, stochastic optimization methods are applied to the problem and recursive algorithms are developed.

$x_i(t)$ will denote the inventory levels of machine i, and $u_i(t)$ stand for production rate (the control) of machine i. Since we are not solving the dynamic programing equations, the demand processes can be quite general. They do not have to be constants although a constant demand model is used for simplicity. In the sequel, formulation for surplus (defined as the difference between accumulative production and accumulated demand) control model is given. Then approximation procedures will be developed and some numerical results will be presented.

3.1 Surplus control model

The two machines are in a cascade form and given by

$$\begin{aligned} \dot{x}_1(t) &= u_1(t) - u_2(t), \\ \dot{x}_2(t) &= u_2(t) - d, \\ x_1(t) &\geq 0. \end{aligned} \tag{3.1}$$

Let the machine operation states be defined by $I_i(t)$, $i = 1, 2$ with

$$I_i(t) = \begin{cases} 1, & \text{machine } i \text{ is working,} \\ 0, & \text{otherwise.} \end{cases}$$

We then have

$$0 \leq u_i(t) \leq u_{i\max}, \ i = 1, 2.$$

where $u_{i\max} > d$. Assume that $u_{1\max} > u_{2\max}$. This scenario is depicted in the following Figure 1.

Surplus at machine i is defined as the difference between accumulative production and accumulated demand, i.e., it is the inventory level (or work in progress) at machine i plus the inventory level of all down stream machines. Let $s_i(t)$ be the surplus for machine i, $i = 1, 2$.

$$s_1(t) = x_1(t) + x_2(t) \text{ and } s_2(t) = x_2(t).$$

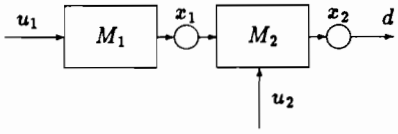


Figure 1: A two-machine system

Notice that the surplus can be positive or negative. If a surplus is negative, a backlog for the production at i is accumulated. With these notations, the system dynamics can also be written as:

$$\begin{aligned} \dot{s}_1(t) &= u_1(t) - d, \\ \dot{s}_2(t) &= u_2(t) - d, \\ s_1(t) &\geq s_2(t), \end{aligned} \tag{3.2}$$

Comparing (3.2) with the one-machine model in the work 1 and 2, the surplus control is more or less like having two machines operating independently.

Let z_i denote the surplus threshold levels of machine i. The control policy is given by:

$$\begin{cases} u_1(t) = \begin{cases} u_{1\max}, & \text{if } s_1(t) < z_1,\ I_1(t) = 1, \\ d, & \text{if } s_1(t) = z_1,\ I_1(t) = 1, \\ 0, & \text{otherwise;} \end{cases} \\ u_2(t) = \begin{cases} u_{2\max}, & \text{if } s_2(t) < z_2,\ s_1(t) - s_2(t) > 0 \text{ and } I_2(t) = 1, \\ d, & \text{if } s_2(t) = z_2,\ s_1(t) - s_2(t) > 0 \text{ and } I_2(t) = 1, \\ 0, & \text{otherwise.} \end{cases} \end{cases} \tag{3.3}$$

The interpretation of the control policies is similar to that of the one machine case. The problem to be investigated is to find the optimal threshold value $z^* = (z_1^*, z_2^*)$ such that the cost functional

$$J(z) = \lim_T \frac{1}{T} E \int_0^T (c_1 x_1(t) + c_2^+ x_2^+(t) + c_2^- x_2^-(t))dt \tag{3.4}$$

is minimized.

3.2 Approximation procedures

The framework and procedures will be presented for the surplus control problems below.

Let $\xi(t) = (x(t), I(t))'$, where $x(t) = (x_1(t), x_2(t))'$ and $I(t) = (I_1(t), I_2(t))'$. As in 2, we shall assume that there is an invariant measure $P^z(\cdot)$ for the process $\xi(t)$ throughout the paper. With this assumption, the cost function can be rewritten as

$$J(z) = \int [c_1 x_1 + c_2^+ x_2^+ + c_2^- x_2^-] P^z(dx). \tag{3.5}$$

By using perturbation analysis methods 6, the sample gradient estimates can be constructed. Let

$$\tilde{\nabla} J_T(z) = (\tilde{\nabla} J_T^1(z), \tilde{\nabla} J_T^2(z))'$$

be the gradient estimate of $\nabla J(z)$.

$$\begin{aligned}\tilde{\nabla} J_T^1(z) &= \frac{1}{T}\left(\int_0^T [c_1 + (c_2^+ - c_1) I_{\{x_2(t)\geq 0, p_1(t)=1\}} - (c_2^- - c_1) I_{\{x_2(t)\leq 0, p_1(t)=1\}}]dt\right) \\ \tilde{\nabla} J_T^2(z) &= \frac{1}{T}\left(\int_0^T [(c_2^+ - c_1) I_{\{x_2(t)\geq 0, p_2(t)=1\}} - (c_2^- - c_1) I_{\{x_2(t)\leq 0, p_2(t)=1\}}]dt\right),\end{aligned} \tag{3.6}$$

where $p_1(t)$ and $p_2(t)$ are auxiliary processes. $p_1(t) = 1$ during perturbation propagation on Machine 2 due to the perturbation on the parameter z_1; $p_1(t)$ is set to zero, otherwise. $p_2(t) = 1$ during perturbation generation on Machine 2 due to perturbation on control parameter z_2; $p_2(t)$ is equal to zero, otherwise. For further use, denote the quantities in the integrands by $h(z, \xi(t))$. Then,

$$\tilde{\nabla} J_T(z) = \frac{1}{T}\int_0^T h(z, \xi(t))dt.$$

Theorem 3.1. *Assuming the existence of $P^z(\cdot)$, suppose that the process $\xi(\cdot)$ is weakly ergodic in the sense for each z and each bounded and measurable function $\pi(z, \xi)$,*

$$\frac{1}{T}\int_0^T \pi(z, \xi(t))dt \overset{T\to\infty}{\longrightarrow} \bar{\pi}(z) \text{ in probability,}$$

where $\bar{\pi}(z)$ denotes the average of $\pi(z, \cdot)$ with respect to the invariant measure $P^z(\cdot)$. Then, the gradient estimate is consistent in that $\tilde{\nabla} J_T(z) \overset{T}{\longrightarrow} \nabla J(z)$ in probability as $T \to \infty$.

Utilizing the above gradient estimates, a recursive algorithm is then developed to approximate the optimal threshold values. The algorithm is of the form

$$Z_{n+1} = Z_n - \varepsilon \text{ (gradient estimate)} .$$

The essence is that the approximating sequence $\{Z_n\}$ is generated recursively. For each n, with threshold value Z_n, a time interval $[0, T_\varepsilon]$ is taken. Following the path of the process involved, a simulation run is performed to get a gradient estimate $\tilde{\nabla} J(Z_n)$. The iteration is given by (1.1) with X_n replaced by Z_n.

Assume that the conditions of Theorem 3.1 are satisfied, and the differential equation

$$\dot{z} = \bar{h}(z) = -\nabla J(z) \tag{3.7}$$

has a unique solution for each initial condition $z(0)$. Let $z^\varepsilon(\cdot)$ be defined by

$$z^\varepsilon(t) = Z_n, \text{ for } t \in [n\varepsilon, n\varepsilon + \varepsilon),$$

i.e., $z^\varepsilon(\cdot)$ is a piecewise constant interpolation of z_n with interpolation interval ε. Suppose that $z_0^\varepsilon \Rightarrow z(0)$. Then $\{z^\varepsilon(\cdot)\}$ is tight in $D^2[0, \infty)$ and any weakly convergent subsequence has limit $z(\cdot)$ which is a solution of the ODE with initial condition $z(0)$.

In addition to the above conditions, assume that the conditions of Theorem 2.2 are satisfied for $\{Z_n\}$. Let $t_\varepsilon \to \infty$ as $\varepsilon \to 0$. Then $z^\varepsilon(t_\varepsilon + \cdot)$ is tight in $D^2[0, \infty)$ and any weak limit is equal to the optimal threshold value z^*.

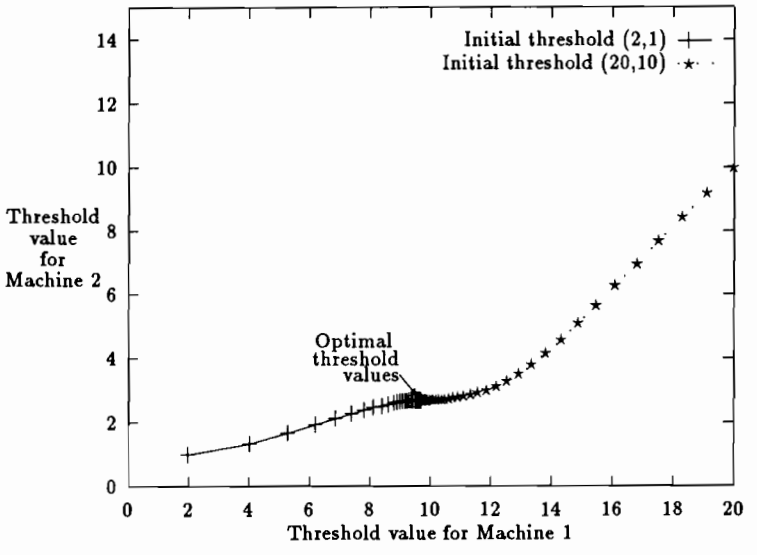


Figure 2: Convergence of the iterates

3.3 Numerical experiments

First, consider the one machine problem 2. The analytical result is given by:

$$z^* = \frac{1}{\frac{\mu}{d} - \frac{\lambda}{u_{2\max} - d}} \ln \frac{u_{\max}\lambda(c^+ + c^-)}{c^+(u_{\max} - d)(\lambda + \mu)},$$
$$J(z^*) = \frac{c^+ d}{\lambda + \mu} + \frac{c^+}{\frac{\mu}{d} - \frac{\lambda}{u_{\max} - d}} \ln \frac{u_{\max}\lambda(c^+ + c^-)}{c^+(u_{\max} - d)(\lambda + \mu)}.$$

Choose $1/\lambda = 0.1$, $1/\mu = 0.125$, $d = 1.0$, $c^+ = 2.0$, $c^- = 10.0$ and $u_{\max} = 2.0$. Then the optimal values are found to be $z^* = 66.96$ and $J(z^*) = 142.89$ by using the above formulae. Now, using our algorithm with $\varepsilon = 0.5$ and $T_\varepsilon = 10,000,100$ replications were obtained. With initial value $Z_0 = 100$, by taking averages of the replications, the approximated values are found to be $\bar{z}^* = 70.04$ (with a 95% confidence interval $[69.42, 70.06]$), and $J(\bar{z}^*) = 145.35$. Similarly, with $Z_0 = 20$, $\bar{z}_2^* = 67.23$ (with a 95% confidence interval $[66.64, 67.80]$), and $J(\bar{z}^*) = 139.34$ were obtained.

Next, Figure 2 demonstrates the convergence of the algorithm for two machine case. By generating contour curves via simulation for each set of threshold values, the approximation obtained in our algorithm is seen to belong to the region of optimality. It seems that the initial condition does not affect the algorithm significantly. Thus the algorithm is robust with respect to the initial data.

4 Further asymptotic results

In this section, further results of Algorithm (1.1) will be obtained. First, order of magnitude error bound is derived and then a local limit theorem is established under additional conditions. The consideration in this section falls into the category of rates of convergence.

4.1 Error estimates

Theorem 4.1. *Assume that the conditions of Theorem 2.1 are satisfied and there is a twice continuously differentiable Liapunov function $V(\cdot)$ such that $V(x) \geq 0$, $V(x) \to \infty$ as $|x| \to \infty$, $V_x'(x)\bar{b}(x) \leq -\lambda V(x)$ for some $\lambda > 0$ and $V_{xx}(\cdot)$ is bounded, where V_x and V_{xx} denote the first and the second derivatives of $V(\cdot)$, respectively. Suppose that for each x,*

$$\left| \sum_{j=n}^{\infty} E^{\mathcal{F}_n} \frac{1}{T_\epsilon} \int_{jT_\epsilon}^{jT_\epsilon+T_\epsilon} (b(x,\xi(t)) - \bar{b}(x)) \right| \leq K \tag{4.1}$$

for some $K > 0$, where $E^{\mathcal{F}_n}$ denotes the conditional expectation with respect to the σ-algebra $\mathcal{F}_n = \sigma\{\xi(u), u \leq nT_\epsilon + T_\epsilon\}$. Assume that

$$|b(x,\xi)|^2 + |\bar{b}(x)|^2 \leq K(1 + V(x)).$$

Then

$$\limsup_n V(X_n) = O(\varepsilon). \tag{4.2}$$

Notice that (4.1) can be written as

$$\left| \int_{nT_\epsilon}^{\infty} E^{\mathcal{F}_n} (b(x,\xi(t)) - \bar{b}(x))\, dt \right| \leq K. \tag{4.3}$$

It is readily seen that if $\xi(\cdot)$ is a ϕ-mixing process with mixing rate $\psi(\cdot)$ such that $\int_0^\infty \psi(t) < \infty$. The mixing inequality (cf. 7 pp. 82) implies that

$$\begin{aligned} & \left| \int_{nT_\epsilon}^{\infty} E^{\mathcal{F}_n} (b(x,\xi(t)) - \bar{b}(t))\, dt \right| \\ \leq\ & 2 \int_{nT_\epsilon}^{\infty} \psi(t - T_\epsilon) dt \leq K. \end{aligned}$$

Notice that the condition is slightly weaker than that of 9. Due to the fact that averaging is used in the iterates, differentiability of $b(\cdot,\cdot)$ need not be assumed.

Proof: We shall use a technique known as perturbed Liapunov function method (8, 11). By virtue of a Taylor expansion, direct calculation leads to

$$\begin{aligned} E^{\mathcal{F}_n} V(X_{n+1}) - V(X_n) &= \varepsilon V_x'(X_n)\bar{b}(X_n) \\ &+ \varepsilon V_x'(X_n) \frac{1}{T_\epsilon} \int_{nT_\epsilon}^{nT_\epsilon+T_\epsilon} (b(X_n,\xi(t)) - \bar{b}(X_n))\, dt + O(\varepsilon^2)(1 + V(X_n)). \end{aligned} \tag{4.4}$$

Define

$$V_1^\epsilon(n) = \epsilon V_x'(X_n) \sum_{j=n}^{\infty} \frac{1}{T_\epsilon} \int_{jT_\epsilon}^{jT_\epsilon+T_\epsilon} (b(X_j, \xi(t)) - \bar{b}(X_j))\, dt \tag{4.5}$$
$$V^\epsilon(n) = V(X_n) + V_1^\epsilon(n).$$

It is easily seen that

$$|V^\epsilon(n)| \le \epsilon K(1 + V(X_n)) \tag{4.6}$$

In addition,

$$\begin{aligned} & E^{\mathcal{F}_n} V_1^\epsilon(n+1) - V_1^\epsilon(n) \\ = \; & -\epsilon V_x'(X_n) \frac{1}{T_\epsilon} \int_{nT_\epsilon}^{nT_\epsilon+T_\epsilon} (b(X_n, \xi(t)) - \bar{b}(X_n))\, dt \\ & + O(\epsilon^2)(1 + V(X_n)). \end{aligned}$$

Consequently, owing to the assumption of this theorem,

$$E^{\mathcal{F}_n} V^\epsilon(n+1) - V^\epsilon(n) \le -\lambda\epsilon V(X_n) + O(\epsilon^2)(1 + V(X_n)).$$

(4.6) then yields that

$$E^{\mathcal{F}_n} V^\epsilon(n+1) \le V^\epsilon(n) - \epsilon\lambda V^\epsilon(n) + O(\epsilon^2)(1 + V^\epsilon(n)).$$

By choosing ϵ small enough, we get

$$E^{\mathcal{F}_n} V^\epsilon(n+1) \le \left(1 - \frac{\epsilon\lambda}{2}\right) V^\epsilon(n) + O(\epsilon^2).$$

Iterating on the above inequality, taking expectation, and letting $n \to \infty$, the desired result follows. □

4.2 A local limit theorem

If the Liapunov function is locally (near x^*) quadratic, then it can be shown that there is an N_ϵ such that $\{U_n = (X_n - x^*)/\sqrt{\epsilon};\ n \ge N_\epsilon\}$ is tight. Define $u^\epsilon(t) = U_n$ for $t \in [(n - N_\epsilon)\epsilon, (n - N_\epsilon + 1))$. In order to obtain a local limit result for the scaled sequence $\{u^\epsilon(\cdot)\}$, in addition to the conditions of Section 2, assume that $b_x(x^*, \xi)$ exists and

$$\begin{aligned} & b(x, \xi) = b(x^*, \xi) + b_x(x^*, \xi)(x - x^*) + o(|x - x^*|); \\ & \frac{1}{T} \int_0^T b_x(x, \xi(t)) \to \bar{b}_x(x) \text{ in probability as } T \to \infty \text{ for each } x; \\ & W_n = \sqrt{\epsilon} \sum_{j=N_\epsilon}^{n} \frac{1}{T_\epsilon} \int_{jT_\epsilon}^{jT_\epsilon+T_\epsilon} b(x^*, \xi(t))dt, \\ & w^\epsilon(t) = W_n \text{ for } t \in [\epsilon(n - N_\epsilon), \epsilon(n - N_\epsilon + 1)); \\ & w^\epsilon(\cdot) \Rightarrow w(\cdot) \text{ a Brownian motion process.} \end{aligned}$$

Suppose that the following stochastic differential equation

$$du = \bar{b}_x(x)u dt + dw \tag{4.7}$$

has a unique solution for each initial condition. The uniqueness is in the sense of in distribution.

Remark: The above conditions require that $b(x,\xi)$ be differentiable at x^*. It essentially allows us to make a 'linearization' around x^*. Many processes (for instance mixing processes with certain mixing rates) satisfying the above weak convergence assumption. Clearly W_n can also be defined by

$$W_n = \sqrt{\epsilon}\frac{1}{T_\epsilon}\int_{N_\epsilon T_\epsilon}^{tT_\epsilon/\epsilon+T_\epsilon} b(x^*,\xi(t))dt. \tag{4.8}$$

Under the above conditions, it can be proved that $\{u^\epsilon(\cdot)\}$ is tight in $D^r[0,\infty)$, and any weakly convergent subsequence has a limit $u(\cdot)$ which is a solution of the stochastic differential equation (4.7).

References

[1] R. Akella and P.R. Kumar, Optimal control of production rate in a failure-prone manufacturing system, *IEEE Trans. Automat. Control* **AC-13** (1986), 116-126.

[2] T. Bielecki and P.R. Kumar, Optimality of zero-inventory policies for unreliable manufacturing systems, *Oper. Res.* **26** (1988), 532-546.

[3] P. Billingsley, *Convergence of Probability Measures*, J. Wiley, New York, 1968.

[4] P. Dupuis and R. Simha, Sampling controlled stochastic approximation, *IEEE Trans. Automat. Control* **AC-36** (1991), 915-924.

[5] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York, 1986.

[6] Y.C. Ho and X. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic, Boston, MA, 1991.

[7] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[8] H.J. Kushner and H. Huang, Asymptotic properties of stochastic approximations with constant coefficients, *SIAM J. Control Optim.* **19** (1981), 87-105.

[9] H.J. Kushner and G. Yin, Asymptotic properties of distributed and communicating stochastic approximation algorithms, *SIAM J. Control Optim.* **25** (1987), 1266-1290.

[10] S.X.C. Lou, S. Sethi, and Q. Zhang, Optimal feedback production planning in a stochastic two–machine flowshop, to appear in *European J. Oper. Res.*

[11] G.C. Papanicolaou, D. Stroock and S.R.S. Varadhan, Martingale approach to some limit theorems, *Proc. of the 1976 Duke Univ. Conference on Turbulence*, M. Reed ed., Duke University Mathematics Series, vol. 3, Durham, NC, 1977.

[12] B.T. Polyak, New stochastic approximation type procedures, *Automat. Remote Control*, **51** (1990), 937-946.

[13] G. Yin, Recent progress in parallel stochastic approximations, in *Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control*, L. Gerencsér and P.E. Caines Eds., 159-184, Springer-Verlag, 1991.

[14] G. Yin, On extensions of Polyak's averaging approach to stochastic approximation, *Stochastics* **36** (1991), 245-264.

An Analysis of Gradient Estimates in Stochastic Network Optimization Problems

Nikolai Krivulin

Two classes of stochastic networks and their performance measures are considered. These performance measures are defined as the expected value of some random variables and cannot normally be obtained analytically as functions of network parameters. We give similar representations for the random variables to provide a useful way of analytical study of these functions and their gradients. The representations are used to obtain sufficient conditions for the gradient estimates to be unbiased. The conditions are rather general and usually met in simulation of the stochastic networks. Applications of the results are discussed and some practical algorithms of calculating unbiased estimates of the gradients are also presented.

1 Introduction

Stochastic network models are widely used in modern engineering, management, biology *etc* to investigate real systems. These models are usually so complicated that can hardly be studied with the help of the analytical methods only. A more fruitful way is to use computer simulation to analyze the networks [1,2,3]. By performing simulation experiments one may get a great amount of information about the network behaviour.

Usually, the main aim of the analysis is to improve a network performance. In order to optimize a performance criterion with respect to network parameters one needs to evaluate it. Simulation provides estimating the criterion as well as its sensitivity (or its gradient, when the parameters are continuous) in a simple way. It is not difficult to obtain estimates provided there exists a simulation model, however each simulation experiment may be very time consumming. Therefore, it is of great importance to develop efficient methods of simulation and estimation.

There are many stochastic optimization procedures which use the data obtained by simulation (see [1] and also a short survey in [4]). In many cases, the procedures that exploit gradient are prefered to those using estimates of the objective function only. The stochastic algorithms which apply unbiased estimates of gradient are often highly efficient. As an example, one can compare the Robbins–Monro procedure with the Kiefer–Wolfowitz one. It is well known [4] that the first procedure based on the unbiased estimates of gradient converges to the solution faster than the second one which approximates the gradient by the finite differences.

In this paper we analyse the problem of unbiased estimation of the gradient of stochastic network performance measures. The paper is based on the results of [5,6]. In Section 1 we describe two classes of stochastic networks and give some examples. We show that the sample performance functions of the networks of both classes may be represented in a similar way. In

fact, these functions are expressed through given ones by using the operations of maximum, minimum and addition.

Section 2 includes a technical result which provides a general representation for the sample performance functions of the networks.

In Section 3 we briefly discuss three methods of estimating gradients, based on simulation data.

The main results are presented in Section 4. Firstly, we introduce a set of functions for which one may obtain unbiased estimates of their gradients. We prove some technical lemmae to state properties of the set. In conclusion, we give the conditions that provide the gradient estimates to be unbiased. These conditions are rather general and usually fulfilled in simulation studies of the stochastic networks.

In Section 5 we show how the results may be applied in practice. Some algoritms of calculating the gradient estimates are described.

2 Stochastic networks and related optimization problems

In this section we present two classes of stochastic networks and discuss optimization problems related to the networks. The performance criterion of the network is normally defined as the expected value of a random variable, $f(\theta,\omega)$, *ie*

$$F(\theta) = E_\omega[f(\theta,\omega)] = E[f(\theta,\omega)],$$

where $\theta \in \Theta \subset R^n$ is a set of decision parameters and ω is a random vector representing the randomness of network behaviour. As a function of the parameters, $f(\theta,\omega)$ is often called sample performance function.

The problem is to optimize the performance measure $F(\theta)$ with respect to $\theta \in \Theta$. In practical problems it is very hard to evaluate the expectation analytically in closed form, even if there is an analytical formula available for $f(\theta,\omega)$. However, it is not difficult to obtain the value of $f(\theta,\omega)$ for any fixed $\theta \in \Theta$ and any realization of ω by using simulation. In that case, one normally use the Monte Carlo approach to estimate the objective function $F(\theta)$ or its gradient.

The main purpose of this section is to show that for many optimization problems, $f(\theta,\omega)$ may be represented in similar algebraic forms. In other words, $f(\theta,\omega)$ is expressed in terms of some given random variables by means of the operations *max*, *min* and $+$. This representation offers the potential for analytical study of the estimates of performance measure gradient. It also provides a theoretical background for efficient algorithms of calculating the estimates.

Activity network. We begin with stochastic activity network models widely used in corporate management in the scheduling of large projects. Consider a project consisting of some activities (or jobs) which must be done to complete it. Each activity is presumed to require a random amount of time for performing it. It is not permitted to begin each activity until some others ones preliminary to it have been completed. One is normally interested in reducing the expected completion time of the whole project.

In order to describe the project as a network, we define an oriented graph $(\mathbf{N},\mathbf{A})$, where $\mathbf{N}$ is the set of nodes and $\mathbf{A}$ is the set of arcs. Each node $i \in \mathbf{N}$ represents the corresponding

activity of the project. For some $i, j \in \mathbf{N}$, the arc (i,j) belongs to $\mathbf{A}$ if and only if the ith activity must precede the jth one directly.

To simplify further formulae we define the set of the father nodes as $\mathbf{N}_F(i) = \{j \in \mathbf{N} | (j,i) \in \mathbf{A}\}$, and the set of the daughter nodes as $\mathbf{N}_D(i) = \{j \in \mathbf{N} | (i,j) \in \mathbf{A}\}$ for every $i \in \mathbf{N}$. In addition, we introduce the set of starting nodes $\mathbf{N}_S = \{i \in \mathbf{N} | \mathbf{N}_F(i) = \emptyset\}$ and the set of the end nodes $\mathbf{N}_E = \{i \in \mathbf{N} | \mathbf{N}_D(i) = \emptyset\}$ of the graph.

Now we have to define the duration of the activities, so that the network would be described completely. Denote the duration of the ith activity by $\tau_i, i \in \mathbf{N}$. We assume τ_i to be a positive random variable, such that $\tau_i = \tau_i(\theta, \omega)$, where $\theta \in \Theta$ is a set of decision parameters and ω is a random vector which represents the random effects involved in realizing the project. The set $\mathbf{T} = \{\tau_i | i \in \mathbf{N}\}$ is presumed to be given.

The sample completion time of the ith activity may be expressed in the form

$$t_i(\theta,\omega) = \begin{cases} \max_{j \in \mathbf{N}_F(i)} t_j(\theta,\omega) + \tau_i(\theta,\omega) & \text{if } i \notin \mathbf{N}_S \\ \tau_i(\theta,\omega) & \text{if } i \in \mathbf{N}_S \end{cases} \tag{1}$$

For the sample completion time of the whole project, we have $t(\theta, w) = \max_{i \in \mathbf{N}_E} t_i(\theta, \omega)$.

In that case, the expected completion time is $T(\theta) = E[t(\theta,\omega)]$, and we wish to minimize $T(\theta)$ with respect to $\theta \in \Theta$.

It is easy to see from (1) that one can represent t as a function of $\tau \in \mathbf{T}$ by using the operations *max* and $+$. To illustrate, consider the simple network depicted in Figure 1.

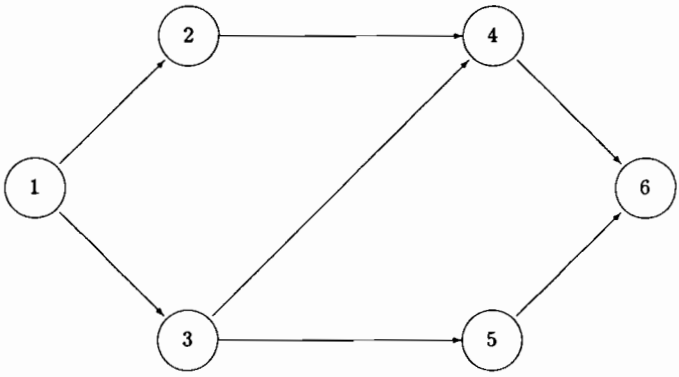


Figure 1. An activity network

For this network, applying (1) successively, we may write the sample completion time as

$$t = \tau_1 + \max\{\max\{\tau_2, \tau_3\} + \tau_4, \tau_3 + \tau_5\} + \tau_6.$$

We will exploit the possibility of t being expressed in such a form in the discussion below.

We conclude this example with the remark about the main difficulty of the activity network optimization problem. It is easy to understand that in the case of general random variables $\tau \in \mathbf{T}$ it is usually very difficult or even impossible to obtain the expected completion time analitically, even if the network is as simple as that in Figure 1. To apply an

optimization procedure in this situation one normally estimate this function or its gradient by using the Monte Carlo approach. Notice, however, that the simulation models of such networks are generally rather simple.

Reliability network. Another class of stochastic network models arises from the reliability investigation of complex interconnected systems in engineering, millitary research, biology etc. Consider a system of elements having bounded random lifetimes. Each element keeps in order until either this element has failed or all those supplying it directly have lost their working conditions. The whole system is presumed to be in order if at least one of the main elements that are supplied by some others but do not supply any element keeps working. An important problem in analyzing this system is to maximize its expected lifetime.

Let $(\mathbf{N}, \mathbf{A})$ be the directed graph discribing the relations between the system elements. In the graph the set of nodes $\mathbf{N}$ corresponds to the set of system elements. If for some $i, j \in \mathbf{N}$, the ith element supplies the jth one directly, then $(i,j) \in \mathbf{A}$. For the graph we retain the notations $\mathbf{N}_F(i), \mathbf{N}_D(i), \mathbf{N}_S$ and $\mathbf{N}_E$ introduced above.

For every element $i \in \mathbf{N}$, we define the lifetime as the random variable $\tau_i(\theta, \omega)$ which depends on the set of decision parameters $\theta \in \Theta$. Assume the set $\mathbf{T} = \{\tau_i\}$ to be given. Now, we may represent the time for the ith element to be in order as

$$t_i(\theta,\omega) = \begin{cases} \min\{\max_{j \in \mathbf{N}_F(i)} t_j(\theta,\omega), \tau_i(\theta,\omega)\} & \text{if } i \notin \mathbf{N}_S \\ \tau_i(\theta,\omega) & \text{if } i \in \mathbf{N}_S \end{cases} \tag{2}$$

The sample and expected lifetimes of the whole system may be written as

$$t(\theta,\omega) = \max_{i \in \mathbf{N}_E} t_i(\theta,\omega) \quad \text{and} \quad T(\theta) = E[t(\theta,\omega)],$$

respectively.

To illustrate this realibility network model consider that depicted in Figure 2 (Ermakov, [1]).

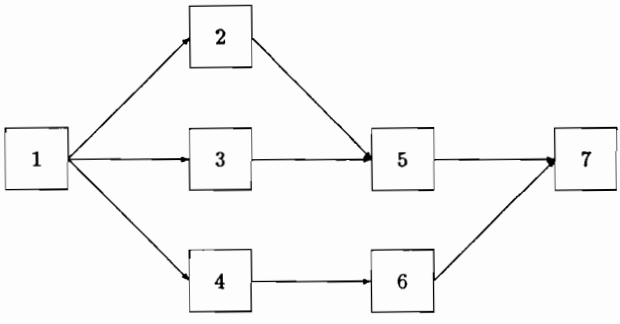


Figure 2. A reliability network

For the sample lifetime of the system, we have from (2)

$$t = \min\{\tau_1, \max\{\min\{\tau_4, \tau_6\}, \min\{\max\{\tau_2, \tau_3\}, \tau_5\}\}, \tau_7\}.$$

We can see that the sample lifetime of such a network has one important property: it may be represented as a function of all $\tau \in \mathbf{T}$ by using only the operations *max* and *min*. Note that the difficulties in solving the problem of expected lifetime maximization are the same as in activity network optimization.

3 An algebraic representation lemma

We have seen that the functions of network performance possesses some algebraic properties. The point is that they may be expressed as a function of given random variables by means of the operations *max*, *min* and $+$. For the networks, one can obtain such representations from recursive equations (1) and (2). In this section we present a general form of the representations, which provides a common way of examining analytical properties of the performance functions in both networks.

In order to simplify futher formulae we introduce the notations $\vee$ for maximum and $\wedge$ for minimum. In addition, we will use the sign $\bigvee$ ($\bigwedge$) to represent an iterated maximum (minimum), *ie*

$$\bigvee_{i=1}^{n} x_i = x_1 \vee \ldots \vee x_n \quad (\bigwedge_{i=1}^{n} x_i = x_1 \wedge \ldots \wedge x_n)$$

Let $\mathbf{X}$ be a set supplied with the operations $+$, $\vee$ and $\wedge$. Without loss of generality we may consider $\mathbf{X}$ to be a set of real numbers. It is easy to extend the result of this section to various sets of real-valued functions and random variables. We assume that the traditional algebraic axioms are fulfilled in $\mathbf{X}$. In particular, we will use the following axioms.

Axiom 1. Distributivity of maximum over minimum.

$\forall x, y, z \in \mathbf{X}$, $(x \wedge y) \vee z = (x \vee z) \wedge (y \vee z)$.

Axiom 2. Distributivity of minimum over maximum.

$\forall x, y, z \in \mathbf{X}$, $(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$.

Axiom 3. Distributivity of sum over maximum and minimum.

$\forall x, y, z \in \mathbf{X}$, $(x \vee y) + z = (x + z) \vee (y + z)$, $(x \wedge y) + z = (x + z) \wedge (y + z)$.

The general form of the representation is determined in the next technical lemma.

Lemma 1. *Let $\varphi(z_1, \ldots, z_p)$ be a function of the variables $z_1, \ldots, z_p$ taking their values in $\mathbf{X}$, φ is defined as a composition of the operations $\vee$, $\wedge$ and $+$. Then φ can be represented as*

$$\varphi(z_1, \ldots, z_p) = \bigvee_{i \in I} \bigwedge_{j \in J_i} \sum_{k=1}^{p} \alpha_{ij}^k z_k,$$

where I and J_i for all $i \in I$ are finite sets of indices, and all α_{ij}^k are integers.

Proof. Without loss of generality we suppose that there is no more than one entry of each variable $z_1, \ldots, z_p$ into the expression. If some variable has two or more entries, we introduce additional ones so that the above presupposition would be fulfilled. Let us prove the lemma by induction on the number of variables.

For $p = 1$, the statement of the lemma is obvious. If $p = 2$, there are three possibilities

$$z_1 \vee z_2, \quad z_1 \wedge z_2 \quad \text{and} \quad z_1 + z_2,$$

and it is clear that the statement is also true.

Assume that the statement of the lemma is true up to some value $p - 1$. Consider an expression φ of p variables. Clearly, there is an operation in the expression that should be performed after the other ones. Denote this operation by the asterisk $*$. In this case, we have $\varphi = \varphi_1 * \varphi_2$, where φ_1 and φ_2 are expressions such that each of them cannot include all the variables $z_1, \ldots, z_p$. By the assumption, the statement of the lemma holds for both φ_1 and φ_2. Now, we have three possibilities for the operation $*$.

1. $\vee$. This is obvious.
2. $\wedge$. It is sufficient to apply Axiom 1.
3. $+$. To obtain the representation in this case, one has to apply successively Axioms 1, 2 and 3.

Consequently, the statement of the lemma is true for $\varphi = \varphi_1 * \varphi_2$. □

4 Estimates of gradient

To optimize the network performance measure $F(\theta) = E[f(\theta, \omega)]$ one often needs information about the gradients $\partial F(\theta)/\partial\theta$. In the absence of analytical formulae for the gradient, Monte Carlo experiments may be performed to estimate its values. There are three general methods of estimating $\partial F(\theta)/\partial\theta$ based on data obtained by simulation [1,3,7]. In the first two methods the gradient is approximated by the finite differences and then estimated by using the Monte Carlo approach. To illustrate these two methods, assume θ to be a scalar and consider the following estimates:

The crude Monte Carlo (CMC) estimate:

$$G_{CMC} = \frac{1}{N\Delta\theta} \sum_{i=1}^{N} (f(\theta + \Delta\theta, \omega_i) - f(\theta, \omega_{N+i}))$$

The common random number (CRN) estimate:

$$G_{CRN} = \frac{1}{N\Delta\theta} \sum_{i=1}^{N} (f(\theta + \Delta\theta, \omega_i) - f(\theta, \omega_i))$$

where $\omega_i, i = 1, \ldots, 2N$ are independent realizations of the random vector ω. The second estimate differs from the first in one respect: in the CRN estimate the random variables ω_i are the same for both $\theta + \Delta\theta$ and θ, whereas in the CMC estimate they are different. Note that each of them requires $2 \times N$ simulation runs (N at the original value θ and N at $\theta + \Delta\theta$). Clearly, in the case of the vector $\theta \in R^n$, one must perform $(n + 1) \times N$ simulation experiments to get each estimate. In [1, pp. 153-154] Ermakov has shown that the finite difference estimates have the mean square error (MSE) which is of order $O(N^{-1/3})$ for G_{CRN} and $O(N^{-1/2})$ for G_{CRN}.

We may somewhat improve the MSE properties of the estimate by using more sophisticated difference formulae, however, the estimates become very expensive in terms of computer time because they require a large number of additional simulation experiments. For example, the following symmetric difference estimate

$$G_{CRN}^{SD} = \frac{1}{2N\Delta\theta}\sum_{i=1}^{N}(f(\theta+\Delta\theta,\omega_i) - f(\theta-\Delta\theta,\omega_i))$$

requires $2 \times N$ simulation runs ($2 \times n \times N$, when $\theta \in R^n$).

An estimate of the third method can be written in the form

$$G = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}f(\theta,\omega_i), \tag{3}$$

provided that the gradient of the sample performance function (sample gradient) exists. It should be noted that, although we may obtain values of the sample performance function by simulation, it can be rather difficult to evaluate its gradient.

Recently, a new technique, called infinitesimal perturbation analysis (IPA), has been developed (Ho *et al.* [2]) as an efficient method of obtaining gradient information. The IPA method yields the exact values of the sample gradient $\partial f(\theta,\omega)/\partial\theta$ by performing one simulation run. The method is based on the analysis of the dynamics of the network and closely connected with the simulation technique. Therefore, one can easily combine an IPA procedure for calculating the sample gradient with a suitable algorithm of network simulation. Such a procedure provides all the partial derivatives of the sample gradient simultaneously during one simulation run. Furthermore, it needs an additional computation cost which is usually very small compared with that required for the simulation run alone.

The key question concerning the IPA method is whether it produces an unbiased estimate of the performance measure gradient. It can easily be shown that if $\partial f(\theta,\omega)/\partial\theta$ is an unbiased estimator of $\partial F(\theta)/\partial\theta$ then estimate (3) has MSE which is of order $O(N^{-1})$. In short, in the case of unbiasedness, this is a very efficient estimate, that provides considerable savings in computation.

In the next section, using the algebraic representation of Section 3, we will examine properties of the network performance functions so as to derive the conditions for estimate (3) of the performance measure gradient to be unbiased.

5 A theoretical background of unbiased estimation

A sufficient condition for the estimate (3) of $\partial E[f(\theta,\omega)]/\partial\theta$ at some $\theta \in \Theta$ to be unbiased is

$$\frac{\partial}{\partial\theta}E[f(\theta,\omega)] = E[\frac{\partial}{\partial\theta}f(\theta,\omega)]. \tag{4}$$

Cao in [7] showed that (4) holds in the case of $f(\theta,\omega)$ being uniformly differentiable at θ w.p.1. Note that such a differentiability property is not easy to verify and hard to interpret for practical systems. A useful way to prove the interchange in (4) is to apply the Lebesgue dominated convergence theorem (Loève [8]). We use this theorem in the following form.

Theorem 2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\Theta \subset R^n$ and $f : \Theta \times \Omega \longrightarrow R$ be a $\mathcal{F}$-measurable function for any $\theta \in \Theta$ and such that the following conditions hold:*

(i) *for every $\theta \in \Theta$, there exists $\partial f(\theta, \omega)/\partial\theta$ at θ w.p.1,*

(ii) *for all $\theta_1, \theta_2 \in \Theta$, there is a random variable $\lambda(\omega)$ defined on the same probability space, with $E\lambda < \infty$ and such that*

$$|f(\theta_1, \omega) - f(\theta_2, \omega)| \leq \lambda(\omega)\|\theta_1 - \theta_2\| \quad w.p.1. \tag{5}$$

Then equation (4) holds on Θ.

As an important consequence, we may state that the function $F(\theta) = E[f(\theta, \omega)]$ is a Lipschitz one with a constant $L = E\lambda$ and continuously differentiable on Θ, provided f satisfies the theorem conditions.

Definition 1. A function $f(\theta, \omega)$ defined on the probability space $(\Omega, \mathcal{F}, P)$ at every $\theta \in \Theta$ belongs to the set $\mathcal{D}_{\Theta,\Omega}$ (or simply $\mathcal{D}$) if and only if it satisfies the conditions of Theorem 2.

Example 1. Random variables which arise from simulation study of networks, can be treated as members of a family of random variables [1]. There are few families one usually applies, namely the Exponential family, the Gaussian family *etc.* Various random variables of a family may be obtained from the standard variable by using a suitable transformation. An ordinary way to transform random variables is based on changing location and scale parameters.

Let $\xi(\omega)$ be the standard random variable of a family. Define

$$f(\theta, \omega) = \theta_1 \xi(\omega) + \theta_2,$$

where $\theta = (\theta_1, \theta_2)^\top \in \Theta \subset R^2$. Let us check whether it holds that $f \in \mathcal{D}$. Obviously, the partial derivatives of f with respect to θ_1 and θ_2 exist for almost all ω and equal

$$\frac{\partial}{\partial\theta_1} f(\theta, \omega) = \xi(\omega) \quad \text{and} \quad \frac{\partial}{\partial\theta_2} f(\theta, \omega) = 1.$$

In addition, it is easy to verify that f satisfies Condition (ii) of Theorem 2 with $\lambda = |\xi| + 1$. If $E|\xi| < \infty$, as is usually the case, then the conditions of Theorem 2 are fulfilled for f and we have $f \in \mathcal{D}$.

The next technical lemmae give the sufficient conditions for the arithmetic operations and the operation *max* and *min* not to break the main properties of the functions from $\mathcal{D}$.

Lemma 3. *Let $f, g \in \mathcal{D}$ and let λ_1 and λ_2 be the random variables that provide Condition (ii) of Theorem 2 for f and g, respectively. Let μ_1, μ_2 and ν be positive random variables. Then the following are satisfied.*

(i) $f + g \in \mathcal{D}$.

(ii) *If α is a bounded random variable, then $\alpha f \in \mathcal{D}$.*

(iii) *If $|f| \leq \mu_1$ and $|g| \leq \mu_2$ hold w.p.1 for any $\theta \in \Theta$ and $E[\lambda_1\mu_2 + \lambda_2\mu_1] < \infty$, then $fg \in \mathcal{D}$.*

(iv) *If $|f| \leq \mu_1$ and $|g| \geq \nu$ hold w.p.1 for any $\theta \in \Theta$ and $E[\frac{\mu_1\lambda_2}{\nu^2} + \frac{\lambda_1}{\nu}] < \infty$, then $\frac{f}{g} \in \mathcal{D}$.*

Proof. Clearly, $f+g$, αf, fg and f/g are measurable functoins of ω and differentiable ones on Θ w.p.1. Since for all of these functions the proofs of inequality (5) are quite similar, we verify it for one of them only. For instance, we examine $h = fg$.

For all $\theta_1, \theta_2 \in \Theta$ we have

$$\begin{aligned} |h(\theta_1,\omega) - h(\theta_2,\omega)| &= |f(\theta_1,\omega)g(\theta_1,\omega) - f(\theta_2,\omega)g(\theta_2,\omega)| = \\ |f(\theta_1,\omega)g(\theta_1,\omega) - f(\theta_2,\omega)g(\theta_1,\omega) + f(\theta_2,\omega)g(\theta_1,\omega) - f(\theta_2,\omega)g(\theta_2,\omega)| &\le \\ |g(\theta_1,\omega)||f(\theta_1,\omega) - f(\theta_2,\omega)| + |f(\theta_2,\omega)||g(\theta_1,\omega) - g(\theta_2,\omega)| &\le \\ (\lambda_1(\omega)\mu_2(\omega) + \lambda_2(\omega)\mu_1(\omega))\|\theta_1 - \theta_2\| \quad \text{w.p.1.} \end{aligned}$$

In short, $|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda(\omega)\|\theta_1 - \theta_2\|$ w.p.1, where $\lambda = \lambda_1\mu_2 + \lambda_2\mu_1$, $E\lambda = E[\lambda_1\mu_2 + \lambda_2\mu_1] < \infty$. By Theorem 2, we conclude $fg \in \mathcal{D}$. □

Notice, from Lemma 3 (i) and (ii) it follows that being closed for the operations of addition and multiplication by bounded random variables, $\mathcal{D}$ is a linear space of functions with these two operations.

Lemma 4. *Let $f, g \in \mathcal{D}$. Suppose that for any $\theta_0 \in \theta$, there exists a neighbourhood $\mathbf{U}_\omega(\theta_0)$ of θ_0 w.p.1 such that one and only one of the following conditions*

(i) $f(\theta,\omega) = g(\theta,\omega)$,

(ii) $f(\theta,\omega) < g(\theta,\omega)$,

(iii) $f(\theta,\omega) > g(\theta,\omega)$

is satisfied for all $\theta \in \mathbf{U}_\omega(\theta_0)$. Then $f \vee g \in \mathcal{D}$ and $f \wedge g \in \mathcal{D}$.

Proof. Consider $h(\theta,\omega) = f(\theta,\omega) \vee g(\theta,\omega)$. It is clear that h is measurable with respect to ω. In order to prove differentiability of h w.p.1 on Θ, we examine an arbitrary $\theta \in \Theta$. There are only two possibility for h not to be differentiable. Firstly, it is possible that the derivative of h at θ does not exist if at least one of the derivatives $\partial f(\theta,\omega)/\partial\theta|_{\theta=\theta_0}$ and $\partial g(\theta,\omega)/\partial\theta|_{\theta=\theta_0}$ does not. In addition, h may not be differentiable at θ if the maximum of the functions f and g changes over from f to g at this point or vice versa. The last case is equivalent to that there exists $\omega \in \Omega$ such that all the neighbourhoods $\mathbf{U}_\omega(\theta_0) \subset \Theta$ contain both points at which $f(\theta,\omega) = g(\theta,\omega)$ and $f(\theta,\omega) \neq g(\theta,\omega)$. By the assumption of the lemma, both of these cases may occur only with zero probability. Therefore, there exists $\partial h(\theta,\omega)/\partial\theta|_{\theta=\theta_0}$ at all $\theta \in \Theta$ w.p.1.

For the function h, the proof will be completed if we show that h satisfies Condition (ii) of Theorem 2. Since $f, g \in \mathcal{D}$, there are random variables λ_1 and λ_2 with $E\lambda_1 < \infty$ and $E\lambda_2 < \infty$ such that the inequalities

$$\begin{aligned} |f(\theta_1,\omega) - f(\theta_2,\omega)| &\le \lambda_1(\omega)\|\theta_1 - \theta_2\| \quad \text{w.p.1} \\ |g(\theta_1,\omega) - g(\theta_2,\omega)| &\le \lambda_2(\omega)\|\theta_1 - \theta_2\| \quad \text{w.p.1} \end{aligned}$$

hold for all $\theta_1, \theta_2 \in \Theta$. Let ω be an arbitrary element of Ω at which both these inequalities hold. Devide Θ into two subsets:

$$\begin{aligned} \mathbf{X}_\omega &= \{\theta \in \Theta | f(\theta,\omega) \ge g(\theta,\omega)\}, \\ \mathbf{Y}_\omega &= \{\theta \in \Theta | f(\theta,\omega) < g(\theta,\omega)\}. \end{aligned}$$

Obviously, it holds

$$|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda_1(\omega)\|\theta_1 - \theta_2\|$$

for all $\theta_1, \theta_2 \in \mathbf{X}_\omega$ and

$$|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda_2(\omega)\|\theta_1 - \theta_2\|$$

for all $\theta_1, \theta_2 \in \mathbf{Y}_\omega$. Assume $\theta_1 \in \mathbf{X}_\omega, \theta_2 \in \mathbf{Y}_\omega$. If $h(\theta_1,\omega) \ge h(\theta_2,\omega)$, we deduce

$$|h(\theta_1,\omega) - h(\theta_2,\omega)| = |f(\theta_1,\omega) - g(\theta_2,\omega)| < |f(\theta_1,\omega) - f(\theta_2,\omega)| \le \lambda_1(\omega)\|\theta_1 - \theta_2\|.$$

Similarly, if $h(\theta_1,\omega) < h(\theta_2,\omega)$, we have $|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda_2(\omega)\|\theta_1 - \theta_2\|$. It follows that $|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda(\omega)\|\theta_1 - \theta_2\|$, $\lambda(\omega) = \lambda_1(\omega) \vee \lambda_2(\omega)$, for all $\theta_1, \theta_2 \in \Theta$. Since this inequality holds for almost all $\omega \in \Omega$, we conclude that

$$|h(\theta_1,\omega) - h(\theta_2,\omega)| \le \lambda(\omega)\|\theta_1 - \theta_2\| \quad \text{w.p.1},$$

and $E\lambda = E[\lambda_1 \vee \lambda_2] \le E\lambda_1 + E\lambda_2 < \infty$.

In other words, h satisfies the conditions of Theorem 2. Consequently, $f \vee g \in \mathcal{D}$. The proof of the statement $f \wedge g \in \mathcal{D}$, is analogous. □

It should be noted that the condition of Lemma 4 is not necessary, as the next example shows.

Example 2. Let $\Theta = [-1,1]$, $(\Omega, \mathcal{F}, P)$ be a probability space, where $\Omega = [0,1]$, $\mathcal{F}$ is the σ–field of Borel sets of Ω and P is the Lebesgue measure on Ω. Consider the following functions:

$$f(\theta,\omega) = -\theta^3 + \omega, \quad g(\theta,\omega) = \theta^2 + \omega$$

and

$$h(\theta,\omega) = f(\theta,\omega) \vee g(\theta,\omega) = \begin{cases} -\theta^3 + \omega & \text{if} \quad -1 \le \theta \le 0 \\ \theta^2 + \omega & \text{if} \quad 0 < \theta \le 1 \end{cases}$$

One can easily verify that for any neighbourhood of $\theta = 0$, there exist both points with $f > g$ and $f < g$ w.p.1. The conditions of Lemma 4 are therefore violated. Nevertheless, h is differentiable at 0 for all $\omega \in \Omega$. In addition, it holds that $h \in \mathcal{D}$.

Corollary 5. *Let $f, g \in \mathcal{D}$. If for every $\theta \in \Theta$ it holds that $f \ne g$ w.p.1, then $f \vee g \in \mathcal{D}$ and $f \wedge g \in \mathcal{D}$.*

Proof. Clearly, the condition of the corollary implies that either $f - g > 0$ or $f - g < 0$ holds at every $\theta \in \Theta$ w.p.1. Since $f, g \in \mathcal{D}$, these two functions are continuous ones of θ w.p.1 as well as $f - g$. Because of continuity, $f - g > 0$ $(f - g < 0)$ holds w.p.1 not only at θ, but also at every points of a neighbourhood of θ. It remains to apply Lemma 4. □

Using Corollary 5 we give the following general conditions for $\mathcal{D}$ to provide closeness with respect to the operations $\vee$ and $\wedge$.

Lemma 6. *Let $f, g \in \mathcal{D}$. If for any $\theta \in \Theta$ it holds that the random variables $f(\theta,\omega)$ and $g(\theta,\omega)$*

(i) *are independent, and*

(ii) *at least one of them is continuous*

then $f \vee g \in \mathcal{D}$ and $f \wedge g \in \mathcal{D}$.

To prove the lemma it is sufficient to see that its conditions lead to that of Corollary 5.

The next two examples show that both conditions of Lemma 6 are essential.

Example 3. Let $(\Omega, \mathcal{F}, P)$ and Θ be defined as in Example 2. Also define

$$f(\theta,\omega) = -\theta + \omega \quad \text{and} \quad g(\theta,\omega) = \theta + \omega.$$

Let us consider the function

$$h(\theta,\omega) = f(\theta,\omega) \vee g(\theta,\omega) = \begin{cases} -\theta+\omega & \text{if } -1 \le \theta \le 0 \\ \theta+\omega & \text{if } 0 < \theta \le 1 \end{cases}$$

It is clear that $f, g \in \mathcal{D}$ and for every $\theta \in \Theta$, the random variables $f(\theta,\omega)$ and $g(\theta,\omega)$ are continuous. Although inequality (5) holds with $\lambda = 1$ for h, this function is not differentiable at $\theta = 0$ for all $\omega \in \Omega$. Therefore, $h \notin \mathcal{D}$.

Example 4. Let $\Theta = [0,1]$, $\Omega_1 = \Omega_2 = [0,1]$ and P be the Lebesgue measure on $\Omega = \Omega_1 \times \Omega_2$. Denote $\omega = (\omega_1, \omega_2)^\top$ and consider the following functions:

$$f(\theta,\omega) = \begin{cases} \frac{1}{2}\theta & \text{if } \omega_1 \le \frac{1}{2} \\ 1 & \text{if } \omega_1 > \frac{1}{2} \end{cases}$$

$$g(\theta,\omega) = \begin{cases} \theta^2 & \text{if } \omega_2 \le \frac{1}{2} \\ 1 & \text{if } \omega_2 > \frac{1}{2} \end{cases}$$

and

$$h(\theta,\omega) = f(\theta,\omega) \vee g(\theta,\omega) = \begin{cases} \max\{\frac{1}{2}\theta, \theta^2\} & \text{if } \omega_1 \le \frac{1}{2} \text{ and } \omega_2 \le \frac{1}{2} \\ 1 & \text{otherwise} \end{cases}$$

One can see that $f, g \in \mathcal{D}$ and for every $\theta \in \Theta$, the random variables $f(\theta,\omega)$ and $g(\theta,\omega)$ are independent. In addition, the condition (ii) of Theorem 2 holds for h with $\lambda = 2$. Nevertheless, $h = \max\{\frac{1}{2}\theta, \theta^2\}$ with probability $\frac{1}{4}$, that is not a differentiable function at $\theta = \frac{1}{2}$. In that case, $h \notin \mathcal{D}$.

Lemma 7. *Let $\mathcal{M}$ be a set of functions from $\mathcal{D}$ such that for all $f, g \in \mathcal{M}$, the conditions of Lemma 4 are fulfilled. Then $\mathcal{M}$ is closed for the operations* max *and* min.

Proof. Let $f, g \in \mathcal{M}$ and let us define $h = f \vee g$. Note that $h \in \mathcal{D}$ by Lemma 4. We have to prove the conditions of Lemma 4 to be satisfied for h and any $u \in \mathcal{M}$.

If u is either f or g, say $u \equiv f$, we may write

$$h - u = f \vee g - f = \begin{cases} g - f & \text{if } f < g \\ 0 & \text{if } f \ge g \end{cases}$$

Since $f, g \in \mathcal{M}$, for any point of Θ, there is a neighbourhood on which only one of the conditions $f - g < 0, f - g = 0$ or $f - g > 0$ holds w.p.1. From the above identity this also holds for $h - u$ on the neighbourhood. Consequently, in this case the conditions of Lemma 4 are fulfilled.

Now we assume $u \in \mathcal{M} \setminus \{f, g\}$. We have

$$h - u = f \vee g - u = \begin{cases} g - f & \text{if } f < g \\ f - u & \text{if } f \ge g \end{cases}$$

Let us examine any $\theta \in \Theta$. Suppose that $f < g$ w.p.1 at θ. Since f, g and u belong to $\mathcal{M}$, there are neighborhoods $\mathbf{U}_\omega(\theta)$ and $\mathbf{V}_\omega(\theta)$ where the conditions of Lemma 4 are fulfilled

for each pairs of functons (f,g) and (g,u), respectively. It follows from the expression of h that the neighborhood $\mathbf{U}_\omega \bigcup \mathbf{V}_\omega(\theta)$ is that Lemma 4 requires for h and u. If it holds that $f \geq g$ or $f = g$ at θ, the reasoning is the same.

In short, we have shown that the conditions of Lemma 4 are satisfied for h and any $u \in \mathcal{M}$, and therefore, $h = f \vee g \in \mathcal{M}$. In the case of minimum the proof is analogous. □

Corollary 8. *If $f_j \in \mathcal{M}$ for every $j = 1,\ldots,N$, then it holds*

$$\bigvee_{i\in I} \bigwedge_{j\in J_i} f_j \in \mathcal{M},$$

where I is a finite set of indices and $J_i \subset \{1,\ldots,N\}$ for every $i \in I$.

This is an immediate consequence of the previous lemma.

The next example is of importance to the main result of the section.

Example 5. Let $f_j \in \mathcal{D}$ for all $j = 1,\ldots,N$. Suppose that at every $\theta \in \Theta$, all the random variables $f_j(\theta,\omega)$ are continuous and independent. Define $\mathcal{L}$ to be a set of linear combinations $\sum_{i\in I} a_i f_i$ with integer coefficients a_i, $i \in I \subset \{1,\ldots,N\}$. Obviously, $\mathcal{L}$ is stable for addition. For all functions $u = \sum_{i\in I} a_i f_i$ and $v = \sum_{j\in J} b_j f_j$ we have $u - v = \sum_{k\in K} c_k f_k$. It is clear that for every $\theta \in \Theta$, $u - v$ is a continuous random variable because of the properties of f (except for the case of all $c_k = 0$ which is obvious). Therefore, it holds that $u - v \neq 0$ w.p.1 at every $\theta \in \Theta$. Similarly as in Corollary 5, one can deduce that u and v satisfy the conditions of Lemma 4. From this we conclude that $\mathcal{L}$ may be treated as an example of $\mathcal{M}$.

One can easily see that the condition of continuity is essential to this reasoning. To illustrate the important role of independence, consider the following functions

$$f(\theta,\omega) = -2\theta + 2\omega, \quad g(\theta,\omega) = \theta - \omega \quad \text{and} \quad u(\theta,\omega) = \theta + \omega$$

under the same assumption as in Example 3. It is easy to verify that the conditions of Lemma 4 are fulfilled for any two functions of them. Nevertheless, the functions u and $v = f + g$ do not satisfy them, as Example 3 has shown.

Now, we may formulate the main result of the section. We first introduce some definitions. Let $\mathcal{A}$ be the algebra of all functions $f : \Theta \times \Omega \longrightarrow R$ being defined on the probability space $(\Omega, \mathcal{F}, P)$ at every $\theta \in \Theta$ with the operations $\vee, \wedge$ and $+$. In other words, this is a closed system of the functions for these operations.

Definition 2. Let $\mathbf{T}$ be a finite subset of functions of $\mathcal{A}$. We define $[\mathbf{T}]_{\mathcal{A}}$ to be the set generated by $\mathbf{T}$ in $\mathcal{A}$, that is the set of all functions being obtained from ones of $\mathbf{T}$ by means of the operations $\vee, \wedge$ and $+$.

Theorem 9. *Let $\mathbf{T} \in \mathcal{D}$. Suppose that for all $\tau \in \mathbf{T}$, $\tau(\theta,\omega)$ are continuous and independent random variables at any $\theta \in \Theta$. Then it holds $[\mathbf{T}]_{\mathcal{A}} \subset \mathcal{D}$.*

Proof. It results from Lemma 1 that every $f \in [\mathbf{T}]_{\mathcal{A}}$ can be represented as

$$f = \bigvee_{i\in I} \bigwedge_{j\in J_i} \sum_{\tau\in\mathbf{T}} a_{ij}^{\tau} \tau,$$

where all a_{ij}^{τ} are integers. It has been shown in Example 5 that the functions of the family $\{\sum_{\tau \in \mathbf{T}} a_k^{\tau} \tau\}_{k=1,2,\ldots}$ satisfy the conditions of Lemma 4. Applying Corollary 8, we conclude that the statement of the theorem is true. □

It is important to note that the conditions of Theorem 9 are rather general and usually fulfilled in the network simulation. In particular, in contrast with the traditional approaches (cp, for example, existing results on the unbiasedness of IPA estimates in [2,3]), we may not restrict ourself to the exponential distribution.

In short, to satisfy the theorem only the following are required for the functions of the set $\mathbf{T}$:

(i) for any $\theta \in \Theta$, all $\tau \in \mathbf{T}$ are continuous and independent random variables;

(ii) each $\tau \in \mathbf{T}$ as a function of θ is differentiable w.p.1 and Lipschitz one with an integrable random variable as a Lipschitz constant.

In the next section we will show how these results can be applied to some problems to verify the unbiasedness of gradient estimates.

6 Applications

Now we discuss the applications of the previous results to optimizing the networks. In particular, we describe algorithms of obtaining sample gradients, based on the algebraic representation of the networks. In this section we keep using the notations $(\Omega, \mathcal{F}, P)$ and Θ for the underlying probability space and the parameter space, respectively.

We begin with the stochastic activity network. Let the duration of the jth activity be represented by the function $\tau_j(\theta, \omega)$. Denote the set of all such functions of the network by $\mathbf{T}$. As we have seen, a sample completion time of the network $t(\theta, \omega)$ may be expressed by functions of $\mathbf{T}$ by using only the operations *max* and $+$. This implies $t \in [\mathbf{T}]_{\mathcal{A}}$.

Suppose that $\mathbf{T} \in \mathcal{D}$ and all $\tau \in \mathbf{T}$ are continuous, and they are independent random variables at every $\theta \in \Theta$. For the mean completion time $T(\theta) = E[t(\theta, \omega)]$, it follows from Theorem 2 that $\frac{1}{N} \sum_{i=1}^{N} \partial t(\theta, \omega_i)/\partial\theta$, where $\omega_i \in \Omega$, is an unbiased estimate of the gradient $\partial T(\theta)/\partial\theta$.

As an example, suppose $\tau(\theta, \omega) = -\theta \ln(1-\omega)$, where $\theta \in R$ and the random variable ω is uniformly distrubuted on $[0,1]$. It is well known [1] that $-\ln(1-\omega)$ has an exponential distribution with mean 1. Similarly as in Example 1, we have $\tau \in \mathcal{D}$. In addition, durations of the activities are normally considered as independent in the probabilistic sense. Our result is, therefore, applicatable in this case.

Now, suppose that there is a simulation procedure for the activity network with L nodes to provide a simulation experiment for any fixed $\theta \in \Theta$ and a realization of ω. One can easily combine it with the following algorithm.

Algorithm 1.

Step (i). At the initial time, fix values of θ and ω; set $g_j = 0$ for $j = 1, \ldots, L$, and set $c = 0$.

Step (ii). At the completion of any activity i, add the value of $\partial\tau_i(\theta, \omega)/\partial\theta$ to g_i and add 1 to c;
if $c = L$, then save g_i as the value of $\partial t(\theta, \omega)/\partial\theta$ and stop; otherwise go to Step (iii).

Step (iii). Determine the set $\mathbf{N}_D(i)$. For every $j \in \mathbf{N}_D(i)$, if all activities of the set $\mathbf{N}_F(j)$ have been completed, then set $g_j = g_i$.

To verify the correctness of Algorithm 1 it sufficies to see that it is simply based on recursive equation (1). Note that Algorithm 1 is similar to those based on the IPA method in [3].

For a reliability network, one can apply Theorem 2 in a similar way. As in Section 2, denote the sample lifetime of a system by $t(\theta,\omega)$. It is not difficult to construct the next algorithm that calculates the sample gradient $\partial t(\theta,\omega)/\partial\theta$.

Algorithm 2.

Step (i). At the initial time, fix values of θ and ω.

Step (ii). At the failure of element i, exclude all nodes representing the elements that are now not able to keep working from the set $\mathbf{N}$ as well as the corresponding arcs from the set $\mathbf{A}$.

Step (iii). If for the reduced set $\mathbf{N}$ it holds $\mathbf{N} \cap \mathbf{N}_E = \emptyset$, then save $\partial\tau_i(\theta,\omega)/\partial\theta$ as the value of $\partial t(\theta,\omega)/\partial\theta$ and stop; otherwise go to Step (ii).

In conclusion, note that both algorithms are rather simple. In fact, they only require calculating gradients of given functions and performing some trivial operations to produce values of the sample gradients. Using these values, one can easily estimate the gradients of the system performance measures so as to apply efficient optimization procedures.

References

[1] Ermakov, S.M. (1975). Die Monte-Carlo-Methode und verwandte Fragen. VEB Deutscher Verlag der Wissenschaften.

[2] Ho, Y.C. (1987). Performance evaluation and perturbation analysis of discrete event dynamic systems. IEEE Trans. Automat. Contr., AC-32, no. 7, 563–572.

[3] Suri, R. (1989). Perturbation analysis: The state of the art and research issues explained via the GI/G/1 queue. In Proc. of IEEE, 77, no. 1, 114–137.

[4] Glynn, P.W. (1986). Optimization of stochastic systems. In Proc. of 1986 Winter Simulation Conf., 52–58.

[5] Krivulin, N.K. (1990). On complex system optimization using simulation. Vestnik Leningrad. Univ. Mat. Meh. Astronom., 2(8), 100–102. (in Russian)

[6] Krivulin, N.K. (1990). Optimization of discrete event dynamic systems by using simulation. Ph.D. Dissertation. St Petersburg University. (in Russian)

[7] Cao, X.R. (1985). Convergence of parameter sensitivity estimates in a stochastic experiment. IEEE Trans. Automat. Contr., AC-30, no. 9, 845–853.

[8] Loève, M. (1960). Probability Theory. Van Nostrand.

Records of Simulated Annealing

Ryszard Zieliński

A sufficient condition is given for the simulated annealing (SA) process not to leave a basin of a local minimum if the basin is deep enough. An improvement of the SA process is suggested and some results of test computations are reported.

1 Introduction

Given a real–valued function f on a set $\mathcal{X}$, let $X = (x_n, n \geq 1)$ be a simulated annealing (SA) process (see e.g. AJMMS 1988 or a very short and clear presentation of the method in Schoen 1991). When looking for a global minimum of f on $\mathcal{X}$ what we are really interested in is the process $Y = (y_n, n \geq 1)$ of records of the process X defined as

$$y_1 = x_1,$$

$$y_n = \begin{cases} x_n, & \text{if } f(x_n) < f(x_{n-1}), \\ y_{n-1}, & \text{otherwise.} \end{cases}$$

The process X can be accepted as satisfactory if the process Y almost surely converges to a global minimum of f.

Given a sequence $(T_n, n \geq 1)$ of positive reals, the general structure of X (without stopping rule) is as follows (without loss of generality we assume that $f > 0$ on $\mathcal{X}$):

1. let x_1 be a point in $\mathcal{X}$;
2. let ξ_n be a random neighbour of x_n;
3. if $f(\xi_n) \leq f(x_n)$ then put $x_n = \xi_n, n = n + 1$, and go to 2;
4. if $f(\xi_n) > f(x_n)$ then
   a. let R_n be a uniform random number in $(0, 1)$;
   b. if $R_n < \exp[-(f(\xi_n) - f(x_n))/T_n]$, then put $x_n = \xi_n$; otherwise leave x_n without change. Put $n = n + 1$ then go to 2.

Let $\mathcal{F}_n = \sigma(x_1, x_2, \ldots, x_n)$. Define the following sequence of random events

$$A_n = \{f(\xi_n) > f(x_n),\ R_n < exp[-(f(\xi_n) - f(x_n))/T_n]\}$$

and let $A = \{A_n, i.o.\}$, where $i.o.$ as usually denotes "infinitely often". It is obvious that if $P(A) = 0$ then the process X will not leave a basin of a local minimum if the basin is deep enough.

2 Results

The following theorem gives us an insight into the structure of SA processes from the point of view of the convergence of their records to a global minimum. As a matter of fact the theorem presents a simple applications of the following version of the Borel–Cantelli lemma: *if* $(\mathcal{F}_n, n \geq 1)$ *is an increasing sequence of* σ*-fields and* $(B_n, n \geq 1)$ *is a sequence of random events such that* $B_n \in \mathcal{F}_n$*, then the events* $\{B_n, i.o.\}$ *and* $\{\sum_{n=1}^{\infty} P(B_{n+1}|\mathcal{F}_n) = \infty\}$ *are almost sure equal* (see e.g. Hall and Heyde (1980), p. 32).

Let $(R_n, n \geq 1)$ be a sequence of random variables uniformly distributed on $(0,1)$. Throughout the paper we assume that given $\mathcal{F}_n$ the random elements $\xi_n, n = 1, 2, \ldots$ and $R_n, n = 1, 2, \ldots$ are (conditionally) independent.

THEOREM .*If for an* $\alpha > 1$

$$\sum_{n=2}^{\infty} P\{f(x_n) < f(\xi_n) < f(x_n) + \alpha T_n \log n | \mathcal{F}_n\} < \infty \tag{1}$$

almost surely, then $P(A) = 0$.

P r o o f. For the random event A_n we have

$$A_n = A_n^{(1)} + A_n^{(2)},$$

where

$$A_n^{(1)} = A_n \cap \{T_n \leq \frac{f(\xi_n) - f(x_n)}{\alpha \log n}\},$$

$$A_n^{(2)} = A_n \cap \{T_n > \frac{f(\xi_n) - f(x_n)}{\alpha \log n}\}.$$

Now

$$P\left(A_n^{(1)}|\mathcal{F}_n\right) \leq P\{R_n < \exp[-\alpha \log n]\} = \frac{1}{n^{\alpha}}$$

and hence $\sum_{n=1}^{\infty} P\left(A_n^{(1)}|\mathcal{F}_n\right) < \infty$.

The probability of $A_n^{(2)}$ may be estimated as

$$P\left(A_n^{(2)}|\mathcal{F}_n\right) \leq P\{f(x_n) < f(\xi_n) < f(x_n) + \alpha T_n \log n | \mathcal{F}_n\},$$

and hence if (1) then $\sum_{n=1}^{\infty} P\left(A_n^{(2)}|\mathcal{F}_n\right) < \infty$. It follows that $P\{A_n, i.o.\} = 0$.

3 Comments

It is clearly seen from the theorem that Y can converge to a global minimum of f only when the probabilities

$$p_n = P\{f(x_n) < f(\xi_n) < f(x_n) + \alpha T_n \log n | \mathcal{F}_n\}$$

are large enough; if p_n tend to zero they should tend not faster than $n^{-\gamma}$ for some $\gamma > 1$ or faster than $(\log n)^{-1}$. This however heavily depends on the function f which typically is out of our control, and on the sequence T_n and the probability laws of the sample points ξ_n which we are able to manipulate.

If $T_n \log n$ tends to zero then the random event

$$\{f(x_n) < f(\xi_n) < f(x_n) + \alpha T_n \log n\}$$

is asymptotically empty; if $T_n \log n = O(n^{-\gamma})$ for some $\gamma > 1$ then the condition (1) is satisfied even for the linear function $f(x) = f(x_n) + c \cdot (x - x_n)$ and ξ_n uniformly distributed over the ball $B(x_n, \rho)$, $\rho > 0$ [E.g. for $f : \mathcal{R}^2 \to \mathcal{R}^1$, if $f(x) = f(x_n) + c \cdot (x - x_n)$ and ξ_n is uniformly distributed over the ball $B(x_n, \rho)$, then $P\{f(x_n) < f(\xi_n) < f(x_n) + \alpha T_n \log n | \mathcal{F}_n\} = P\{x_n < \xi_n < x_n + \alpha T_n \log n / c | \mathcal{F}_n\} < \alpha T_n \log n \rho / c$.]

If, as in Bochaczevsky et al. (1986) and Brooks and Verdini (1988), $T_n = \lambda \cdot f(\xi_n)$ for a $\lambda > 0$, then

$$\{f(\xi_n) < f(x_n) + \alpha T_n \log n\} = \{(1 - \alpha\lambda \log n) f(\xi_n) < f(x_n)\}$$

which is asymptotically an almost sure event. It follows that then the performance of

$$P\{f(x_n) < f(\xi_n) | \mathcal{F}_n\}$$

as n tends to infinity is crucial.

4 An improvement

To ensure the almost sure convergence of records it may be enough to design the sequence $(\xi_n, n \geq 1)$ in such a way that for infinitely many n's the support of the probability distribution of ξ_n is the whole $\mathcal{X}$, which gives us a result analogous to that for random search for global minima in the Theorem 3.3 in Zieliński and Neumann (1983). To see how the application of the idea from that theorem can improve the SA procedure consider the following example.

5 Example

Suppose the problem is to find a global minimum of the following function

$$f(x, y) = \sin(3.14x) \cdot \sin(3.14y) + 0.1(x + y), \quad 0 \leq x \leq 1, \ 0 \leq y \leq 1.$$

There exists exactly one local (and global) minimum at the point $(0,0)$. Suppose that the SA process with $T_n = 1/ln(n)$ starts from the point $(1,1)$ where the function has its local maximum. The problem is that to achieve the global minimum the SA process has to go around the maximum of the function in the center of the unit square. The contour map of the function on the unit square is presented in Fig. 1. Suppose that the candidate ξ_n is distributed uniformly over the rectangular $(x_n - \lambda/2, x_n + \lambda/2) \cap ([0,1] \times [0,1])$ with some $\lambda \in (0,1)$. In our numerical experiment we have chosen $\lambda = 0.1$; then the area of the support of the random candidate is equal to a number between 0.0025 and 0.01 which is rather large if one takes into account that the ratio of the volume of the unit ball in $\mathcal{R}^k$ and the volume of the "unit" interval $[-1,1]^k$ is equal to, e.g., 0.00249 for $k = 10$ and $2.5 \cdot 10^{-8}$ for $k = 20$.

Parallel we considered the modification of the SA process consisting in that every $s = 10$ steps the candidate has the uniform distribution over the whole domain $[0,1]^2$. Typical result after $n = 1000$ steps is shown in Fig. 2 where the classical SA process is presented in the upper part and the modified process in the bottom part of the picture. In the left-hand-side part of Fig. 2 the trajectories of processes X and in the right-hand-side those of the process Y are presented (the origin is situated in the northwest vertex).

In the Table 1 the number of steps needed in the two processes to achieve the ε-neighborhood of the point of global minimum with $\epsilon = 0.1$ in ten experiments (with the same random numbers in every pair of experiments) are given. The mean number of steps in 100 experiments for the original SA process was equal 2,815.85 and that for the modified one 158.26 (in 1 out of 100 experiments the original process terminated earlier then the modified one).

Table 1

4605	442
1362	209
680	122
3324	86
748	55
1008	103
820	30
10354	192
1584	229
6032	40

Table 2

5022	790
3769	223
9564	6942
712	1762
1468	542
29757	10107
12286	13534
11953	11650
11255	4833
9249	1910

As a second example consider the function

$$g(x,y) = (x^2 + y^2)[(1-x)^2 + (1-y)^2] + 0.3x - 0.1y - [(x+1)(2-y)]^{-8}.$$

The function has a rather flat local minimum at the point $(0, 0.031)$ with $f(0, 0.031) = 0.969$ and a deep global minimum at the point $(0,1)$ with $f(0,1) = 0.9$. The function is presented in Fig. 3. When starting the search process from the origin the superiority of the modified procedure is not so apparent because there exist a short direct passage from the starting point to the point of global minimum, easy to follow by standard SA process. Table 2 for this function is constructed in full analogy to the Table 1 for the function $f(x,y)$. The mean number of steps in 100 experiments for the original SA process was equal to 10,049.11

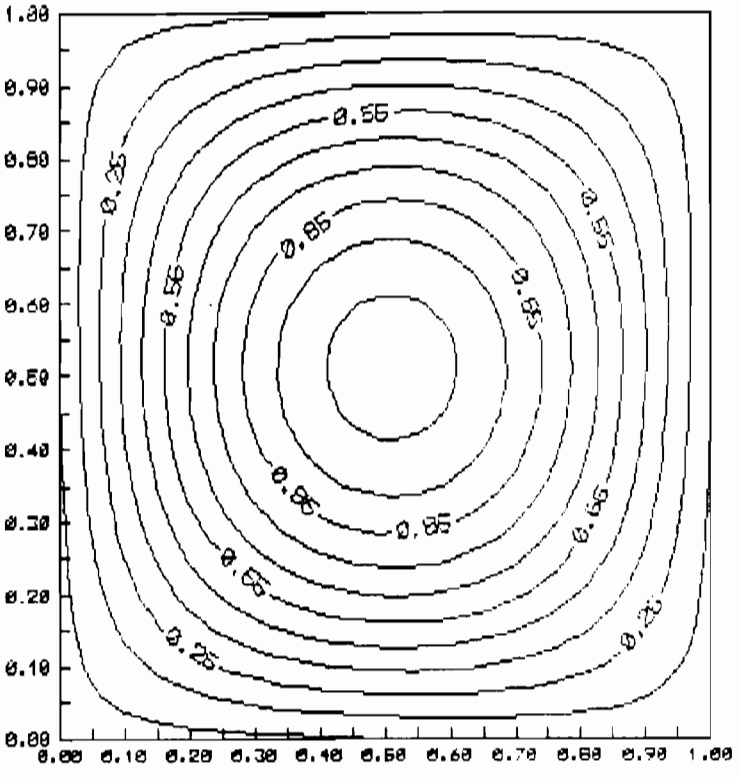


Figure 1

and that for the modified one 3,734.37 (in 28 out of 100 experiments the original SA process terminated earlier then the modified one).

In my opinion neither version of SA can beat the well known multistart method with a good algorithm for descending to local minimum. A statistical analysis of that method is presented in Zieliński (1981).

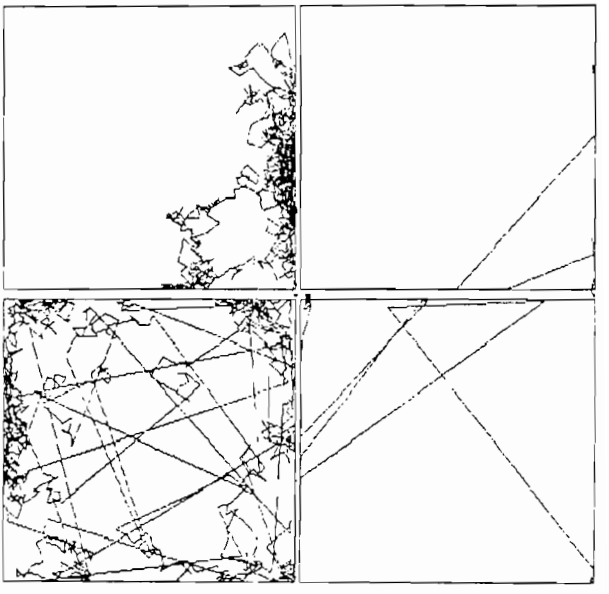


Figure 2

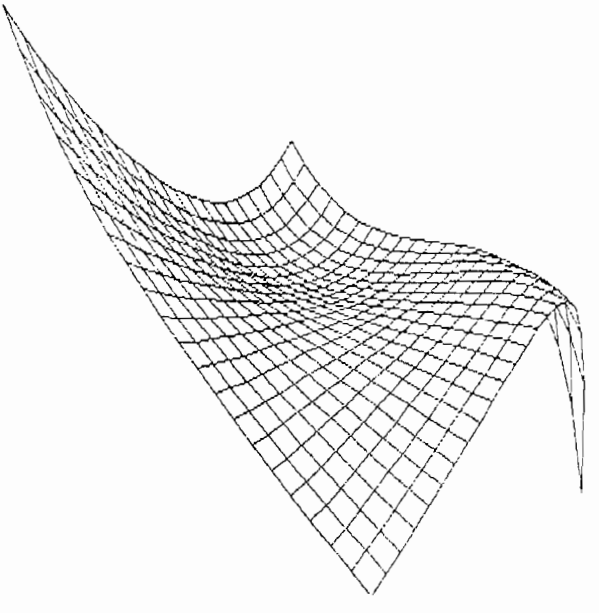


Figure 3

References

AJMMS (1988), *American Journal of Mathematical and Management Sciences,* vol. 8, Nos. 3 & 4

Bochaczevsky, I.O., Johnson, M.K., Stein, M.L. (1986), *Generalized Simulated Annealing for Function Optimization.* Technometrics 28, 28, 209–217

Brooks, D.G., Verdini, W.A. (1988), *Computational Experience with Generalized Simulated Annealing over Continuous Variables.* AJMMS 8, Nos. 3 & 4, 425–449

Hall, P., Heyde, C.C. (1980), *Martingale Limit Theory and Its Applications.* Academic Press

Schoen, F. (1991), *Stochastic Techniques for Global Optimization: A survey of Recent Advances.* Journal of Global Optimization 1, 207–228

Zieliński, R. (1981), *A statistical estimate of the structure of multi-extremal problem.* Mathematical Programming 21, 348–356. North Holland.

Zieliński, R., Neumann, P. (1983), *Stochastische Verfahren zur Suche nach dem Minimum einer Funktion.* Akademie–Verlag, Berlin

Markov Sequences as Optimization Algorithms

A.S. Tikhomirov

The matter of the article is further investigation of random search methods presented in [1]. An upper bound on the number of steps of search methods guaranteeing the attainment of the accuracy ϵ with the reliability γ is considered.

1 Introduction. Statement of the problem and preliminary results.

As it was shown in [1], there exists such a homogeneous Markov random search sequence (m.r.s.s.) that it takes in average only $O(\ln^2 \epsilon)$ computations of objective function $f : X \to R$ to determine $x_0 = \arg\max f$ with given accuracy ϵ. Here the next portion of features of such m.r.s.s. is proposed.

First of all we give the upper bound on the number of steps of our m.r.s.s., guaranteeing the attainment of the accuracy ϵ with the reliability γ. The other part of the work is devoted to the minimization of the constant, multiplying $\ln^2 \epsilon$ in the estimates obtained. Here we show, how the *a priori* information on the objective function f gives us the opportunity to diminish the complexity of random search method.

For the purpose of convenience we begin with some definitions and preliminary results borrowed from [1]. Proofs of some assertions are omitted for the sake of brevity.

Let (X, ρ) be a compact metric space equipped with Borel σ-algebra $\mathcal{B}_X$. We suppose that there exists such a measure $\mu : \mathcal{B}_X \to [0,1]$ that the following conditions are fulfilled:

CX1. If $D_\epsilon(x) = \{y \in X : \rho(x,y) < \epsilon\}$ then $0 < \mu(D_\epsilon(x)) \leq 1$ for all $x \in X$ and $\epsilon > 0$.

CX2. If $S_\epsilon(x) = \{y \in X : \rho(x,y) \leq \epsilon\}$ then $0 < \mu(S_\epsilon(x)) = \mu(S_\epsilon(z))$ for all $x, z \in X$ and $\epsilon > 0$.

Thus $\varphi : [0,1] \to [0,1]$ defined by $\varphi(r) = \mu(S_r(x))$ does not depend on x and is a *cadlag* nondecreasing function.

CX3. For every $x \in X$ and $\epsilon \geq 0$ $\partial S_\epsilon(x) = \{y : \rho(x,y) = r\}$.

We suppose that $diam\, X = \mu(X) = 1$. In addition we require that

CX4. $\varphi \in C([0,1])$.

CX5. There exists such a function $b : [0,\infty) \to (0,1]$ that for any $a \geq 1$ and $r \geq 0$ with $0 \leq ar \leq 1$

$$\varphi(r) \geq b(a)\varphi(ar). \tag{1}$$

One can easily deduce from (1) that for all r

$$\varphi(r) \geq Ar^B \tag{2}$$

with some positive constants A and B.

Let $f : (X, \mathcal{B}_X) \to (R, \mathcal{B}_R)$ be some measurable bounded function satisfying the following conditions:

CF1. The function f achieves its maximum value at the unique point $x_0 = \arg\max f(x) \in X$.

CF2. Function f is continuous at the point x_0.

CF3. For each $r > 0$ $\sup\{f(x); x \in S_r^c(x_0)\} < f(x_0)$.

Let $(\Omega, \mathcal{F}, P)$ be some probability space. A sequence of random elements $\{\xi_i\}_{i\geq 0}$, $\xi_i : (\Omega, \mathcal{F}) \to (X, \mathcal{B}_X)$ will be further called as a random search sequence (r.s.s.) acting in $(X, \mathcal{B}_X)$. If this sequence is a Markov one with respect to the filtration $\mathcal{F}_n = \sigma(\xi_0, \ldots, \xi_n)$ it will be named a Markov random search sequence, if for each $i > 0$ $f(\xi_i) \geq f(\xi_{i-1})$ P-a.s. it will be called a monotone one.

Further, denoting $B_f(x) = \{y \in X : f(y) \geq f(x)\}$ we consider the homogeneous Markov family $(\{\xi_i\}_{i=0}^{\infty}, P_x, x \in X)$ with the transition function

$$R(x, .) = \delta_x(.)P(x, B_f^c(x)) + P(x, . \cap B_f(x)). \tag{3}$$

Here $P(x, .)$ is a probability measure for each $x \in X$ and $P(., A)$ is B_X-measurable for each $A \in B_X$. Surely $P_x(\xi_0 = x) = 1$. For the sake of brevity we shall write S_δ instead of $S_\delta(x_0)$.

Let us define the aim of our random search method as a determination of the point x_0 with a given accuracy $\epsilon > 0$. Thus we are interested in the fact that ξ_i hits a set S_ϵ on some step i. As in [1] we introduce a family of sets

$$M_\delta^f = M_\delta = \{y \in S_\delta : f(y) > \ f(z) \text{ for every } z \in S_\delta^c\}.$$

It is easy to see that $M_{\delta_1} \supset M_{\delta_2}$ for $\delta_1 > \delta_2$ and $f(x) > f(y)$ for $x \in M_\delta$ and $y \notin M_\delta$. As we suppose our r.s.s. to be monotone one it remains in M_δ once hitting it. Thus we are interested in events $\{\xi_i \in M_\epsilon\}$ and not in $\{\xi_i \in S_\epsilon\}$.

Further we shall often be concerned with a function

$$F_f(r) = F(r) = \mu(M_r)/\mu(S_r) = \mu(M_r)/\varphi(r).$$

F may be considered as a characteristics of the objective function f. As in [1], we shall suppose that

CF4. $\inf\{F(r), 0 < r \leq 1\} = F > 0$.

One can easily demonstrate the following proposition:

Lemma 1.1 . Let $X \in R^d$, $x_0 \in intX \neq \emptyset$, $f \in C^2(S_\delta(x_0))$ for some $\delta > 0$ and the matrix of second derivatives of f is non-degenerate at the point x_0. Then CF4 takes place and

$$F(r) \to F_0 = \sqrt{\prod_{i=1}^{d} \frac{\lambda_{\max}}{\lambda_i}} \qquad as\ r \to 0\ ,$$

where λ_i are eigen-values of the matrix $\partial^2 f(x_0)$ and $\lambda_{\max} = \max \lambda_i < 0$.

Let $\tau_\epsilon = \min\{n : \xi_n \in M_\epsilon\}$. Surely τ_ϵ is a nice characteristics of the r.s.s. Main results of the work may be described as follows: there exists such a homogeneous m.r.s.s. that for every $\alpha \in R^1$ and some positive constants $C_1(X,F)$ and $C_2(X,F)$ the inequality

$$\limsup_{\epsilon \downarrow 0} \sup_{x \in X} P_x(\tau_\epsilon > C_1(X,F)\ln^2 \epsilon + \alpha\sqrt{C_2(X,F) \mid \ln^3 \epsilon \mid}\,) \le 1 - \Phi(\alpha)$$

takes place, where Φ is $\mathcal{N}(0,1)$ distribution function. The transition function R is chosen to minimize C_1.

2 Asymptotic behaviour of τ_ϵ

Given $1 \le n < \infty$ let $\{r_i\}_{i=0}^n$ be some sequence of real numbers such that $r_0 = 1$, $r_{i-1} > r_i$, $r_n = \epsilon$. Let $M_0 = X$, $M_i = M_{r_i}$, $N_n = M_n$, $N_i = M_i \backslash M_{i+1}$, $m_i = \mu(M_i)$ and $\tau_i = \min\{k : \xi_k \in M_i\}$. First of all we shall demonstrate several preliminary results.

Lemma 2.1 . Let the sequence of numbers $\{u(i,k)\}$, $0 \le i \le k$, $1 \le k \le n$ satisfy the following conditions: $u(i,k) \le \inf(P(y, M_k), y \in N_i)$, $0 \le i < k$, $1 \le k \le n$ and $u(k,k) = 1$. Then the inequality

$$P_x(\xi_l \in M_k^c) \le \sum_{i=1}^{k} (u(i,k) - u(i-1,k)) P_x(\xi_{l-1} \in M_i^c) \tag{4}$$

takes place for all $l \ge 1$.

Proof. As we suppose our r.s.s. to be monotone one it remains in M_i once hitting it. Therefore we have

$$P_x(\xi_l \in M_k^c) = \sum_{i=0}^{n} P_x(\xi_l \in M_k^c, \xi_{l-1} \in N_i) = \sum_{i=0}^{k-1} P_x(\xi_l \in M_k^c, \xi_{l-1} \in N_i) \le$$

$$\le \sum_{i=0}^{k-1} (1 - \inf_{N_i} P(y, M_k)) P_x(\xi_{l-1} \in N_i) \le \sum_{i=0}^{k-1} (1 - u(i,k)) P_x(\xi_{l-1} \in M_i \backslash M_{i+1}) =$$

$$= \sum_{i=1}^{k-1} (1 - u(i,k))(P_x(\xi_{l-1} \in M_{i+1}^c) - P_x(\xi_{l-1} \in M_i^c)) + (1 - u(0,k)) P_x(\xi_{l-1} \in M_1^c) =$$

$$= (1 - u(k-1,k)) P_x(\xi_{l-1} \in M_k^c) + \sum_{i=1}^{k-1} (u(i,k) - u(i-1,k)) P_x(\xi_{l-1} \in M_i^c).$$

The proof is complete.

Let $\{\alpha\}_{i=2}^n$ and $\{\theta\}_{i=1}^n$ be two sequences of independent random variables such that $P(\alpha_i = 0) = q_i$, $P(\alpha_i = 1) = 1 - q_i$, $P(\theta_i = k) = u_i(1-u_i)^k$, $k = 0,1,2,\dots$, where $q_i, u_i \in (0,1)$.

Let $t_1 = 1 + \theta_1$, $t_i = \alpha_i(1 + \theta_i)$ and $T_k = \sum_{i=1}^k t_i$.

Theorem 2.1. Let $q_i = m_i/m_{i-1}$ for all $2 \le i \le n$ and $\inf(P(y, M_k); y \in N_i) \ge u_{i+1} m_k / m_{i+1}$ for all $0 \le i < k \le n$. Then inequalities

$$P_x(\tau_k > l) \le P(T_k > l) \tag{5}$$

take place for all $k \geq 1,\ l \geq 0$.

Proof. In the case of $l = 0$ (5) is evident. We shall concern the case $l \geq 1$. As the inequality (4) takes place for $P_x(\tau_k > l)$ all we need is to demonstrate the following relation:

$$P(T_k > l) = \sum_{i=1}^{k}(u(i,k) - u(i-1,k))P(T_i > l-1) \tag{6}$$

with $u(i,k) = u_{i+1}m_k/m_{i+1}$. If $k = 1$ then (6) can be easily demonstrated. We shall concern the case $k \geq 2$. Surely (6) reduces to (7):

$$P(T_k = l) = \sum_{i=1}^{k-1} u(i,k)P(T_i \leq l-1,\ T_{i+1} > l-1) + u(0,k)P(T_1 > l-1). \tag{7}$$

We shall concern the cases of $l = 1$ and $l > 1$ separately. If $l = 1$ then (7) reduces to $P(T_k = 1) = u(0,k)$ that is easy to verify:

$$P(T_k = 1) = P(\sum_{i=1}^{k} t_i = 1) = u_1 \prod_{i=2}^{k} m_i/m_{i-1} = u_1 m_k/m_1 = u(0,k).$$

Let $l > 1$. With the help of simple computations one may obtain that

$$\sum_{i=1}^{k-1} u(i,k)P(T_i \leq l-1,\ T_{i+1} > l-1) + u(0,k)P(T_1 > l-1) =$$

$$= \sum_{i=1}^{k-1} P(T_k = l, t_{i+1} \neq 0, t_{i+2} = 0, \ldots, t_k = 0) + P(T_k = l, t_2 = 0, \ldots, t_k = 0).$$

Let $A_i = \{T_k = l, t_{i+1} \neq 0, t_{i+2} = 0, \ldots, t_k = 0\}$ for all $1 \leq i \leq k-1$, $A_k = \{T_k = l, t_2 = 0, \ldots, t_k = 0)$ and $A = \bigcup_{i=1}^{k} A_i$. It is easy to see that $A = \{T_k = l\}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Therefore $P(T_k = l) = \sum_{i=1}^{k} P(A_i)$. The proof is complete.

As it may be easily seen

$$ET_k = 1/u_1 + \sum_{i=2}^{k}(1 - q_i)/u_i$$

$$\text{and}\quad DT_k = 1/u_1^2 + \sum_{i=2}^{k}(1 - q_i^2)/u_i^2 - ET_k.$$

Let us discuss some properties of $m : (0,1] \to [0,1]$ which is defined by $m(r) = \mu(M_r)$.

Lemma 2.2. m is a right continuous nondecreasing function such that $0 < m(r) \leq \varphi(r)$.

Proof. As CF1-CF3 take place we need to show only right continuous feature of m. Let $M_{r+0} = \bigcap_{n=1}^{\infty} M_{r+1/n}$. We are going to prove that $M_r = M_{r+0}$. Surely $M_r \subset M_{r+0}$. Let us prove the opposite insertion. Let $x \in M_{r+0}$, then $x \in M_{r+1/n}$ for all n, therefore $\rho(x, x_0) \leq r + 1/n$ and $f(x) > f(z)$ for all $z \in S^c_{r+1/n}$ with all n. So we have $\rho(x, x_0) \leq r$ and $f(x) > f(z)$ for all $z \in \bigcup_{n=1}^{\infty} S^c_{r+1/n}$. As $\bigcup_{n=1}^{\infty} S^c_{r+1/n} = S^c_r$ then $x \in M_r$. So

$m(r) = \mu(M_r) = \mu(M_{r+0}) = \mu(\bigcap_{n=1}^{\infty} M_{r+1/n}) = \lim_{n\to\infty} \mu(M_{r+1/n}) = \lim_{n\to\infty} m(r+1/n)$. The proof is complete.

Lemma 2.3. Let f be a continuous function and $\mu(f^{-1}(c)) = 0$ for all $c \in R$. Then m is continuous at (0,1].

Proof. Let $M_{r-0} = \{x \in D_r : f(x) > f(z) \text{ for all } z \in D_r^c\}$. By the compactness of D_r^c we have $M_{r-0} = \{x \in D_r : f(x) > \max\{f(y);\ y \in D_r^c\}\}$. It is easy to see that $M_{r-0} = \bigcup_{n=k}^{\infty} M_{r-1/n}$, where $k = [1/r] + 1$. As f and φ are continuous we have $\mu(M_r \backslash M_{r-0}) = \mu(\{x \in D_r : f(x) = \max\{f(y);\ y \in D_r^c\} \text{ and } f(x) > f(z) \text{ for all } z \in S_r^c\}) \le \mu(\{x \in D_r : f(x) = \max\{f(y);\ y \in D_r^c\}\})$. The proof is complete.

Now let us consider $P(x,.)$. We suppose here that there exists the Radon-Nikodym derivative $p(y,x) = P(x,dy)/\mu(dy)$ such that $p(y,x) = \pi(\rho(y,x))$ where $\pi : [0,1] \to (0,\infty)$ is nonincreasing strictly positive function.

We put here $r_1 = 1/2$. Let us estimate $P(x, M_k)$ for $x \in N_i$ and $i < k$. We have $P(x, M_k) \ge \mu(M_k) \inf\{\pi(\rho(x,y));\ x \in S_{r_i},\ y \in S_{r_k}\} \ge m_k \pi(r_i + r_k) \ge m_k \pi(2r_i - 0)$.

Let $g : [0,1/2] \to (0,\infty)$ be defined by

$$g(r) = \begin{cases} \pi(0), & r = 0 \\ \pi(2r-0), & r \in (0,1/2]. \end{cases} \tag{8}$$

As p is a density we have $1 = \int_X \pi(\rho(x,y))\mu(dy) = \int_0^1 \pi(r)d\varphi(r) = \int_0^{1/2} g(r)d\varphi(2r)$.

Let the sequences $\{r_i(n)\}_{i=0}^n$ be such that $r_0 = 1$, $r_1 = 1/2$, $r_{i-1} > r_i$, $r_n = \epsilon$ and $\max_{2\le i\le n}(r_{i-1}(n) - r_i(n)) \to 0$ as $n \to \infty$. Let $\{t_i(n)\}$, $\{T_i(n)\}$ be defined by Theorem 2.1 with $u_1(n) = m(r_1)g(r_1)$ and $u_i(n) = m(r_i(n))g(r_{i-1}(n))$.

Let finally

$$I(m,g,\epsilon) = \frac{1}{m(r_1)g(r_1)} + \int_{(\epsilon,1/2]} \frac{1}{g(r)} d(-\frac{1}{m(r)}),$$

$$D(m,g,\epsilon) = \frac{1}{(m(r_1)g(r_1))^2} + \int_{(\epsilon,1/2]} \frac{1}{g^2(r)} d(-\frac{1}{m^2(r)}) - I(m,g,\epsilon).$$

Surely $ET_n \to I(m,g,\epsilon)$, $DT_n \to D(m,g,\epsilon)$ as $n \to \infty$.

One can easily deduce from Theorem 2.1 the following assertion:

Theorem 2.2.

$$E_x\tau_\epsilon \le I(m,g,\epsilon). \tag{9}$$

We are going to discuss the behavour of τ_ϵ while ϵ tends to zero.

Lemma 2.4. Let the sequences $\{r_i\}_{i=0}^n$, $\{t_i\}_{i=1}^n$, $\{T_i\}_{i=1}^n$ satisfy mentioned conditions, and $n = n(\epsilon)$ tends to infinity as ϵ tends to zero. Let $\{r_i(n)\}$ be chosen in such a manner that with some $0 < q < 1$ the inequalities $r_i(n) \ge q r_{i-1}(n)$ take place. Let CF4 be valid, $v(n) = \max_{1\le i\le n} u_i(n)$, $w(n) = \min_{1\le i\le n} u_i(n)$ and $v(n)/(w(n) \mid \ln\varphi(\epsilon) \mid) \to 0$ as $\epsilon \to 0$.

Then the central limit theorem is valid for T_n.

Proof.As t_i are independent random elements all we need is to show that Lyapounoff condition is fulfilled (see[2]). The proof is simple and therefore is omitted.

Theorem 2.3. Let CF4 be valid and $P(x,.)$ has a density $p(y,x) = \pi(\rho(x,y))$, where

$$\pi(r) = \lambda(\epsilon)c(r)\begin{cases} \varphi^{-1}(\epsilon), & r \in [0,\epsilon] \\ \varphi^{-1}(r), & r \in (\epsilon,1] \end{cases}$$

with $0 < A \le c(r) \le B < \infty$ for some constants A and B.

Then the inequality

$$\limsup_{\epsilon \downarrow 0} \sup_{x \in X} P_x(\tau_\epsilon > I(m,g,\epsilon) + \alpha\sqrt{D(m,g,\epsilon)}) \le 1 - \Phi(\alpha) \tag{10}$$

takes place, where Φ is the $\mathcal{N}(0,1)$ distribution function and

$$I(m,g,\epsilon) \le C_1(X,F)\ln^2\epsilon, \tag{11}$$

$$D(m,g,\epsilon) \le C_2(X,F) \mid \ln^3\epsilon \mid . \tag{12}$$

The proof of (10) is seen immediately from lemma 2.4. The proof of inequalities (11) and (12) is mainly the same as in [1] and therefore is omitted.

3 Optimization of $I(m,g,\epsilon)$

Now we are interested in the estimator of $E_x\tau_\epsilon$ given by Theorem 2.2.

Let $\mathcal{M}$ be the set of all nonincreasing left continuous strictly positive functions g satisfying the condition $\int_0^{1/2} g(r)d\varphi(2r) = 1$, and $\mathcal{M}_{A,B}$ be a subset of $\mathcal{M}$ such that $A \le g \le B$. Our aim is to find $g \in \mathcal{M}$ that minimize $I(m,g,\epsilon)$ for given m and ϵ. We propose m to be continuous.

It is evident that $\inf\{I(m,g,\epsilon);\ g \in \mathcal{M}\} \in [0,\infty)$.

Lemma 3.1. Let $g \in \mathcal{M}$ and

$$p(x) = \frac{q(x)}{\int_0^{1/2} q(x)d\varphi(2x)}, \text{ where } q(x) = \begin{cases} g(\epsilon - 0), & x \in [0,\epsilon] \\ g(x), & x \in (\epsilon, 1/2]. \end{cases} \tag{13}$$

Then the following three assertions are fulfilled: 1) $p \in \mathcal{M}$; 2) $I(m,p,\epsilon) \le I(m,g,\epsilon)$; 3) if there exists such $x \in (0,\epsilon]$ that $p(x) \ne g(x)$ then $I(m,p,\epsilon) < I(m,g,\epsilon)$.

Proof. The first relation is evident. As $I(m,p,\epsilon) = I(m,g,\epsilon)\int_0^{1/2} q(x)d\varphi(2x)$ the proof of two other relations is complete.

Lemma 3.2. Let $A = m(\epsilon)/m(1/2)$, $B = \varphi^{-1}(2\epsilon)$ and $g \in \mathcal{M}\backslash\mathcal{M}_{A,B}$. Then there exists such $p \in \mathcal{M}_{A,B}$ that $I(m,p,\epsilon) < I(m,g,\epsilon)$.

Proof. As $A \leq 1 < B$ we may choose either $p \equiv 1$ or p defined by (13). The rest is trivial.

Theorem 3.1. Let A and B be taken from lemma 3.2. Then there exists such $p \in \mathcal{M}_{A,B}$ that $I(m,p,\epsilon) = \inf\{I(m,g,\epsilon);\ g \in \mathcal{M}\}$.

The proof may be done with the help of the first Helly Theorem and Lebesgue Theorem on majorized convergence (see[3]).

One can easily deduce from the second Helly Theorem (see[3]) that $\min\{I(m,g,\epsilon);\ g \in \mathcal{M}\}$ possesses a stability property with respect to small perturbations of m and φ. Therefore we propose m and φ to be smooth functions.

Let $\mathcal{N}$ be the set of all nonincreasing left continuous strictly positive densities on $[0,1]$. Making the change of variables $x = \varphi(2y)$ and using the notation $n(x) = m(\varphi^{-1}(x)/2)$, $p(x) = g(\varphi^{-1}(x)/2)$, $h(x) = \sqrt{n'(x)/n^2(x)}$ and $\delta = \varphi(2\epsilon)$ we have $p \in \mathcal{N}$ and

$$I(m,g,\epsilon) = \frac{1}{n(1)p(1)} + \int_{\delta}^{1} \frac{h^2(x)}{p(x)}dx.$$

For the sake of brevity we shall write $I(p)$ instead of $I(m,g,\epsilon)$.

Theorem 3.2. There exists such $p \in \mathcal{N}$ that the following propositions are fulfilled: $I(p) = \min\{I(g); g \in \mathcal{N}\}$, $p \in C([0,1])$ and $p = q/\lambda$ where $q \equiv 1$ or

$$q(x) = \begin{cases} h(b_0), & x \in [a_0, b_0] \\ h(a_j), & x \in [a_j, b_j], \quad j \in J \\ h(x), & \text{otherwise.} \end{cases}$$

Here $a_0 = 0$, $\delta \leq b_0 < 1$, $J \subset N = \{1,2,\ldots\}$, $(a_j,b_j) \cap (a_i,b_i) = \emptyset$ for all $i \neq j$,

$$h^2(b_0) = \frac{1}{b_0}\int_{\delta}^{b_0} h^2(x)dx, \quad h^2(a_j) = \frac{1}{b_j - a_j}\int_{a_j}^{b_j} h^2(x)dx \quad \text{for } b_j < 1$$

and $$h^2(a_j) = \frac{1}{1-a_j}\left(\int_{a_j}^{1} h^2(x)dx + \frac{1}{n(1)}\right) \quad \text{for } b_j = 1.$$

The proof is rather long and therefore is omitted.

The following simplest example will conclude the matter. Let $X = [0,1)^d$ be equipped with such a metrics: if $x, y \in X$, $x = (u_1,\ldots,u_d)$, $y = (z_1,\ldots,z_d)$ then $\rho(x,y) = 2\max_{1\leq i\leq n} \min(|\ u_i - z_i\ |,\ 1 - |\ u_i - z_i\ |)$. Evidently $diam X = 1$.

Let μ be the Lebesgue measure on Borel subsets of X. We have $\varphi(r) = \mu(S_r(x)) = r^d$ for each $0 \leq r \leq 1$.

Let $f_1(x) = -\sum_{i=1}^{d} a_i^2(u_i - 1/2)^2$ where $x = (u_1,\ldots,u_d)$ and $\{a_i\}$ is some sequence of strictly positive real numbers. We consider the family of functions f_δ with $\delta \in [\sqrt[d]{2}\epsilon, 1]$ such that $f_\delta(x) = \max\{\ f_1(x),\ \sup\{f_1(y);\ y \in S_\delta^c\}\ \}$. Evidently $x_0 = \arg\max f_\delta(x) = (1/2,\ldots,1/2)$. It is easy to see that

$$F_{f_\delta}(u) = F_\delta(u) = F\begin{cases} 1, & u \in [0,\delta] \\ (\delta/u)^d, & u \in (\delta,1] \end{cases}$$

with $F = C_d 2^{-d} \prod_{i=1}^{d} a_{\min}/a_i$, where C_d is the Euclidean volume of the unit ball in R^d, $a_{\min} = \min a_i$.

The density minimizing the estimator of $E_x \tau_\epsilon$ given by Theorem 2.2 will be further called as optimal density. One can easily deduce from Theorem 3.2 the following assertion:

Statement. Let $p_\delta(y,x) = \pi_\delta(\rho(x,y))$ and

$$\pi_\delta(u) = \lambda_\delta \begin{cases} (a\epsilon)^{-d}, & 0 \le u \le a\epsilon \\ u^{-d}, & a\epsilon < u \le R(\delta) \\ b^{-1}(\delta), & R(\delta) < u \le 1, \end{cases}$$

where $a = 2\sqrt[d]{2}$, $R(\delta) = 2\min\{\delta, a^{-1}\}$ and $b(\delta) = \sqrt{R^d(\delta) - R^{2d}(\delta)}$.

Then p_δ is the optimal density for f_δ and $E_x \tau_\epsilon \le I(m, g_\delta, \epsilon) = 2^d d^2 (|\ln \epsilon| + s(\delta))^2$, where $s(\delta) = \ln(R(\delta)/a) + (1 + b(\delta) R^{-d}(\delta))/d$, g_δ is defined by (8) and $m_\delta = F_\delta \varphi$.

We see that the optimal density doesnot depend on $\{a_i\}$, but δ must be known in advance. When the *a priori* information is absent, we may use p_1 that is an optimal density for all f with $F_f \equiv const$. To compare p_1 with optimal density p_δ we put $C(\epsilon, \delta) = I(m_\delta, g_1, \epsilon)/I(m_\delta, g_\delta, \epsilon)$. It is easy to see that $C(\epsilon, \delta) \to d/2 \,|\ln \epsilon|\, v(\epsilon)$ as $\delta \to \sqrt[d]{2}\epsilon$, where $v(\epsilon) \to 1$ as $\epsilon \to 0$. As $C(\epsilon, \delta)$ is large for small δ, we conclude that optimal density may be much better then the standard one.

References

[1] Nekrutkin, V.V. and Tikhomirov, A.S. (1992). Several results on random search methods in metric spaces. Stochastic Optimization and Design, 1, N1.

[2] Loeve, M. (1963). Probability theory. 3rd ed. Van Nostrand,Priceton, N.J., 685.

[3] Kolmogorov, A.N. and Fomin, S.V. (1989). Elements of function theory and functional analysis. M., Nauka., 624. (in Russian).

Simple Genetic Algorithms for Environmental Modelling

Juhani Kettunen and Mika Jalava

A category of methods devised in the field of artificial intelligence known as genetic algorithms (GAs) is suggested for use in nonlinear estimation and experimental design. The performance of a set of GAs is analysed using test problems. Elitist selection with the single point crossover strategy is considered as the most promising combination of the genetic operators for nonlinear applications. The mutation operator is not studied systematically, but according to empirical results, the algorithms do not work without it. The results indicate that GAs are promising tools for solving nonlinear optimization problems, even when the objective function is discrete, multimodal or flat in shape. An advantage of GAs is that they facilitate heuristic sensitivity analysis simultaneously with a search for the optimum. Furthermore, they are simple and straightforward to apply.

1 Introduction

Environmental modelling is often subject to badly behaving estimation and experimental design problems. This is, especially, the case when nonlinear, mechanistic, numerically solved (partial) differential equations form the computational basis for the modelling. High parameter correlation and structural multicollinearity make the models poorly identifiable. Physically-based parameterization leads to the loss functions of estimation that are demanding, frequently discrete and nonquadratic. The problems are often ill-posed and even the necessary conditions of optimality are non-existent.

Nonlinear, mechanistic modelling is, however, one of the dominant trends in environmental management. It is therefore important to seek out feasible techniques for facilitating model-oriented data design and analysis also in these troublesome cases.

From engineering point of view, experimental design and parameter estimation contribute nonlinear programming problems that are featured by multimodal, flat and discrete goal functions and complex constraints, all characteristics that make it unreliable to apply traditional search or programming procedures. Thus, there is a practical need to develop new optimization techniques for environmental applications.

The goal of this study was to seek simple and feasible algorithms for solving the optimization problems involved in nonlinear, mechanistic modelling. Because of the encouraging results obtained in engineering, e.g. in robotics (Davidor, 1991), a genetics based approach was chosen as a starting point for this study. A set of simple algorithms was constructed and compared using test problems. The main aim was to find a feasible combination of genetic operators for the environmental management.

2 Genetic algorithms and operators

Genetic algorithms (GAs) are search techniques imitating the mechanics of natural selection and genetics. The motivation for research into GAs has been the robustness of evolutionary processes. Genetic algorithms differ from calculus-based search procedures in four ways (Goldberg 1989):

- They work with a coding of parameter sets, not the parameters themselves
- GAs search from a population of points, not a single point
- GAs make only use of information about objective function, not derivatives or other auxiliary knowledge.
- GAs use probabilistic transition rules

Among the basic elements of GAs are strings coding the information about the phenotype, i.e. the parameter or decision vector. In this study, the decision vector was treated as a table of n positive integers consisting of 32-bit positions in sequence. Each of the integers coded one element of the vector, but in the genetic operations the table was treated like a binary string of length $n \cdot 32$ bits.

Genetic optimization proceeded in steps. After initialization, three successive genetic operators selection, crossover and mutation were activated repeatedly. The sequence of the operators was as follows.

Initialization. GAs were initialized by generating a starting population of size m. In the test runs, the starting population consisted of m decision vectors with n elements chosen randomly among the points x_i of the feasible region. Before the operations, the elements were transformed into the binary space.

Selection. Two different selection strategies were considered. They are referred to in the following as elitist (EL) and roulette wheel (RO) selection. The operation of elitist selection was straightforward. The fitness of candidate-solutions was computed, candidates were ranked, and a fixed proportion of the best solutions (e.g. 10%) was selected as the genetic parents for the following generations. The roulette wheel strategy differed from the elitist model in that the parents were selected randomly by weighting their relative fitnesses. The fitness in this context meant the value of the goal function.

Crossover. A new generation, the offspring population of size m, was reproduced by mixing the binary information of the parents. Two different operators were considered. The model that is referred to below as the single point (SPO) crossover strategy operated as follows. The integer tables of parent pairs – mothers and fathers – were cut at a randomly chosen position and two offsprings were reproduced from each parent pair. The first of them inherited the first part of the mother's and the second part of the father's binary information, and the second the second part of the mother's and the first part of the father's code. Because the coding system, only one element of the binary string was completely rebuilt up of a mixture of the parents' information. The rest of the elements were present in the previous generation either in father's or mother's binary string.

The alternative crossover strategy considered here is referred as random crossover (RAN). It worked like the single point strategy but each bit of each element of the integer table given to the offspring could originate from either the mother or the father depending on the result

of the lottery. Thus, the RAN strategy introduced more diversity into the offspring than did the SPO strategy.

Mutations completed the GAs. These were motivated by the desire to increase genetic diversity, and with the aim of improving the global nature of the search. In the study mutation played a minor role. Random mutations were introduced, but their probability was kept as low as 10^{-3}. Technically, the mutations were treated like the crossover, introducing them piecewise into the code for each of the parameters.

3 Tests problems and tests

Four simple GAs consisting of different selection and crossover operators (TABLE 1) were compared, the aim being to find feasible guidelines for nonlinear modelling and programming. The test problems were selected to represent typical characteristics of the nonlinear experimental design and estimation for mechanistic modelling, namely the effects of discreteness, multimodality and flatness of the loss function. Criteria for the choice of problems were also uniqueness and ease of inference. This was obtained by choosing problems with known behaviour and extrema.

TABLE 1 The combinations of genetic operators being compared

Reproductive plan	Selection strategy	Crossover strategy
ROSPO	Roulette wheel	Single point crossover
RORAN	Roulette wheel	Random crossover
ELSPO	Elitist	Single point crossover
ELRAN	Elitist	Random crossover

The following four test problems were used to compare the algorithms:

F1:	$f_1(x_i) = \text{Min} \sum_{i=1}^{n} x_i^2$	$-5.12 \leq x_i \leq 5.12$
F2:	$f_2(x_i) = \text{Min} 100(x_1 - x_2)^2 + (1 - x_1)^2$	$2.048 \leq x_i \leq 2.048$
F3:	$f_3(x_i) = \text{Min} \sum_{i=1}^{5} \text{Int}(x_i)$	$-5.12 \leq x_i \leq 5.12$
F4:	$f_4(x) = \text{Max}(10 - 5000x)/e^{500x}$	$0.0 \leq x \leq 1$

Of these, F1, F2 and F3 were adopted from De Jong (1975, see also Goldberg, 1989) and F4 was chosen to represent a programming problem in which the optimum is located within an extremely narrow interval. F1 is a wellposed quadratic objective function, and was considered as the reference for the comparison. F2, which is the well-known Rosenbrock (1960) function, featured an optimization problem with the flat objective function. Problem F3 was included in the study because of its discreteness.

Two types of tests were performed.They are referred to in the following as *test* 1 and 2. *Test* 1 was chosen to analyse the relations between the number of generations and the population size and to study the effect of the population size on the rate of convergence. The test problems were solved by repeating each search 100 times with a starting population consisting of m randomly chosen decision vectors. Ten different population sizes (m ranging from 100 to 1000) were studied. Number of decision variables in test problem F1 was 3.

The sequence of the genetic operations was repeated until a termination rule was encountered. In this study four types of termination criteria were considered. The run was terminated, if any of the following termination rules was activated:

- No progress had taken place during the last five generations
- 10% of the fittest solutions were equal bit-for-bit
- 10% of the solutions resulted in the same value of the objective function
- When the deviation from the known true solution did not exceed a given level

Test 2 was performed to analyse the dependence of number of generations on the dimension n of the decision vector. The comparison was made only using test problem F1. The number of decision variables ranged from 3 to 15. The search was repeated 30 times with each size of the decision vector and runs were terminated when the sum of squared terms x_i^2 was less than predetermined constant: $x_i^2 < 0.01$.

4 Results

Test 1: Test problem F1 appeared to be an easy task for any of the operator combinations. If the average population size exceeded 500, all of the algorithms reached the optimum in fewer than 15 generations (Fig 1(a)). Both algorithms comprising the elitist selection strategy performed clearly better than the roulette wheel selectors. Elitist selection using the single-point crossover strategy (ELSPO) found the optimum rather easily, on average within two generations, when the population size exceeded 200 (Fig 1(a)). Also the standard error of the mean of the number of generations (Fig 1(b)) was, undoubtedly, lowest when the ELSPO strategy was applied, while the rest of the algorithms accomplished the task equally well in terms of this measure.

Similar results to those obtained in test F1 were also obtained in test F3. When the population size was more than 500-600 in this 5-dimensional search, the algorithms discovered the optimum after a reasonable number of generations.

The superior performance of elitist selection with single-point crossover, however, was even clearer in test F3 than in problem F1. Any other ranking of the algorithms was impossible in test F3. This is easily seen in Figs. 1(c) and 1(d).

Problems F2 and F4 exhibited slightly different behaviour from the problems F1 and F3. Because of the large flat domains in the objective functions, the termination criteria were barely reachable with small population sizes, and the diversity of the genetic information disappeared in such cases before acceptable discrepancies between obtained and true solutions were achieved. An increase in population, however, solved the problem. In particular, both elitist models performed extremely well when the population size reached 600 (Fig. 2). Elitist selection with single-point crossover was also unquestionably the best combination of genetic operators in these tests. Elitist selection with random crossover was equally clearly second in all tests that facilitated ranking. Roulette wheel selection did not perform very well in any of the test runs.

Test 2: The ranking of the algorithms in *test 2* reminded that of *test 1* (Fig 3). The elistist models performed better than the roulette selectors and the larger the dimension grew the

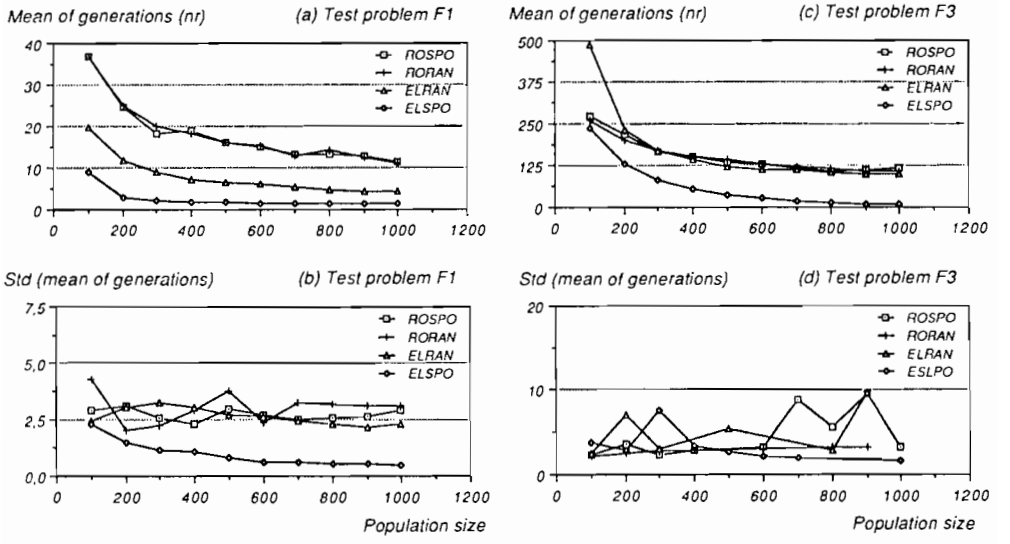


Figure 1: Mean number of generations as a function of population size and the corresponding standard error of the mean. Test problems F1 and F3.

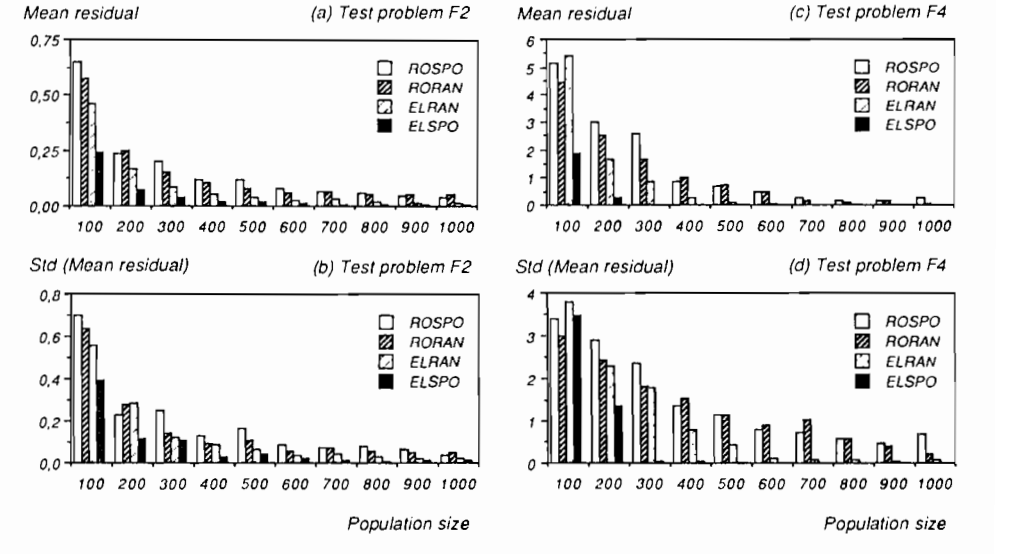


Figure 2: Discrepancy between the computed and true optimum (residual) as a function of population size, and the corresponding standard error of the mean

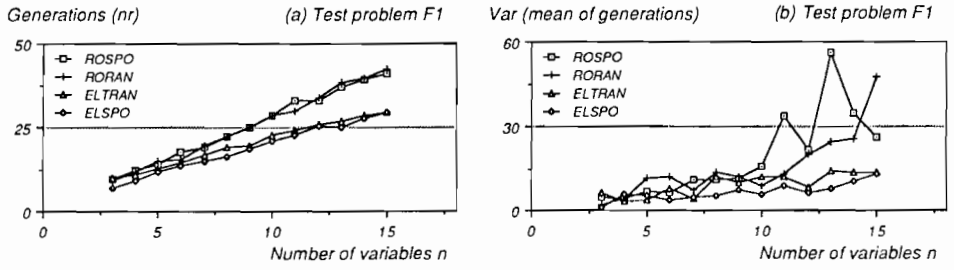


Figure 3: Mean number of generations as a function of decision vector size and the corresponding variance of the mean. Test problem F1.

clearer was the ranking. When the dimension of decision vector, however, exceeded 10 also the variation increased remarkably with roulette wheel algorithms whereas the elitist models were very stable throughout the study.The crossover strategy played minor role in the system performance.

Mutation was not studied at length, but empirical observations during the tests indicated that minor mutations close the optimum were more beneficial than any other types of mutation. It was also clear throughout the tests that the mutation operator was a necessary part of the system. If it was not applied, the termination criteria (1) and (2) became active at the very start of the search, which thus collapsed. This was the case especially with the elitist algorithms.

5 Discussion and conclusions

According to the results, the elitist selection strategy was the most competitive element in simple GAs. This result was clear in all the tests. Intuitively, it is understandable and expectable, but it might equally well reflect the fact that all the test functions, whether they were multimodal, discrete or flat, were deterministic by nature. Thus, the more randomly oriented operators, especially random crossover strategies, may have suffered from the biased choice of objective functions. From the experimental design point of view the result is not discouraging, because the designs are normally bound to deterministic formulations.

Computationally, the outcome in favour of elistist selection, is appealing because it is a much more economical selection operator than the roulette wheel strategy. For example the number or sorting it requires is only a fraction required by the RO strategy.

The flatness of goal functions should be allowed for by choosing a population size that is large enough to conserve diversity. This result, too, is intuitively clear. The proper allocation

of resources between the population size and the computer capasity required by successive generations, however, requires laborious tuning of the algorithm. As a matter of fact, even our small test exercise could have been excessive without powerful main frame facilities.

From a practical experimental design point of view the capability of simple GAs to deal with flat and discrete objective functions is the most promising result of the exercise. With nonlinear designs it is almost impossible to avoid virtually unidentifiable design problems, and thus the ability to deal with flatness combined with the fact that genetic algorithms operate on the population basis facilitate two things. Firstly, the computation is not sensitive to singularities as are traditional search algorithms. Secondly, the population basis quite probably reveals the domain of structural unidentifiability.

According to the intuitive experience of the study, the mutations are necessary operators of the simple genetic algorithms. The result might partly explain why conservative strategies as elitist selection is preferred in the tests. Possibly, in our tests, the mutations introduce the necessary diversity in the solutions rather than the crossover.

We treated the binary information of the decision vector as the binary string consisting of n successive integers, 32 bits in lenght. The system can also be made more flexible. Because of the fact that goal function evaluation and genetic operations are carried out in different spaces, the real and the binary, transformations are necessary between them. Thus, integer table may code more than n decision variables and each integer more than one element, if desired. In addition to this, the binary code of elements may be variable in length.

Computationally GAs are easily and compactly programmable. The downside of this is that the amount of computational effort tends to increase rapidly. This increase can be prevented by combining GAs with more tradional approaches, as shown by De Souza et al (1991).

Acknowledgements

The study was financed by the Laboratory of Hydrology and Water Resources Management, Helsinki University of Technology. The English was revised by Mr. Michael Garner. The authors also greatly appreciate the ideas and comments of Mr. Hannu Sirviö and the referee of the article.

References

Davidor, Y. 1991. Genetic algorithms and robotics. A heuristic strategy for optimization. World Scientific. Singapore.

De Jong, K.A. 1975. An analysis of the behaviour of a class of genetic adaptive systems. Doctoral dissertation. University of Michigan.

De Souza, P.S.&Talukdar, S.N. 1991. Genetic algorithms in asynchronous teams. In: Belew, R.K&Booker, L.B. (Eds.): Proceedings of the fourth international conference on Genetic Algorithms. Morgan Kaufmann Publ. San Mateo, CA.

Goldberg, D, E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. AddisonWesley. 412p.

Rosenbrock, H.H. 1960. An Automatic Method for Finding the Greatest and Least Value of a Function. Computer Journal, Vol.3:175.

Covering Based on a Stratified Sample

Maxim V. Chekmasov and Marina V. Kondratovich

Zhigljavsky [4] stated the problem of improving the upper estimates for the densities of coverings of the space. He proposed to use different kinds of dependent samples which generate coverings by the standard rule, that use the independent sample, described by Rogers [1]. In this paper we demonstrate, that the use of stratified sample is preferable to the independent one. Therefore, we confirm that this is a promising subject for study.

Many algorithms of global random search consist of a series of iterations. At any iteration the random points are generated according to some distribution. The common way to obtain the random points is independent random choice of them. This is the same as to use the independent sample. The classical method to decrease the variance of the Monte-Carlo estimations for integrals is the use of the stratified sample. The gist of stratified sample is a division of a region into some number of disjoint subsets and independent generation of random points in each of them.

In paper [3] it is proved that the stratified sample dominate independent one according to a number of criteria natural for global random search. In paper [2] it is shown that the stratified sample is admissible in the set of continuous functions. It is shown in this paper that stratified sample-based covering is more economical than independent sample-based one. The reader can also consult [4] for the full explanation of these results.

Let K be a bounded set of positive measure $\mu(K)$ and T be the numerable system $\{K+a_i\}$ of traslations of K by vectors $a_1, a_2, \ldots$ Generally, system T does not form covering for the whole space. Let us clear what part of the space is covered by the elements of T. According to [3], the relative measure of the part of space uncovered by the elements of T is defined as follows.

For each cube C with sides parallel to the axes and of length $s(C)$ define

$$\sigma_+(T,C) = \frac{1}{\mu(C)}\Sigma\mu(K+a_i), \text{ where } K+a_i \cap C \neq \emptyset;$$

$$\sigma_-(T,C) = \frac{1}{\mu(C)}\Sigma\mu(K+a_i), \text{ where } K+a_i \text{ a subset of } C;$$

$$\Phi(T,C) = 1 - \mu(\bigcup_{i=1}^{\infty}(K+a_i)\cap C)/\mu(C).$$

Here $\Phi(T,C)$ is the relative measure of the part of the cube uncovered by the sets of T. Further,

$$\sigma_+(T) = \overline{\lim_{s(c)\to+\infty}}\ \sigma_+(T,C)$$

$$\sigma_-(T) = \underline{\lim_{s(c)\to+\infty}}\ \sigma_-(T,C)$$

$$\Phi_+(T) = \overline{\lim_{s(c)\to+\infty}} \Phi(T,C)$$

Small values of $\Phi_+(T)$ signify that the sets of T cover the most part of the space.

Let C be the cube defined by the inequalities $0 \le x_i \le s(C)$, $i = 1,...,n$. Let σ be a positive real number, $\mu(K)/\mu(C) = \sigma/N$, where N is an integer. Let $b_1, b_2, ...$ is grid of points with coordinates being integer multiple of $s(C)$, $x_1, x_2, ..., x_N$ is system of N points lying in C. Then according to [1] (theorem 1.5) for the system $T = T(x_1, x_2, ..., x_N) = \{K + x_i + b_j\}$ $(i = 1,2,...,N,\ j = 1,2,...)$ we have $\sigma_+(T) = \sigma_-(T) = \sigma$.

The system of points $x_1, x_2, ..., x_N$ can be chosen in various manners. In [1] the points $x_1, x_2, ..., x_N$ form the independent sample, i.e. x_i are independent realizations of the random variable uniformly distributed in the cube C. In [3] it is shown that for the independent sample

$$\Phi_+^{ind} = E\Phi_+(T(x_1, x_2, ..., x_N)) = (1 - \sigma/N)^N.$$

Let the points $x_1, x_2, ..., x_N$ form the stratified sample, i.e. $C = \bigcup_{i=1}^{N} C_i$ where $\mu(C_i) = 1/N$, $\mu(C_i \cap C_j) = 0$, $i \ne j$, x_i $(i = 1, ..., N)$ the realizations of the uniform random vector in C_i : $P_i(A) = NP(A \cap C_i)$. Denote

$$\Phi_+^{str} = E\Phi_+(T(x_1, x_2, ..., x_N)).$$

Theorem. For all bounded sets K and positive real number σ $\Phi_+^{str} \le \Phi_+^{ind}$ and there exist the sets K and real positive number σ for which $\Phi_+^{str} < \Phi_+^{ind}$.

Proof. Let $\tau(x)$ be the indicator function of K. Let K be a subset of a cube of the side of length $s(C)$. Let us suppose (without loss of of generallity) $s(C) > 2s(K)$ then the sets $K + x_i + b_j$ and $K + x_i + b_k$ with $j \ne k$ have no common points and the indicator function of the set $\bigcup_{j=1}^{\infty} (K + x_i + b_j)$ is equal to $\sum_{j=1}^{\infty} \tau(x - x_i - b_j)$.

Therefore the indicator function of the set E, composed by the points not belonging to any set of system T, is equal to

$$\sigma(x) = \prod_{i=1}^{N}(1 - \sum \tau(x - x_i - b_j)).$$

Then for all cubes G with sides parallel to the coordinate axes, we have

$$\sigma(T,G) = \frac{1}{\mu(G)} \int_G \sigma(x)dx.$$

Because the integrated function is periodic for all coordinates with period $s(C)$, we have

$$\sigma_+(T(x_1, ..., x_N)) = \overline{\lim}\sigma_+(T,G) = \frac{1}{\mu(G)} \int_G \sigma(x)dx.$$

Let us find the value $\sigma_+(T(x_1, ..., x_N))$ averaged on every possible stratified sample $x_1, ..., x_N$.

$$\Phi_+^{str} = E\sigma_+(T(x_1, ..., x_N)) = \int_{C_1} \frac{N}{\mu(C)} dx_1 ... \int_{C_N} \frac{N}{\mu(C)} \sigma_+(T(x_1, ..., x_N))dx_N$$

$$= \frac{N^N}{\mu(C)^{N+1}} \int_{C_1} dx_1 ... \int_{C_N} dx_N \int_C \prod_{i=1}^{N} (1 - \sum_{j=1}^{\infty} \tau(x - x_i - b_j)) dx$$

$$= \frac{N^N}{\mu(C)^{N+1}} \int_C dx \left(\prod_{i=1}^{N} \int_{C_i} (1 - \sum_{j=1}^{\infty} \tau(x - x_i - b_j)) dx_i \right)$$

$$= \frac{N^N}{\mu(C)^{N+1}} \int_C dx \prod_{i=1}^{N} \left(\mu(C_i) - \sum_{j=1}^{\infty} \int_{C_i} \tau(x - x_i - b_j) dx_i \right)$$

$$= \frac{N^N}{\mu(C)^{N+1}} \int_C dx \prod_{i=1}^{N} \left(\frac{\mu(C)}{N} - \mu(K \cap C_i) \right)$$

$$= \frac{N^N}{\mu(C)^N} \prod_{i=1}^{N} \left(\frac{\mu(C)}{N} - \mu(K \cap C_i) \right) = \prod_{i=1}^{N} (1 - \mu(K \cap C_i)/\mu(C_i))$$

Let us find the maximal value of expression $\prod_{i=1}^{N} (1 - N \frac{\mu(K \cap C_i)}{\mu(C)})$ under the condition that

$$\mu(K \cap C_1)/\mu(C) + \mu(K \cap C_2)/\mu(C) + ... + \mu(K \cap C_N)/\mu(C) = \sigma/N.$$

Using Lagrange factor's method, it is easy to prove that the maximal value is reached at

$$\mu(K \cap C_1)/\mu(C) = ... = \mu(K \cap C_N)/\mu(C) = \sigma/N^2$$

and equals to

$$\prod_{i=1}^{N} (1 - N\sigma/N^2) = (1 - \sigma/N)N$$

Thus, it is proved that $\Phi_+^{str} \leq (1-\sigma/N)^N = \Phi_+^{ind}$. We have $\mu(K)/\mu(C) = \sigma/N$, $\mu(C_i)/\mu(C) = 1/N$, i.e. $\mu(K)/\mu(C_i) = \sigma$.

The value of function Φ_+^{str} is determined by the set K, and number σ - relative measure of the set K and the partition. Evidently, there exist such sets K that only $l < N$ values of $\mu(K \cap C_i)$ are nonzero, i.e. $\Phi_+^{str} < \Phi_+^{ind}$. This ends the proof.

The advantage of using the stratified sample is greater when possible number of intersections between the set K and sets C_i decreases.

Let us consider the case of covering the plane ($n = 2$) by sets K - circles; the sets C_i are squares of measure $\mu(C_i) = 1/N$. Then, for average value of relative measure Φ_+^{str}, the following cases are possible: (i) $0 < \sigma \leq \pi$, i.e. the circle radius is less than square's C_i side. Only four values are non-zero: $\mu(K \cap C_i)/\mu(C_i) = (\frac{1}{4}\sigma/N)/(1/N)$. Therefore $\Phi_+^{str} = (1-\sigma/4)4$

(ii) $\pi < \sigma \leq 2\pi$, i.e. the circle radius is greater than square C_i side but less than its diagonal. In this case 12 values of $\mu(K \cap C_i)/\mu(C_i)$ are non-zero. After corresponding calculations the following value can be obtained:

$$\Phi_+^{str} = (1 + (\sigma \cos^{-1} \sqrt{\pi/\sigma})/\pi - \sigma/4 - \sqrt{\sigma/\pi - 1})^4 \cdot (1 - (\sigma \cos^{-1} \sqrt{\pi/\sigma})/(2\pi))^8.$$

(iii) If $\sigma > 2\pi$, there exist such i that $\mu(K \cap C_i) = \mu(C_i)$ and then $\Phi_+^{str} = 0$.

There are cited below the results of computer experiment for estimating relative measure of uncovering plane ($n = 2$) by sets of K circles.

σ	N	stratified	independent
0.2	16	0.800	0.816
	64	0.806	0.821
0.6	25	0.486	0.563
	64	0.497	0.569
2	16	0.051	0.135
	64	0.060	0.173

Similarly, it is possible to obtain an expression for Φ_+^{str} in the case $n > 2$ and K is sphere.

References

[1] Rogers C.A. Packing and covering, Ch.3., Cambridge University Press, 1964.

[2] Ermakov S.M., Zhigljavsky A.A., Kondratovich M.V. Reduction of the problem of random estimating for an extremum of function. Doclady AN SSSR (Reports of the Soviet Academy of Sciences), 1988, vol.302, No.4, p.796-8.

[3] Ermakov S.M., Zhigljavsky A.A., Kondratovich M.V. Comparison of some random search for global extremum procedures. USSR J. Comp. Math. and Math. Phys., 1989, vol.29, No.2, p.163-70.

[4] Zhigljavsky A.A. Global Random Search Kluwer Academic Publishers, 1991.

On Average-Optimal Quasi-Symmetrical Univariate Optimization Algorithms

Luc Pronzato and Anatoly A. Zhigljavsky

1 Introduction

This paper deals with the problem of minimizing the length of an interval containing the scalar argument x^* at which a function $f \in \mathcal{F}$ reaches its minimum value. The class $\mathcal{F}$ considered corresponds to inverse-unimodal functions over a given initial interval I_0, i.e.

$f : I_0 \to \mathcal{R}$

$\exists x^* \in I_0 \mid f$ strictly decreasing for $x \leq x^*$ and strictly increasing for $x > x^*$,

or else strictly decreasing for $x < x^*$ and strictly increasing for $x \geq x^*$.

The restriction to functions symmetrical at x^* will be used in Section 3, but no further assumption on f (such as regularity) will be considered. Our aim is to determine an interval I_N, after N evaluations of f in I_0, which is guaranteed to contain $x^* = \arg\min_{x \in I_0} f(x)$ and which is of minimal length. We assume that the evaluations of f are performed without errors. The result of an evaluation will be called an observation. Let $\mathcal{I}_k$ denote the information obtained after k evaluations of f at the points $x_1, x_2, \ldots, x_k$, i.e. $\mathcal{I}_k = \{a, b, x_1, x_2, \ldots, x_k, f(x_1), f(x_2), \ldots, f(x_k)\}$, with $\mathcal{I}_0 = \{a, b\}$ ($I_0 = [a, b]$). The information $\mathcal{I}_k$ is used to define I_{k+1} which is guaranteed to contain x^*, according to the rule

$$I_{k+1} = [l(\mathcal{I}_k), u(\mathcal{I}_k)] .$$

This is done by removing from I_0 the parts that are not consistent with the observations, i.e. where we are sure that x^* cannot lie (see e.g. 4 for a precise definition of consistency in search problems). We consider nonrandomized *sequential* procedures, i.e. procedures for which the points where f is evaluated are chosen sequentially on the basis of the information collected so far, what we denote by

$$x_{k+1} = g_{k+1}(\mathcal{I}_k) ,$$

with g_{k+1} a deterministic function. Note that the rule for constructing I_k is independent of k, contrary to that for constructing x_{k+1}. The total number N of evaluations that is allowed is fixed in advance. A *strategy* S_N is thus defined by

$$S_N = \{l, u, g_k, k = 1, \ldots, N\} .$$

$\mathcal{S}^N$ will denote the set of all nonrandomized strategies S_N that possess the consistency property $x^* \in I_N$.

We are interested in the determination of the best strategy, in some sense connected with the length L_N of the final interval. The paper will show that classical approaches (based on

worst-case performances) are not optimal on average (which may be more reasonable from a practical point of view, as suggested in 5). Although the problem considered here is mainly theoretical (golden search and parabolic approximation are the approaches classically used in practice to solve univariate optimization problems), it provides an interesting test-case for average-optimal procedures. *Minimax optimality* is considered in Section 2. Restricting the class $\mathcal{F}$ to functions symmetrical at x^*, and introducing a noninformative prior distribution on the location of x^*, we consider an *average-optimal* approach in Section 3. Worst-case and average performances are compared in Section 4.

2 Minimax optimality

A strategy S_N^* will be said ϵ-minimax if it satisfies

$$\sup_{f \in \mathcal{F}} L_N(f, S_N^*) \leq \inf_{S \in \mathcal{S}^N} \sup_{f \in \mathcal{F}} L_N(f, S) + \epsilon .$$

The problem of existence of S_N^* is solved in 2, 3, and we shall only give here a summary of the results. This will be useful for the definition of an average-optimal procedure in Section 3.

Two evaluations of f inside I_k are required in order to eliminate any part of the interval, as illustrated by Figure 1. For that reason the procedure is called of second order, in opposition to first-order searchs (such as the determination of the root of a monotonous function over an interval, see 3).

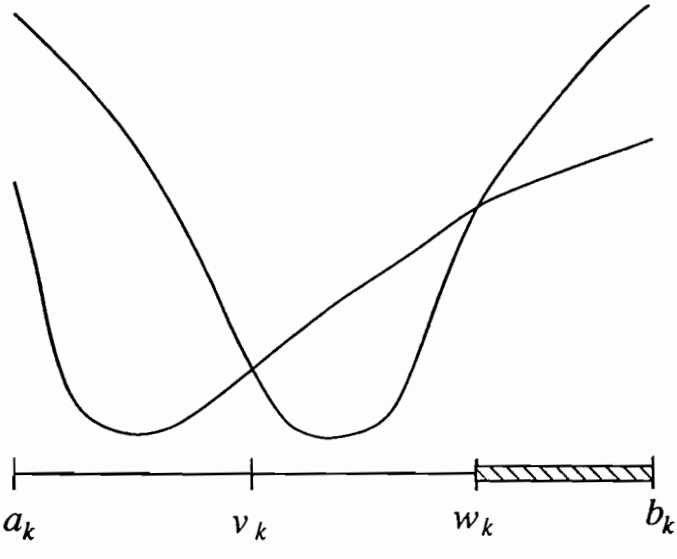


Figure 1: Two evaluations inside I_k are needed to eliminate a part of it.

Let v_k, w_k denote the two points in the interior of $I_k = [a_k, b_k]$ where f is evaluated, with $v_k < w_k$. The rules l, u of the strategy are defined by

$$I(k+1) = \begin{cases} [a_k, w_k] & \text{if } f(v_k) < f(w_k), \\ [v_k, b_k] & \text{otherwise.} \end{cases} \tag{1}$$

Since one observation does not permit to eliminate anything, we have $L_1 = L_0$. Note that either v_k or w_k still belongs to I_{k+1}, so that a single new observation is required at each step to shorten the interval.

The determination of S_N^* can be considered as a *dynamic programming problem* 1, with terminal cost $\sup_{f \in \mathcal{F}} L_N(f, S)$. The forward-in-time (i.e. forward-in-k) problem then corresponds to choosing the sequence of points $x_1, x_2, \ldots, x_N$. As usual, the optimal solution is obtained with a backward-in-time approach. Given $I_{N-1} = [a_{N-1}, b_{N-1}]$ (i.e. once $N-1$ evaluations of f in I_0 have been performed), the optimal location of the last two points is $x_{N-1}^* = \frac{1}{2}(a_{N-1} + b_{N-1})$, $x_N^* = x_{N-1} + \epsilon$, which gives $\sup_{f \in \mathcal{F}} L_N = \frac{1}{2} L_{N-1} + \epsilon$. Note that ϵ must be chosen as small as possible, with the constraint $x_N^* \neq x_{N-1}^*$. It is therefore only related to the precision of the calculations, and we shall assume $\epsilon = 0$ in what follows . Define new normalized coordinates z_k as

$$z_k = \frac{x_k - a_k}{b_k - a_k},$$

where x_k belongs to the k-th interval $I_k = [a_k, b_k]$, so that $z_k \in [0,1], \forall k$. The minimax strategy S_2^* for $N = 2$ is thus defined by $z_1^* = \frac{1}{2}$, $z_2^* = \frac{1}{2}$. Moving backward-in-time, one can show that the minimax optimization procedure consists in successively splitting the remaining interval into equal parts. One thus has, in an obvious manner,

$$L_{k-1}^* = L_k^* + L_{k+1}^*, \tag{2}$$

and

$$z_{k-1}^* = \frac{L_{k-1}^* - L_{k+1}^*}{L_{k-1}^*} = \frac{L_k^*}{L_{k-1}^*}, \; k \geq 1,$$

(see 2 for a detailed proof of the minimax optimality of this procedure). Let $\tilde{}$ denote the backward-ordering operator, i.e.

$$\tilde{z}_k = z_{N-k+1},$$

where k corresponds to the number of function evaluations still allowed ($\tilde{z}_N$ and $\tilde{z}_1$ then respectively correspond to the first and the last one). One has $\tilde{z}_k^* = \frac{L_{N-k+2}^*}{L_{N-k+1}^*}$, $k \geq 2$, with $\tilde{z}_1^* = \tilde{z}_2^* = \frac{1}{2}$. Using the recurrence equation (2), one can calculate $L_{N-k}^*, k = 2, \ldots, N-1$, and $\tilde{z}_N^*$ gives the location of the initial evaluation. The successive locations of the evaluations are then given by the $\tilde{z}_k^*$'s. They correspond to a *symmetrical algorithm*, i.e. each new point is chosen symmetrical of the point already present in the interval with respect to the midpoint of the interval. This gives $\tilde{z}_{k-1}^* = (1 - \tilde{z}_k^*)/\tilde{z}_k^*$ if $\tilde{z}_k^* \in [\frac{1}{2}, 1]$, and $\tilde{z}_{k-1}^* = 1 - \tilde{z}_k^*/(1 - \tilde{z}_k^*)$ otherwise. Taking $\tilde{z}_k^*$ in $[\frac{1}{2}, 1]$, one has: $L_{N-1}^* = 2L_N^*, L_{N-2}^* = 3L_N^*, L_{N-3}^* = 5L_N^*, L_{N-4}^* = 8L_N^*, \ldots, \tilde{z}_1^* = \tilde{z}_2^* = \frac{1}{2}, \tilde{z}_3^* = \frac{2}{3}, \tilde{z}_4^* = \frac{3}{5}, \tilde{z}_5^* = \frac{5}{8}, \ldots$

Remark 1

(i) This procedure is often called Fibonacci algorithm, *due to the fact that* $L_{N-k}^* = U_{k+2} L_N^*$, *with* $(U_k)_k$ *the sequence of Fibonacci numbers,* $U_0 = 0, U_1 = 1, U_k = U_{k-1} + U_{k-2}$, $k \geq 2$.

(ii) Although we do not want to discuss the admissibility problem here, note that the procedure is inadmissible, since it can be improved by a reduction of ϵ or by using the fact that whenever the two smallest values of the function in the interval are equal, x^* must lie in the interval defined by the two arguments 2, 3. However, the second improvement corresponds to a very unlikely situation, while the first one is almost negligible.

3 Average optimality

The class $\mathcal{F}$ is now restricted to functions symmetrical at x^*, and we give a noninformative uniform prior distribution π_0 to x^* over I_0. The shape of the distribution is not modified when parts of I_0 are eliminated, so that the distribution of x^* remains uniform. If the symmetry assumption about f is completely taken into account when propagating the distribution π_k of x^*, π_k does not spread over the whole interval I_k, $k > 0$. However, bearing in mind the minimax case, in what follows we consider π_k as uniform over I_k. A strategy S_N^{**} will then be said average-optimal if it satisfies

$$E_{x^*}\{L_N(f, S_N^{**})\} \leq \inf_{S \in \mathcal{S}^N} E_{x^*}\{L_N(f, S)\}\,.$$

The class $\mathcal{S}^N$ of the strategies considered will be further restricted to quasi-symmetrical procedures.

Using average performances as a criterion for choosing the procedure requires being able to evaluate the expected length of an interval. This can be done easily using previous assumptions. Consider for example an interior point x of $I_k = [a_k, b_k]$. From the symmetry assumption about f, one has

$$\text{Prob}\{f(x) > f(a_k)\} = \text{Prob}\{x^* < \frac{x + a_k}{2}\}\,,$$

which gives, from the uniform distribution of x^*,

$$\text{Prob}\{f(x) > f(a_k)\} = \frac{\frac{1}{2}(x + a_k) - a_k}{b_k - a_k}\,. \tag{3}$$

The determination of S_N^{**} is considered again as a dynamical programming problem to be solved with the Bellman optimality principle. Moving backward in time, we first consider the last step, i.e. the situation where $I_{N-1} = [a_{N-1}, b_{N-1}]$ is given, with an evaluation of f already performed at some $x'_j \in]a_{N-1}, b_{N-1}[, j \leq N-1$. Using (3) and (1), we can evaluate $E_{x^*}\{L_N\}$ as a function of x'_j and x_N. Minimizing it with respect to x_N, we obtain x_N as a function of x'_j, as illustrated by Figure 2. Using normalized coordinates, one gets

$$z_N = \begin{cases} \frac{1}{2} - z'_j & \text{if } 0 \leq z'_j \leq \frac{1}{4}, \\ z'_j & \text{if } \frac{1}{4} \leq z'_j \leq \frac{3}{4}, \\ \frac{3}{2} - z'_j & \text{if } \frac{3}{4} \leq z'_j \leq 1. \end{cases}$$

$E_{x^*}\{L_N\}$ can then be calculated as a function of x_{N-1}, which gives

$$E_{x^*}\{L_N\} = L_{N-1}\,\rho_{N-1}\left(\frac{x_{N-1} - a_{N-1}}{b_{N-1} - a_{N-1}}\right)\,,$$

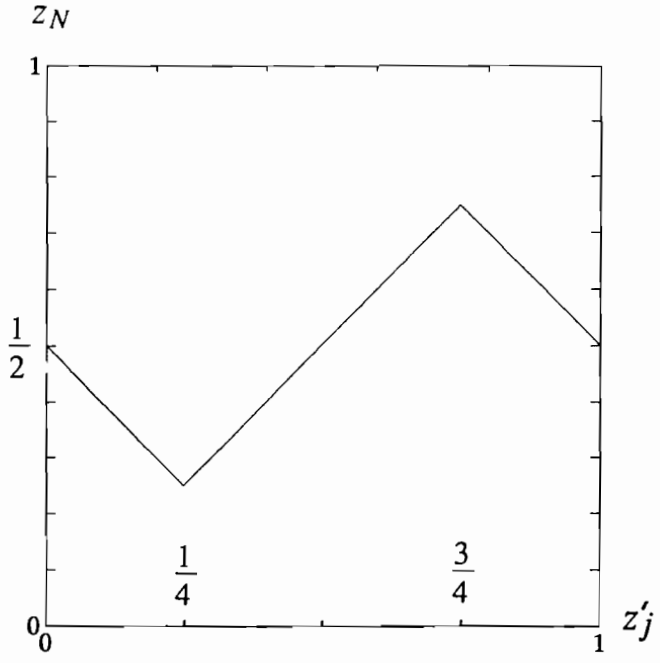


Figure 2: Choice of the last point (z_N in normalized coordinates) as a function of that already present in the interval (z'_j in normalized coordinates).

with

$$\rho_{N-1}(z) = \begin{cases} \frac{7}{8} - z & \text{if } 0 \le z \le \frac{1}{4}, \\ 2z^2 - 2z + 1 & \text{if } \frac{1}{4} \le z \le \frac{3}{4}, \\ z - \frac{1}{8} & \text{if } \frac{3}{4} \le z \le 1. \end{cases} \tag{4}$$

When $N = 2$, this obviously gives $z_1^{**} = z_2^{**} = \frac{1}{2}$ for the optimal strategy (see Figure 4 for a plot of $\rho_{N-1}(z)$).

Remark 2 *A practical implementation of these ideas would require the introduction of a value $\epsilon \neq 0$ in order to obtain $x_2^{**} = x_1^{**} + \epsilon \neq x_1^{**}$, so that, as in the minimax case, the notion of ϵ-optimality might be considered here.*

Suppose now that three observations are to be performed. $E_{x^*}\{L_3\}$ can be calculated as a function of x_1 and x_2 when x_3 is chosen as indicated by Figure 2. A contour plot is given in Figure 3. The function is symmetrical with respect to the diagonal $x_2 = x_1$. Symmetrical procedures are obtained along the line $x_2 = 1 - x_1$. The optimal choice for (x_1, x_2) is given by $(0.6325, 0.3675)$ or its symmetrical $(0.3675, 0.6325)$. These point are indicated by stars on the figure. They satisfy $x_2^{**} = 1 - x_1^{**}$. Following the optimal choice indicated by Figure 2, x_3^{**} is equal to x_2^{**}.

The expression of $E_{x^*}\{L_3\}$ could theoretically be used to express the optimal value of x_2 as a function of x_1, leading one step backward to the case where four observations are to be performed. However, analytical expressions then become intractable, and in what follows we

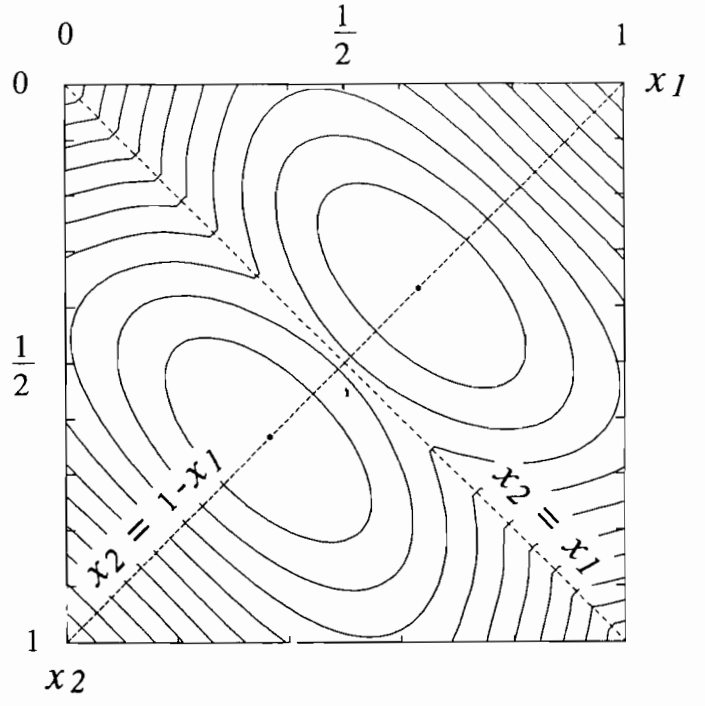


Figure 3: Contour plot of $E_{x^*}\{L_3\}$ as a function of x_1 and x_2 (x_3 is chosen according to Figure 2 and $I_0 = I_1 = [0,1]$).

shall restrict our attention to suboptimal *symmetrical algorithms* (quasi-symmetrical, to be more precise, since the last point x_N is chosen according to Figure 2). One can then easily prove by induction that

$$E_{x^*}\{L_N\} = L_k \rho_k(z_k), k \leq N-2,$$

with

$$\rho_k(z) = \begin{cases} z\rho_{k+1}\left(\frac{1-z}{z}\right) & \text{if } \frac{1}{2} \leq z \leq 1, \\ (1-z)\rho_{k+1}\left(\frac{z}{1-z}\right) & \text{if } 0 \leq z \leq \frac{1}{2}, \end{cases}$$

and ρ_{N-1} defined by (4). Figure 4 presents the functions $\rho_{N-1}, \rho_{N-2}, \ldots, \rho_{N-6}$.

Consider now the determination of the minima $z^{**}_{N-k} = \arg\min_{z\in[\frac{1}{2},1]} \rho_{N-k}(z)$, $k = 1, \ldots$. This will ultimately define the average-optimal quasi-symmetrical strategy as follows. We use again the backward ordering operator $\tilde{}$, so that $\tilde{z}^{**}_k = z^{**}_{N-k+1}$ and $\tilde{\rho}_k(z) = \rho_{N-k+1}(z)$, with k the number of evaluations allowed. The optimal initial point for these evaluations is then given by $\tilde{z}^{**}_k$, and the successive points are chosen symmetrically with respect to the midpoint of the interval, except for the last one.

The functions $\tilde{\rho}_k(z)$ have several local minima in $[\frac{1}{2}, 1]$ (see Figure 4), and we are interested in the global ones $\tilde{z}^{**}_k, k \geq 3$. One can easily show by induction that in the neighborhood of $\tilde{z}^{**}_k$ the function $\tilde{\rho}_k(z)$ can be written as

$$\tilde{\rho}_k(z) = \frac{\alpha_k z^2 + \beta_k z + \gamma_k}{\delta_k z + \epsilon_k}, k \geq 3.$$

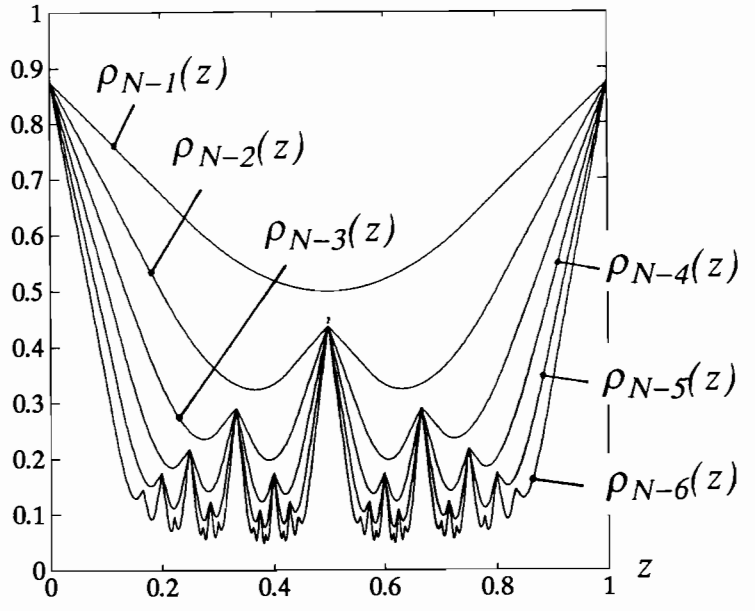


Figure 4: Expected final length $E_{x^*}\{L_k\}$ as a function of the location of the initial point $(E_{x^*}\{L_k\} = L_0\rho_{N-k+1})$.

Similarly, one gets around $\tilde{z}_{k+1}^{**}$

$$\tilde{\rho}_{k+1}(z) = z\frac{\alpha_k\left(\frac{1-z}{z}\right)^2 + \beta_k\frac{1-z}{z} + \gamma_k}{\delta_k\frac{1-z}{z} + \epsilon_k}, k \geq 3\,,$$

which defines a recurrence equation for $\theta_k = (\alpha_k, \beta_k, \gamma_k, \delta_k, \epsilon_k)^T$,

$$\theta_{k+1} = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}\theta_k\,. \tag{5}$$

One can also prove by induction that $\tilde{z}_k^{**}$ is given by

$$\tilde{z}_k^{**} = -\frac{\epsilon_k}{\delta_k} + \frac{\sqrt{2}}{\delta_k\sqrt{\alpha_k}}\,. \tag{6}$$

Table 1 gives the successive values of θ_k and $\tilde{z}_k^{**}$, $k = 1, \ldots, 10$.

k	α_k	β_k	γ_k	δ_k	ϵ_k	$\tilde{z}_k^{**} = z_{N-k+1}^{**}$
1						0.5
2						0.5
3	5	-6	2	1	0	0.6325
4	13	-16	5	-1	1	0.6078
5	34	-42	13	2	-1	0.6213
6	89	-110	34	-3	2	0.6167
7	233	-288	89	5	-3	0.6185
8	610	-734	233	-8	5	0.6178
9	1597	-1974	610	13	-8	0.6181
10	4181	-5168	1597	-21	13	0.6180

TABLE 1: Initial point $\tilde{z}_k^{**}$ for S_k^{**} when k evaluations are allowed.

Remark 3

(i) One can easily check that $\lim_{k\to\infty} \tilde{z}_k^{**} = \alpha = \frac{1}{2}(\sqrt{5}-1) = 0.6180\ldots$, *which corresponds to the Golden-Section algorithm suggested in 2, 3. The term* $-\epsilon_k/\delta_k$ *in (6) corresponds to the sequence of the minimax-optimal algorithm, i.e. more precisely* $-\epsilon_k/\delta_k = \tilde{z}_{k-3}^{*}$, *and one has again* $\lim_{k\to\infty} \tilde{z}_k^{*} = \alpha$. *The eigenvalues of the matrix defining the recurrence equation (5) are given by* $-1, 2+\alpha, 1-\alpha, -1-\alpha, \alpha$, *which provides another indication that the number* α *plays here a special role.*

(ii) Contrary to what occurs for minimax optimality, the values of $\tilde{z}_k$ *corresponding to the use of the symmetrical algorithm initialized in* $\tilde{z}_N^{**}$ *do not coincide with the* $\tilde{z}_k^{**}$*'s.*

(iii) This procedure is not optimal for $N > 3$ *if the restriction to symmetrical procedures is removed. Consider e.g. the case* $N = 4$, $I_0 = [0,1]$, *with* $\tilde{z}_4^{**} = 0.6078$. *Take* $\tilde{z}_3$ *symmetrically, so that* $\tilde{z}_3 = (1 - \tilde{z}_4^{**})/\tilde{z}_4^{**} = 0.6454$. *Three evaluations have to be performed in* $[0, \tilde{z}_4^{**}]$, *with* $\tilde{z}_3$ *already fixed (or equivalently in* $[0,1]$ *with* $\tilde{z}_3$ *already fixed). One can easily check that the optimal choice for* $\tilde{z}_2$ *(with* $\tilde{z}_1$ *obtained from* $\tilde{z}_2$ *according to Figure 2) does not coincide with* $(1 - \tilde{z}_3^{**})/\tilde{z}_3^{**}$.

(iv) Remark 1 (ii) about admissibility applies again here: the second improvement suggested could be implemented (although it would be useful only in very unlikely situations).

4 Comparison between minimax and average optimality

One can compare on the same plot the function ρ_{N-k} (which gives the expected final length for the quasi-symmetrical average-optimal procedure) with the final length for the minimax procedure, both expressed as functions of the initial point where f is evaluated (see Figure 5). The new procedure is seen to perform slightly better than the minimax procedure in the average sense (it is the least one could ask for!).

The expected length $E_{x^*}\{L_N\}$ can be calculated analytically. Assume $L_0 = 1$ and start from $\tilde{z}_N^{**}$ given by (6). Then $\tilde{x}_{N-1} = 1 - \tilde{z}_N^{**} = \frac{\epsilon_N}{\delta_N}\frac{\epsilon_{N-1}}{\delta_{N-1}} - \frac{\sqrt{2}}{\delta_N\sqrt{\alpha_N}}$, and $\tilde{x}_{N-2} = \tilde{z}_N^{**} - \tilde{x}_{N-1}$. It can be shown by induction that

$$\tilde{x}_{N-k} = \left|\frac{\epsilon_{N-k}}{\delta_N}\right| + \frac{\sqrt{2}}{\delta_N\sqrt{\alpha_N}}\delta_{k+3}$$

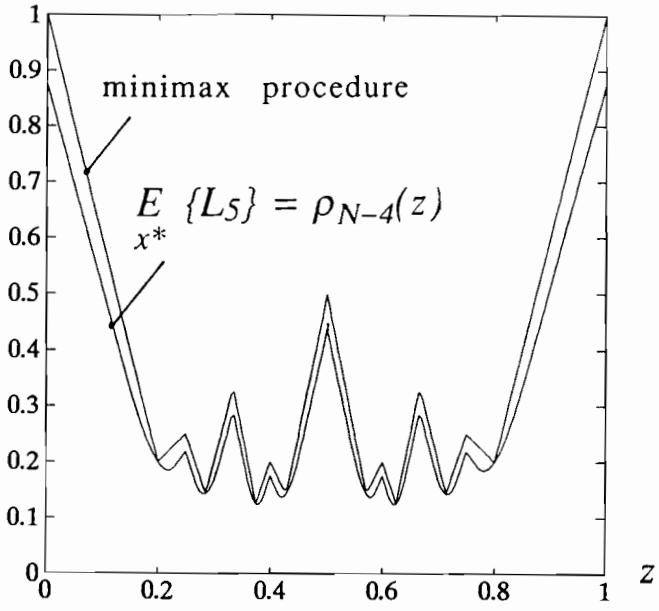


Figure 5: Comparison between the expected final length for the quasi-symmetrical average-optimal procedure and the final length for the minimax procedure, both as functions of the location of the initial point (f is evaluated 5 times, $L_0 = 1$).

which yields $\tilde{x}_3 = \frac{\sqrt{2}}{\sqrt{\alpha_N}}$ and $\tilde{x}_4 = \frac{1}{|\delta_N|} + \frac{\sqrt{2}}{\delta_N\sqrt{\alpha_N}}\delta_{N-1}$. This gives $\tilde{x}_2 = \tilde{x}_4 - \tilde{x}_3 = \frac{1}{|\delta_N|} + \frac{\sqrt{2}}{\sqrt{\alpha_N}}\frac{\delta_{N+1}}{\delta_N}$, and $E_{x^*}\{L_N\}$ is known as function of $\tilde{x}_2$. From (4) one gets $E_{x^*}\{L_N^{**}\} = 2\tilde{x}_2^2/\tilde{x}_3 - 2\tilde{x}_2 + \tilde{x}_3$, and the length in the worst case is $\max_{f\in\mathcal{F}} L_N^{**} = \tilde{x}_2$.

Figure 6 presents the relative decrease in expected length $(L_N^* - E_{x^*}\{L_N^{**}\})/L_N^*$, when the average-optimal procedure (quasi-symmetrical algorithm) is used instead of the minimax one (Fibonacci algorithm). As could be expected from Figure 5, the gain in performances is only marginal. Also indicated on Figure 6 is the relative increase in maximal length (i.e. in the worst case concerning the location of x^*) $(\max_{f\in\mathcal{F}} L_N^{**} - L_N^*)/L_N^*$, which is quite larger.

The price paid for a slight improvement in the average sense is thus an important deterioration in minimax performances. However, worst-case performances could correspond to a rather poor criterion in practice if, as suggested in 5, they are obtained for atypical values of x^* that belong to a subset of I_0 with zero measure. Moreover, the use of a symmetrical procedure certainly puts a curb on the performances in the average sense, and other non-symmetrical procedures are under investigation.

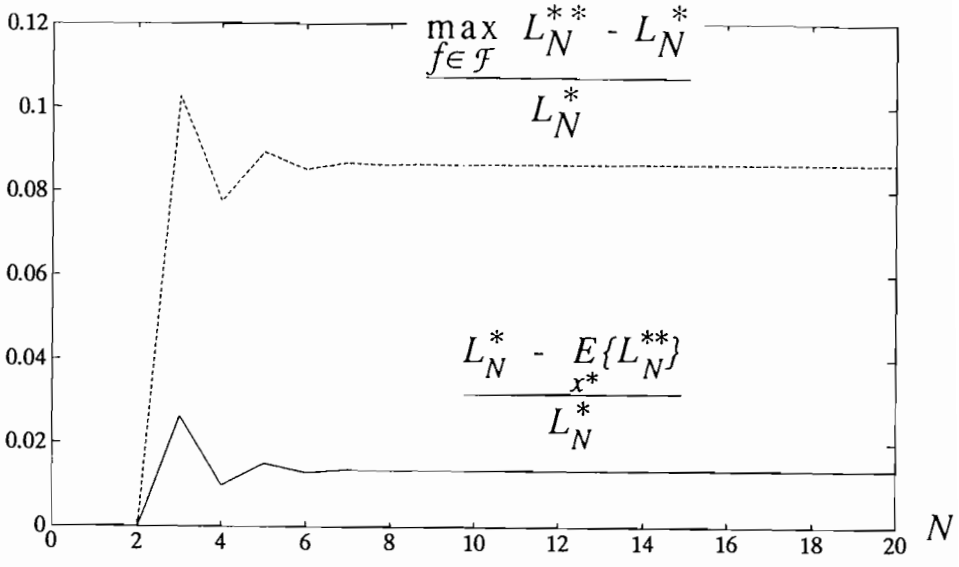


Figure 6: Relative decrease in expected length and increase in maximal length when S_N^{**} is used instead of S_N^*.

References

[1] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, N.J., 1957.

[2] J. Kiefer. Sequential minimax search for a maximum. *Proc. Am. Math. Soc.*, 4:502–506, 1953.

[3] J. Kiefer. Optimum sequential search and approximation methods under minimum regularity assumptions. *J. Soc. Indust. Appl. Math.*, 5(3):105–136, 1957.

[4] J. O'Geran, H. Wynn, and A. Zhiglyavsky. Search. *Acta Applicandae Mathematicae*, 25:241–276, 1991.

[5] H. Wynn and A. Zhiglyavsky. Chaotic behaviour of search algorithms. Proc. 3rd International Conf. Workshop on Model-Oriented Data Analysis, this volume, May 1992.

The Game-Theoretical Model of an Economy

Tatiana Kulakovskaja and Adnan Shamon

This paper deals with generalized n-person cooperative market games derived from nonbalanced pure exchange economies. The economic motivation for such investigation is a problem of fair sharing rules in situation of deficit on the market of commodities.

1 The formal model

Consider a market with a set $I = \{1, 2, \ldots, n\}$ of economical agents (countries, firms, individuals and so on) and a set $K = \{1, 2, \ldots, m\}$ of commodities. Let R^m_+ be the commodities space. The agent $i \in I$ has a supply-vector $a_i = (a_i^1, \ldots, a_i^m) \in R^m_+$ and a demand-vector $b_i = (b_i^1, b_i^2 \ldots, b_i^m) \in R^m_+$. For simplicity suppose $a_i^k b_i^k = 0$, $k = 1, \ldots, m$, so if $a_i^k \neq 0$ then the agent i is a seller of a commodity k, and if $b_i^k \neq 0$ then he is a buyer of this commodity. We suppose, that utility function of an agent is additive and homogenues with respect to commodities and so we can restrict ourself by definition of the utilities for unit quantities of commodities. Let the player i has utility-vector $u_i = (u_i^1, u_i^2 \ldots, u_i^m) \in R^m_+$ as a seller, and utility vector $w_i = (w_i^1, w_i^2 \ldots, w_i^m) \in R^m_+$ as a buyer. We can interpret u_i^k as a minimal price for selling of a commodity k (for example , it's production cost) by the agent i, and, analoqous w_i^k – as a maximal price of a commodity k for buying. As in our model every agent has only one role (seller or buyer) we put $u_i^k = 0$ if $a_i^k = 0$ and $w_i^k > \max_{j: a_i^j > 0} w_i^j$ if $b_i^k = 0$.

Denote the market

$$M = \langle \{a_i\}_{i \in I}, \{b_i\}_{i \in I}, \{u_i\}_{i \in I}, \{w_i\}_{i \in I}, \{c_i\}_{i \in I} \rangle,$$

where $c_i \in R^1_+$ is the initial capital of the agent i.

Propose that u_i^k, w_i^k and c_i, $i \in I$, $k \in K$ are expressed in the same monetary units.

2 The properties of the cooperative game generated by the market

Associate the cooperative game with this market.

Let $\xi_i = (\xi_i^1, \ldots, \xi_i^m)$, $\xi_i^k \leq a_i^k$, $k = 1, \ldots, m$, $\eta_i = (\eta_i^1, \ldots, \eta_i^m)$, $\eta_i^k \leq b_i^k$, $k = 1, \ldots, m$, $i \in I$, $\sum\limits_{i=1}^{n} \xi_i^k = \sum\limits_{i=1}^{n} \eta_i^k$, $k = 1, \ldots, m$, where ξ_i^k is the quantity of a commodity k given by the player i to another ones, and η_i^k is the quantity of commodity k received by player i.

The collection $\{(\xi_i, \ldots, \xi_n); (\eta_i, \ldots, \eta_n)\}$ is named **the distribution** and is denoted (ξ, η). The set of all distributions is denoted $D(I)$.

The subsets of I are called **coalitions** in cooperative game theory. Let $S \subset I$ be the coalition and denote

$$D(S) = \{(\xi, \eta) \in D(I) | \xi_i^k = \xi_i^k = 0,\ k = 1, \ldots, m,\ i \notin S\}.$$

If the distribution $(\xi, \eta) \in D(S)$, we name it **S-distribution** and denote (ξ^S, η^S). In accordance with distribution (ξ, η) the coalition S has the total profit

$$v(S, \xi, \eta) = \sum_{k=1}^{m} \sum_{i \in S} \eta_i^k w_i^k - \sum_{k=1}^{m} \sum_{i \in S} \xi_i^k u_i^k$$

Denote $v(S) = \max\limits_{(\xi,\eta) \in D(S)} v(S, \xi, \eta)$. Put $v(\emptyset) = 0$

Then there is the cooperative game $\Gamma(M) =< I, v >$. If $a_i^k = b_i^k = 0$ for $i \in I$, $k \neq k_0$, then we name this game one-producted and denote Γ^{k_0}. We will mark all notions for Γ^{k_0} by the letter k_0 $(D^{k_0}(I), D^{k_0}(S), v^{k_0}(S)$ and so on)

So far as the distribution of every commodity can be choose independently from others it is easy to verify the folloving.

Proposition I.

$$v(S) = \sum_{k=1}^{m} v^k(S)$$

Let us research one-product game in detail. We will omit the marker k_0 in this part of the paper.

Let $N_1 \subset N$ is a set of a sellers, and $N_2 \subset N$ is a set of a buyers of the product, and $N_1 \cup N_2 = N$.

For simplicity, we suppose that

$$N_1 = \{1, \ldots, n_0\}, N_2 = \{n_0 + 1, \ldots, n\},\ 1 < n_0 < n,$$

$$u_1 \leq u_2 \leq \ldots u_{n_0}$$

$$w_{n_0+1} \leq w_{n_0+2} \leq \ldots \leq w_n$$

Utilites u_k, $k > n_0$, and w_k, $k < n_0$ are not considered.

$$D(I) = \left\{ \begin{array}{lll} (\xi, \eta) & = & \\ & = & (\xi_1, \ldots, \xi_{n_0}; \\ & & \eta_{n_0+1}, \ldots, \xi_n) \in R^n \end{array} \middle| \begin{array}{l} 0 \leq \xi_i \leq a_i \\ 0 \leq \eta_j \leq b_j \\ \sum\limits_{i=1}^{n_0} \xi_i = \sum\limits_{j=n_0+1}^{n} \eta_j \end{array} \right.$$

$$D(S) = D(I) \cap \{x \in R^n | x_i = 0 \quad i \notin S\}$$

$$v(S, \xi, \eta) = \sum_{j \in S \cap N_2} \eta_j w_j - \sum_{i \in S \cap N_1} \xi_i u_i$$

$$v(S) = \max_{(\xi,\eta) \in D(S)} v(S, \xi, \eta)$$

If $S \cap N_1 = \emptyset$ or $S \cap N_2 = \emptyset$ then $D(S) = \{(0, \ldots, 0)\}$, and $v(S) = 0$. Thus $v(\{i\}) = 0$. The coalition S is named **active**, if $v(S) \neq 0$.

The game Γ is called

(a) the game with deficit, if $\sum_{i=1}^{n_0} a_i < \sum_{j=n_0+1}^{n} b_j$

(b) the balanced game, if $\sum_{i=1}^{n_0} a_i = \sum_{j=n_0+1}^{n} b_j$

(c) the game with surplus, if $\sum_{i=1}^{n_0} a_i > \sum_{j=n_0+1}^{n} b_j$

3 The trivial S-distribution

For every active coalition $S \subseteq I$ we will describe now the algorithm of calculating a S-distribution $(\bar{\xi}^S, \bar{\eta}^S)$, that we will name a **trivial S-distribution** (Bondareva 1989)

Step 1. Let i_1, j_1 are such that

$$u_{i_1} = \min_{i \in S \cap N_1} u_i,\ w_{j_1} = \max_{j \in S \cap N_2} w_j$$

$$x_{i_1 j_1} = \begin{cases} \min(a_{i_1} b_{j_1}) & \text{if } u_{i_1} \leq w_{j_1} \\ 0 & \text{otherwise} \end{cases}$$

Step 2. If $x_{i_1 j_1} = 0$ then "stop"

If $x_{i_1 j_1} = a_{i_1}$ then $S \Rightarrow S \setminus \{i_1\},\ b_{j_1} = b_{j_1} - a_{j_1}$

If $x_{i_1 j_1} = b_{j_1}$ then $S \Rightarrow S \setminus \{j_1\},\ a_{i_1} = a_{i_1} - b_{j_1}$

Step 3. If $S \cap N_1 = \emptyset$ or $S \cap N_2 = \emptyset$ then "stop"

else goto step 1 with the new S, a_{i_1}, b_{j_1}

Let l be a number of iteration on that the process stops. It will occur in one of the three cases:
(a) $x_{i_l j_l} = 0$ and $u_{i_l} > w_{i_l}$
(b) $S \cap N_1 = \emptyset$;
(c) $S \cap N_2 = \emptyset$.

We put

$i(S) = i_{l-1},\ j(S) = j_{l-1}$ in the case (a)

$i(S) = i_l,\ j(S) = j_l$ in the case (b),(c)

$S_1 = \{i \in S \cap N_1 | i \leq i(S)\}$

$S_2 = \{j \in S \cap N_2 | j \geq j(S)\}$

The pair $(i(S), j(S))$ is named **S-marginal.**

$$\bar{\xi}_i^S = \sum_{j \in S_2} x_{ij}, \quad i \in S_1$$

Let $\bar{\eta}_j^S = \sum_{i \in S_1} x_{ij}, \quad j \in S_2$

$$\bar{\xi}_i^S = \bar{\eta}_j^S = 0 \text{ for other } i, j \in N$$

It is obvious that $(\bar{\xi}^S, \bar{\eta}^S)$ is S-distribution and we call it **trivial.**

Proposition 2. The trivial S-distribution has the following properties:

(a) $\bar{\xi}_i^S = a_i, \quad i \in S_1, \ i \neq i(S)$

(b) $\bar{\eta}_j^S = b_j, \quad j \in S_2, \ j \neq j(S)$

(c) $\sum_{i \in S \cap N_1} \bar{\xi}_i^S = \sum_{j \in S \cap N_2} \bar{\eta}_j^S = \min\left(\sum_{i \in S_1} a_i, \sum_{j \in S_2} b_j\right)$ (d) $\bar{\xi}_{i(S)}^S = a_{i(S)}$ or $\bar{\eta}_{j(S)}^S = b_j(S)$ but both equalities are true only if $\sum_{i \in S_1} a_i = \sum_{j \in S_2} b_j$

(e) $v(S, \bar{\xi}^S, \bar{\eta}^S) = \max_{(\xi,\eta) \in D(S)} v(S, \xi, \eta) = v(S)$

Proof. Properties (a)-(d) obviously followes from algorithm. Let consider (e). Denote $x^S = \|x_{ij}\|_{i,j \in S}$ the plan of exchanges between the members of coalition S, where x_{ij} is a quantity of product tranfered from agent i to agent j. If $i \notin S \cap N_1$, $j \notin S \cap N_2$, or $u_i > w_j$ then $x_{ij} = 0$.

It is clear, that

$$\xi_i^S = \sum_j x_{ij}, \ \eta_j^S = \sum_i x_{ij}$$

$$v(S, \xi, \eta) = \sum_{j \in S \cap N_2} \eta_j^S w_j - \sum_{i \in S \cap N_1} \xi_i^S u_i = \sum_{i,j \in S} x_{ij}(w_j - u_i)$$

We have an optimization problem:

$$\max \sum_{i,j \in S} x_{ij}(w_j - u_i)$$

$$\sum_i x_{ij} \leq b_j, \ j \in S \cap N_2$$

$$\sum_j x_{ij} \leq a_i, \ i \in S \cap N_1$$

$$x_{ij} \geq 0, \quad i, j \in S$$

$$x_{ij} = 0, \text{ if } i \in S \setminus N_1, \text{ or } j \in S \setminus N_2, \text{ or } u_i > w_j$$

Let the plan $\bar{x}^S$ be a solution of this problem. $\bar{\xi}_i^S = \sum_j \bar{x}_{ij}; \bar{\eta}_j^S = \sum_i \bar{x}_{ij}$. It is easy to show, that $(\bar{\xi}_i^S, \bar{\eta}_j^S)$ is the trivial S-distribution. In such form our problem is the sort of transport one, and $\bar{x}^S$ is so-called North-West plan for it. This plan is optimal in our case, although that is not so for arbitrary transport problem . Q.E.D.

Remark, that if all positive $c_{ij} = w_j - u_i$ are different then trivial S-distribution is unique solution of the corresponding optimization problem, and so we have a rather strict way of achievement of maximal total profit for coalition.

For the market with m commodities the trivial I-distribution $(\bar{\xi}, \bar{\eta})$ is a collection of similier distributions for every commodities, that is $(\bar{\xi}, \bar{\eta}) = \{(\bar{\xi}^k, \bar{\eta}^k)\}_{k=1}^m$, and it is unique optimal distribution for whole market.

4 Balanced distribution and balanced prices

Now consider the market with the fixed price-vector $p = (p^1, \ldots, p^m)$. The prices with the distribution define **allocation** $x = (x_1, \ldots, x_n)$ of the total profit, where

$$x_i = x_i(p, \xi, \eta) = \sum_{k=1}^{m} \xi_i^k (p^k - u_i^k) + \sum_{k=1}^{m} \eta_i^k (w_i^k - p^k)$$

$H(\Gamma) = \{x(p, \xi, \eta);\ (\xi, \eta) \in D(I)\}$ is the set of allocations.

For one-product game we have

$$x_i(p, \xi, \eta) = \xi_i(p - u_i),\ i \in N_1$$

$$x_j(p, \xi, \eta) = \eta_j(w_j - p),\ j \in N_2$$

We call the allocation vector x **balanced** if

$$\sum_{i \in S} x_i \geq v(S), \quad \text{for all } S \subset I,\ S \neq I$$

The corresponding distribution (ξ^*, η^*) and the price-vector p^* also are named balanced.

Theorem 1. (Bondareva 1990)

The trivial I-distribution $(\bar{\xi}, \bar{\eta})$ is balanced one. The balanced price-vector p^* can be choosen arbitrary from the set

$$P^* = \left\{(p^1, \ldots, p^m) | u_{i^k(I)}^k \leq \underline{p}^k \leq p^k \leq \overline{p}^k \leq w_{j^k(I)}^k\right\}$$

It is desirable to connect the balanced prices and the balanced distribution with competitive equilibrium in the perfect market with transferable utilities and money.

Put

$$U_i(\xi, \eta) = \sum_{k=1}^{m} \left[u_i^k (a_i^k - \xi_i^k) + w_i^k \eta_i^k + p^k \xi_i^k - p^k \eta_i^k\right]$$

$$B(p, c) = \left\{(\xi, \eta) \in D(I) | \sum_{k=1}^{m} (p^k \eta_i^k - p^k \xi_i^k) \leq c_i,\ i = 1, \ldots, n\right\}$$

$B(p, c)$ is named a **budget set** .

Theorem 2. A vector C^* exists such that for $c_i \geq c_i^*$, $i = 1, \ldots, n$ the trivial distribution $(\bar{\xi}, \bar{\eta})$ and corresponding balanced price-vector p^* form a competitive equlibrium in the market M with utility functions $U_i(\xi, \eta)$, $i = 1, \ldots, n$ and budget set $B(p, c)$.

5 Computer experiments and concluding remarks

The trivial distribution is accessible only for perfect market, and the problem remains to look for another suitable distributions, for example another balanced distributions.

Let us consider the following optimization problem

$$\text{extr} f(\xi, \eta, p)$$

$$\sum_{i \in S} \sum_{k=1}^{m} \left(\xi_i^k (p^k - u_i^k) + \eta_i^k (w_i^k - p^k) \right) \geq v(S) \quad \text{for } S \subset I,\ S \neq I$$

$$(\xi, \eta) \in D(I);\ p^k \geq 0,\ k = 1, \ldots, m$$

This problem is rather difficult because we deal with the large system of nonlinear inequalities. Our purpose of the optimization would be one of the followings:
(a) $\max v(I, \xi, \eta)$; (b) $\min v(I, \xi, \eta)$; (c) $\min\limits_{k=1}^{m} \alpha_k p^k$; and so on.

We know the answer in the case (a) – trivial distribution and corresponding prices. In other cases we have the results of computer experiments. The problem becomes linear if we consider vector p as a parameter-vector. But the method of linear programming is effective only for a small n. The another way is one of the methods of random search. But the calculating difficulties become hopeless already for a market with 15-20 agents.

The next idea is the introduction of a coalitional structure. The existence of such structure could be explained by communicational restictions or unperfect unformation for real market.

Our computer system includes the complex of procedures which allow for given market M and determinated coalitional structure to construct trivial distribution and the set of corresponding balanced prices, to calculate $v(S)$ for all possible coalitions, to build another balanced distributions, to solve in dialoque would the given price-vector p be balanced for any distribution and to answer another questions about given market M.

References

Ichiishi T. (1983) Game theory for economic analysis. Academic press, New York.

Bondareva O.N (1989) A cooperative game "supply-demand" as a model of an non-balanced market. Vestnik LGU, 22.

LIST OF CONTRIBUTORS

- A.C. Atkinson, Department of Statistical and Mathematical Sciences, London School of Economics, London WC2A 2AE, England, United Kingdom.
- A.G. Bart, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- M.V. Chekmasov, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- N.P. Clochkova, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- S.M. Ermakov, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- V.V. Fedorov, Department of Applied Statistics, School of Statistics, 352 Classroom-Office Building, 1994 Buford Avenue, St.Paul, Minnesota, MN 55108-6042, USA.
- V.N. Fomin, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- P. Hackl, Department of Statistics, University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria.
- R.-D. Hilgers, Institute of Medical Documentation and Statistics, University of Cologne, Joseph-Stelzmann-Str.9, D-W-5000 Köln 41, Germany.
- M. Jalava, Helsinki University of Technology, Laboratory of Hydrology and Water Resources Management, Rakentajanaukio 4A, SF–02150 Espoo, Finland.
- J.N. Kashtanov, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- J. Kettunen, Helsinki University of Technology, Laboratory of Hydrology and Water Resources Management, Rakentajanaukio 4A, SF–02150 Espoo, Finland.
- C.P. Kitsos, Department of Statistics, Athens University of Business and Economics, 76 Patision St., 104 34 Athens, Greece.
- M.V. Kondratovich, Vinnitsa State Pedagogical Institute, Krasnoznamennaya st.32, Vinnitsa, 286004, Ukraine.
- V.M. Kozhanov, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- V.P. Kozlov, State Optical Institute, Postamtskay 11–21, 190000 St. Petersburg, Russia.
- A.E. Kraskovsky, Institute of Transport Problems, Russian Academy of Sciences, V.O., 12 linia, 199178 St.Petersburg, Russia.
- N. Krivulin, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.

- T. Kulakovskaja, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- C. Kulcsár, Laboratoire des Signaux et Systèmes, CNRS-ESE, Plateau du Moulon, 91192 Gif sur Yvette cedex, France.
- J. Kunert, FB Statistik, Universität Dortmund, PF 500500, W4600, Dortmund, Germany.
- J. López-Fidalgo, Departamento de Matemática Pura y Aplicada, Universidad de Salamanca, Plaza de la Merced 1-4, 37008 Salamanca, Spain.
- S.X.C. Lou, Faculty of Management, University of Toronto, 246 Bloor St. W., Toronto, Canada M5S 1V4, Canada. This research was supported in part by grants from URIF and MRCO.
- A.V.Makshanov, Naval Academy, Vyorskaya nab. 73/1, 197045 St.Petersburg, Russia.
- V.B. Melas, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- Ch. Müller, Freie Universität Berlin, Fachbereich Mathematik, WE 1, Arnimallee 2-6, D-W-1000 Berlin 33, Germany.
- W.G. Müller, Department of Statistics, University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria.
- A.K. Musrati, Department of Statistics, University of Glasgow, Faculty of Science, Glasgow G12 8QQ, Scotland, United Kingdom.
- A.C. Ponce de Leon, Department of Statistical and Mathematical Sciences, London School of Economics, London WC2A 2AE, United Kingdom. On leave of absence from University of Rio de Janeiro State (UERJ), Brazil. This research was partially supported by *CAPES/MEC* (Brazil), while the first author pursued a PhD degree at the LSE.
- L. Pronzato, Laboratoire I3S, CNRS-URA 1376, Bat. 4, 250 rue Albert Einstein, Sophia-Antipolis, 06560 Valbonne, France.
- R. Schwabe, Freie Universität Berlin, Fachbereich Mathematik, 1. Mathematisches Institut, Arnimallee 2–6, D-W-1000 Berlin 33, Germany.
- A. Shamon, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- G.I.Simonova, Triniti, Troitsk, Moscow Region, Mier "B", 34-7, 142092, Russia.
- A.S. Tikhomirov, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.
- B. Torsney, Department of Statistics, University of Glasgow, Faculty of Science, Glasgow G12 8QQ, Scotland, United Kingdom.
- Yu.N. Tyurin, Moscow State University, Leninskie gory, 117234 Moscow, Russia.

- J. Volaufová, Institute of Measurement Science, Slovak Academy of Sciences, Dubravska 9, 842 19 Bratislava, Slovakia.

- E. Walter, Laboratoire des Signaux et Systèmes, CNRS-ESE, Plateau du Moulon, 91192 Gif sur Yvette cedex, France.

- H.P. Wynn, Dept. of Mathematics, City University, Northampton Square, London, EC1V OHB, England, United Kingdom.

- H.M. Yan, Faculty of Management, University of Toronto, 246 Bloor St. W., Toronto, Canada M5S 1V4, Canada.

This research was supported in part by grants from URIF and MRCO.

- G. Yin, Department of Mathematics, Wayne State University, Detroit, MI 48202, U.S.A.

This research was supported in part by the National Science Foundation.

- A.A. Zhigljavsky, Department of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaja sq. 2, St.Petersburg, Petrodvorets, 198904, Russia.

- R. Zieliński, Institute of Mathematics, Polish Academy of Science, P.O. Box 137, ul. Sniapeckich 8, 00–950 Warszaw, Poland.

Supported by Grant KBN 2–1168–91–01 UW GR–101.

Contributions to Statistics

The series "Contributions to Statistics" contains publications in statistics and related fields. These publications are primarily monographs and multiple author works containing new research results, but conference and congress reports are also considered.

Apart from the contribution to scientific progress presented, it is a notable characterisitic of the series that actual publishing time is very short thus permitting authors and editors to present their results without delay.

Manuscripts could be sent directly to Physica-Verlag:

Physica-Verlag GmbH & Co.
P. O. Box 10 52 80
69042 Heidelberg, FRG

ISBN 3-7908-0711-7
ISBN 3-387-91457-9