TIMS

Studies in the Management Sciences

Volume 19

OPTIMIZATION IN STATISTICS

S. H. ZANAKIS
J. S. RUSTAGI
editors

North-Holland

OPTIMIZATION IN STATISTICS

With a view towards applications in Management Science and Operations Research

Edited by

S.H. ZANAKIS
J.S. RUSTAGI

under the Departmental Editorship of
Arie Y. Lewin for *Management Science*

1982

NORTH-HOLLAND PUBLISHING COMPANY—AMSTERDAM · NEW YORK · OXFORD

This North-Holland/TIMS series is a continuation of the Professional Series in the Management Sciences, edited by Robert E. Machol.

ISBN: 0444 86544 6

Published by:

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM, NEW YORK, OXFORD

Sole distributors for the U.S.A. and Canada:

ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE
NEW YORK, NY 10017

Printed in The Netherlands

OPTIMIZATION IN STATISTICS

TIMS STUDIES IN THE MANAGEMENT SCIENCES

Editor in Chief

ROBERT E. MACHOL

Volume 19

NORTH-HOLLAND PUBLISHING COMPANY—AMSTERDAM·NEW YORK·OXFORD

PREFACE

Need for optimization arises often in statistics implicitly or explicitly. Although significant developments have been made in both fields, their interface has not received the attention it deserves. Statisticians on one hand and management scientists/operations researchers on the other have made many important contributions to this interdisciplinary subject through their own *separate* publications and conferences. There has been very little participation in each other's activities – even those devoted to the common subject. The first two symposia dedicated to "Optimizing Methods in Statistics" (Rustagi) were attended mostly by statisticians, while participants in "Optimization in Statistics" sessions at recent ORSA/TIMS National Meetings (Zanakis) were primarily operations researchers/management scientists.

Encouraged by the responses of participants in these activities, we felt that a *TIMS Studies in the Management Sciences* volume devoted to this subject would bring statisticians and management scientists closer together and stimulate further interest in this important area. A call for papers was announced in late 1979/early 1980 through national and international journals of statistics, operations research, management science, and industrial engineering. We sought original and expository papers in the area of optimization as utilized in statistical problems arising in the above disciplines. Development of new and intelligent application or comparison of existing optimization algorithms to important statistical problems were encouraged.

This volume contains the papers which survived the usual *Management Science* refereeing and editing processes. We believe that the volume will appeal to management scientists, operations researchers, statisticians, and industrial engineers. It can be used as a reference or textbook in colleges and universities.

The papers in this volume are grouped in three sections according to the statistical applications area: (a) regression and correlation; (b) multivariate data analysis and design of experiments; and (c) statistical estimation, reliability and quality control. An introduction to the articles of each section is provided, as well as an introduction to the overall subject. The reader is assumed to be familiar with elements of statistics, especially in the categories listed above, and with elements of optimization theory, especially linear programming. No great mathematical sophistication is required for reading this book – a knowledge of algebra and elementary calculus will suffice for most of the chapters.

As editors of this volume, we would like to express our sincere appreciation to the authors, whose work and interest made this volume possible; to about fifty referees, whose detailed and critical reviews were indispensable to improve the quality of so many different papers; to Professor Robert E. Machol, Editor-in-Chief of the TSMS publication series, for reviewing carefully the draft of this volume; and to our secretaries at the West Virginia College of Graduate Studies, Florida International University and The Ohio State University for efficiently handling the large volume of correspondence during all phases of this effort.

Miami, Florida — Stelios H. Zanakis
Columbus, Ohio — Jagdish S. Rustagi *

September 1981 — Editors

* The work by J.S. Rustagi was partially supported by contract no. N00-14-78-C-0543 (NR 042-403) of the Office of Naval Research.

TABLE OF CONTENTS

II. MULTIVARIATE DATA ANALYSIS AND DESIGN OF EXPERIMENTS

III. PARAMETER ESTIMATION, RELIABILITY AND QUALITY CONTROL

TIMS/Studies in the Management Sciences 19 (1982) 1–3

A PRELUDE TO OPTIMIZATION IN STATISTICS

S.H. ZANAKIS and J.S. RUSTAGI

Lack of certainty creates a need for statistics in any field of human endeavor. Statistical techniques are widely used in many areas of study, including sociology, medicine, business, and engineering – to mention only a few. All statistical problems contain elements of optimization. In general, this necessitates estimation of some unknown parameters (variables) that will optimize some measure of performance while often satisfying certain restrictions (constraints) on the parameters and input/output data.

The following examples occur frequently in statistics and are examined in more detail later in this volume.

- Linear and nonlinear curve fitting by regression analysis for predictions that minimize some measure of error (absolute, squared, maximum, etc.).
- Design of experiments and sample surveys that minimize costs while they restrict errors, or vice versa.
- Clustering, classification or discrimination of multivariate data according to different attributes in order to minimize improper handling of data and maximize estimation accuracy.
- Estimation of parameters of statistical distributions using criteria such as maximum likelihood, minimum chi-square, minimum variance, etc. (Bard [2] 1979, Zanakis [13,14]). The last criterion is used also in optimal response surface designs as well as in canonical correlation and principal component analysis of multivariate data (linear transformations of variables to maximize or minimize their correlations).
- Test of statistical hypotheses that minimize the probability of type II error while holding the probability of type I error at a certain level (Neyman–Pearson [6] problem).
- Quality control design problems arise in both acceptance sampling (e.g. minimize inspector errors) as well as in process control (e.g. detect outliers, and design minimum cost control charts).
- Reliability design and maintenance policies that minimize costs and restrict probabilities of breakdown or vice versa.

In contrast to classical statistics, Bayesian statistics is concerned with statistical procedures when some prior information is available on the unknown parameter. Bayesian statistics uses this prior information with the sample to derive the posterior distribution on the parameters. The procedures are based on this posterior distribution. The evaluation of these procedures often requires the assumption of a loss function. Such procedures require the application of optimization techniques.

Optimization methods to solve all previous problems may be classified into one of the following categories: classical (using calculus), numerical, variational (including dynamic programming) and other mathematical programming (constrained linear, nonlinear and integer optimization). The availability of high-speed computers has accelerated the growth of management science/operations research and produced many algorithmic and software developments in optimization. The use of such new optimization techniques in statistics, however, has not been extensive, primarily due to four reasons.

First, the topics of optimization and statistics have not been sufficiently blended in graduate education. Secondly, no books integrating the two subjects have been available until recently (Rustagi [9] and Arthanari and Dodge [1] have provided expositions of variational and mathematical programming techniques in statistics). Thirdly, most of the optimization techniques have been developed by nonstatisticians and it is true that there always exists some time lag between the development of some techniques (e.g. mathematical programming) and their use in another area (e.g. statistics). Finally, discourses on the optimization–statistics interface have been scattered in the literature, and until recently professional society meetings on this interface were infrequent.

The organization of two statistical symposia on "Optimizing Methods in Statistics" (Rustagi [8,11]) and recognition of the importance of this interdisciplinary area for management scientists and operations researchers (e.g. see Zanakis [13,15]) have increased the awareness of the field. We were very pleased to see our first "Optimization in Statistics" session at the Joint National Meeting of the Operations Research Society of America and The Institute of Management Sciences on 2 April 1976 followed by similar sessions in each subsequent semiannual joint national conferences of these two societies, and more recently of the American Institute for Decision Sciences. A special issue of *Communications in Statistics* was devoted to optimization in statistics (Rustagi [10]) and an international conference on "Optimizing Methods in Statistics" was sponsored in 1977 by the Bernoulli Society for Statistics in the Physical Sciences.

The present volume is an outgrowth of all these efforts. Contributors come from a variety of disciplines, including statistics, operations research/management science, business administration, industrial engineering, etc. We hope that this effort to cross the boundaries of single disciplines on this common theme will be soon followed by others.

In order to restrict this volume to a manageable size, it was decided not to include subjects of the opposite interface, namely use of statistics in optimization. Interest in this subject has been renewed, as is evident from the inclusion of such sessions in several recent ORSA/TIMS national meetings (Golden and

Zanakis). Considerable attention has been given recently to the development and application of statistical inference procedures for point and interval estimation of the true (global) optimum when this result cannot be guaranteed by the optimization technique used. (See Sielken and Monroe [12] for a comprehensive exposition). This is a vital need when solving nonconvex mathematical programming problems, or when employing heuristics, for example to solve large, otherwise unsolvable, integer programming problems (Zanakis and Evans [15]). Other efforts have included search procedures and stopping rules guided by Bayesian learning (Hill [3]), sequential hypothesis testing (Robbins [7]), or regression curve-fitting for response surface analysis (Myers [4]). Use of statistics to handle uncertainties in linear programming (stochastic and chance-constrained programming) have been well known for some time. The growing interest in the subject of statistics in optimization will, hopefully, result soon in a book on this topic.

References

[1] T.S. Arthanari and Y. Dodge, *Mathematical Programming in Statistics* (Wiley, New York, 1981).

[2] Y. Bard, *Nonlinear Parameter Estimation* (Academic Press, New York, 1974).

[3] J.D. Hill, "A Search Technique for Multimodal Surfaces", *IEEE Trans. on Syst. Sc. & Cybernetics* 5 (1969) 2–8.

[4] R.H. Myers, *Response Surface Methodology* (Allyn and Bacon, Boston, 1971).

[5] M.F. Neuts (Ed.), *Algorithmic Methods in Probability,* TIMS Studies in the Management Sciences, vol. 7 (North-Holland, Amsterdam, 1977).

[6] J. Neyman and E.S. Pearson, "Contributions to the Theory of Testing Statistical Hypothesis", Stat. Res. Mem. I (1936) 1–37 and II (1938) 25–57.

[7] T.C. Robbins, "Go–Stop: Global Optima by a Sequential Testing Optimization Procedure", Paper presented at 39th National ORSA Meeting, Dallas, 5–7 May 1971.

[8] J.S. Rustagi (Ed.), *Optimizing Methods in Statistics* (Academic Press, New York, 1971).

[9] J.S. Rustagi, *Variational Methods in Statistics* (Academic Press, New York, 1976).

[10] J.S. Rustagi (Ed.), "Optimization in Statistics", *Communications in Statistics,* Special Issue (1978).

[11] J.S. Rustagi (Ed.), *Optimizing Methods in Statistics* (Academic Press, New York, 1979).

[12] R.L. Sielken, Jr. and H. Monroe, "How Near is a Near-Optimal Solution: Confidence Limits for the Global Optimum", Technical Reporet no. 62, Institute of Statistics, Texas A&M University (May 1980).

[13] S.H. Zanakis, "Optimization and Statistics: A Synergism", Southeast TIMS Meeting Proceedings (1973) 427–434.

[14] S.H. Zanakis, "Computational Experience with Some Nonlinear Optimization Algorithms in Deriving Maximum Likelihood Estimates for the Three-Parameter Weibull Distribution", in: M.F. Neuts, ed., *Algorithmic Methods in Probability,* TIMS Studies in The Management Sciences, vol. 7 (North-Holland, Amsterdam, 1977) 63–77.

[15] S.H. Zanakis and J.R. Evans, "Heuristic Optimization: Why, When and How to Use It", *Interfaces* 11 (October 1981) 84–91.

PART I

REGRESSION AND CORRELATION

TIMS/Studies in the Management Sciences 19 (1982) 7–9

INTRODUCTION TO CONTRIBUTIONS IN REGRESSION AND CORRELATION

S.H. ZANAKIS and J.S. RUSTAGI

Analysts, experimenters, and decision-makers use regression analysis extensively in many areas of business, science, and technology. Surveys of Fortune 500 firms have shown that regression analysis is the most often used technique in operations research/management science, particularly in the areas of accounting (Jagetia [7]), quality control and maintenance (Ledbetter and Cox [8]).

Correlation analysis provides a measure of linear relationship (association) among a set of variables *without* regard to cause–effect direction. Regression analysis, however, requires such a direction for prediction purposes: i.e. the response (dependent variable $\boldsymbol{y}$) is assumed to be related to one or more independent variables ($\boldsymbol{x}$) through a functional form or model $\boldsymbol{y}=f(\boldsymbol{p}, \boldsymbol{x})$. The parameters, $\boldsymbol{p}$, of the model are estimated by minimizing, over a set of n observations $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$, some function of deviations of the hypothesized model response $\hat{\boldsymbol{y}}$ from the observed value $\boldsymbol{y}$, e.g.

$$\text{minimize} \sum_{i=1}^{n} |y_i - \hat{y}_i|^K.$$

The power constant K is usually set at $K=1$ (the least-absolute deviation criterion or L_1 norm, also called the Tchebycheff criterion), $K=2$ (the least-squares deviation criterion or L_2 norm), or $K=\infty$ (the least maximum deviation criterion or L_∞ norm, which is equivalent to minimize $\max|y_i-\hat{y}_i|$). There is extensive literature on various methods as applied to regression problems using different optimization criteria.

The classical procedure of least squares has been used extensively because of its analytical tractability and its highly developed theory and widespread literature. For a more complete account the reader is referred to Rao [9], Draper and Smith [4], and Chatterjee and Price [3]; to Wesolowsky [11] for a computer-oriented treatment; and to Harter [6] for a historical exposition.

The criterion of minimizing the sum of absolute deviations is preferable to that of least squares in the presence of large disturbances (outliers) or when the classical assumption of a normal distribution of error, $e_i=y_i-\hat{y}_i$, is violated due to "contamination" or "heavy tails". Nevertheless, the use of the least