THEORY AND DECISION LIBRARY

SERIES B: MATHEMATICAL AND STATISTICAL METHODS

V. SERDOBOLSKII

MULTIVARIATE STATISTICAL ANALYSIS

A High-Dimensional Approach

KLUWER ACADEMIC PUBLISHERS

MULTIVARIATE STATISTICAL ANALYSIS

A High-Dimensional Approach

by

V. Serdobolskii

Department of Applied Mathematics,
Moscow Institute of Electronics and Mathematics,
Moscow, Russia

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 0-7923-6643-3

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands

MULTIVARIATE STATISTICAL ANALYSIS

THEORY AND DECISION LIBRARY

General Editors: W. Leinfellner (Vienna) and G. Eberlein (Munich)

Series A: Philosophy and Methodology of the Social Sciences

Series B: Mathematical and Statistical Methods

Series C: Game Theory, Mathematical Programming and Operations Research

SERIES B: MATHEMATICAL AND STATISTICAL METHODS

VOLUME 41

Scope: The series focuses on the application of methods and ideas of logic, mathematics and statistics to the social sciences. In particular, formal treatment of social phenomena, the analysis of decision making, information theory and problems of inference will be central themes of this part of the library. Besides theoretical results, empirical investigations and the testing of theoretical models of real world problems will be subjects of interest. In addition to emphasizing interdisciplinary communication, the series will seek to support the rapid dissemination of recent results.

PREFACE

In the last few decades the accumulation of large amounts of information in numerous applications has stimulated an increased interest in multivariate analysis. Computer technologies allow one to use multi-dimensional and multi-parametric models successfully. At the same time, an interest arose in statistical analysis with a deficiency of sample data. Nevertheless, it is difficult to describe the recent state of affairs in applied multivariate methods as satisfactory. Unimprovable (dominating) statistical procedures are still unknown except for a few specific cases. The simplest problem of estimating the mean vector with minimum quadratic risk is unsolved, even for normal distributions. Commonly used standard linear multivariate procedures based on the inversion of sample covariance matrices can lead to unstable results or provide no solution in dependence of data. Programs included in standard statistical packages cannot process 'multi-collinear data' and there are no theoretical recommendations except to ignore a part of the data. The probability of data degeneration increases with the dimension n, and for $n > N$, where N is the sample size, the sample covariance matrix has no inverse. Thus nearly all conventional linear methods of multivariate statistics prove to be unreliable or even not applicable to high-dimensional data.

This situation is by no means caused by lack of the necessary advancing theoretical support of multivariate analysis. The traditional Fisher approach was developed for classical problems with simple models and arbitrarily large samples. The principle requirement on statistical estimators was consistency, i.e., convergence to true values for a fixed model, as the sample size increases. Traditionally, statistical procedures are developed by a substitution of consistent estimators into the extremal theoretical solutions (the 'plug-in' procedure).

However, the component-wise consistency does not provide satisfactory solutions to the problems of the multivariate analysis. In the case of a high dimension, the cumulative effects of estimating a large number of parameters can lead to a substantial loss of quality and to the breakdown of multivariate procedures.

It is well known that classical mathematical investigations in multivariate statistical analysis were reduced to the calculation of some exact distributions and their functions under the assumption that the observations are normal. The well developed traditional asymptotic theory of statistics is oriented to one-dimensional and low-dimensional problems. Its formal extrapolation to multi-dimensional problems (by replacing scalars by vectors without analyzing specific effects) enriched the statistics neither with methods, nor with qualitatively new theoretical results. One can say that central problems of the multivariate analysis remain unsolved.

Some essential progress has been achieved after a number of investigations in 1970–1974 pioneered by A.N. Kolmogorov. He suggested a new asymptotic approach differing by a simultaneous increase of the sample size N and the dimension n of variables so that the ratio n/N tends to a constant. This constant became a new parameter of the asymptotic theory. In contrast to the traditional asymptotic approach in mathematical statistics, this new approach was called the *increasing dimension asymptotics* (see the handbooks [1, 2]). The investigation of terms of the order of magnitude n/N led to the discovery of a series of new specific phenomena in high-dimensional problems such as accumulation of estimation errors, appearance of finite biases and multiples, and a certain normalization effect when, under some 'restricted dependence conditions', all distributions prove to be equivalent to normal distributions with respect to functionals uniformly depending on variables. In particular, this means that standard quality functions of multivariate procedures prove to be approximately distribution-free and that, at last, we obtain a tool for comparing different versions of procedures.

An investigation of the leading terms of the increasing dimension asymptotics led to the construction of a systematic theory of multivariate analysis characterized by other settings, specific problems, and results of interest for applications. A statistical problem in which the dimension of observations is comparable to the sample size may be called an *essentially multivariate problem.* The statistical analysis taking into account finite effects produced by the estimation of a large number of parameters and related to the solution of essentially multivariate problems may be called the *essentially multivariate analysis.*

The central idea of the investigation of essentially multivariate effects is to study relations between empirical distribution functions of true parameters and of their estimators. Limit equations are derived that connect spectral functions of sample covariance matrices and of true covariance matrices. Such relations proved to be of a special interest for the essentially multivariate approach, since they present a device for a regular construction of improved estimators in different multivariate problems. Using these relations one can first single out non-random leading parts of quality functionals involved in multivariate analysis and then construct their consistent estimators. To obtain an improved procedure, it suffices to maximize these estimators.

The book consists of an Introduction and twelve chapters. The introduction presents historical aspects and the line of development of main ideas. In Chapter 1 the reader will recall the fundamentals of the theory of multivariate analysis in the case when the underlying distributions are normal. In Chapters 2–11 the results of the original investigations are presented. These chapters are mostly independent of each other and written so that they can be read separately.

I hope that specialists in mathematical statistics will be interested in this new branch of the theory of statistics and in the new phenomena investigated. The essentially multivariate statistics is different in its approach, in its special techniques, and in its results of a new kind.

Applied statisticians and users of statistical software will be interested in more efficient methods of practical multivariate analysis that can be developed by using essentially multivariate methods. In fact, nearly all existing software for applied multivariate analysis is now obsolete. The essentially multivariate technique promises to provide stable, uniformly consistent with respect to the number of variables, approximately non-improvable methods whose quality does not depend on distributions.

Students of mathematics obtain a text-book, unique for today, for studying the recently created theory of more efficient methods of multivariate analysis. For a new generation of mathematicians this theory may undoubtedly serve as a reliable basis for their future success in the science of 21st century.

I would like to express my sincere gratitude to Yurii Vasilievich Prokhorov for his attention, invariable support of my investigations, and wise recommendations. Also I am heartily thankful to Victor Matveevich Bukhshtaber for an enthusiastic attitude and a suggestion to write this book.

V.I. Serdobolskii

CONTENTS

INTRODUCTION

Here we briefly outline the history and the development of the essentially multivariate approach in statistics, and principle features of its ideas, notions, and achievements.

Let us introduce necessary notations. We consider n-dimensional populations $\mathfrak{S}$; vectors $\mathbf{x}$ from $\mathfrak{S}$ are called observations. We denote the expectation operator by $\mathbf{E}$, and the function of variance by $\operatorname{cov}(\cdot)$. Let $\Sigma = \operatorname{cov}(\mathbf{x}, \mathbf{x})$ denote the covariance matrix. We consider samples $\mathfrak{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ from $\mathfrak{S}$ of size N and use sample means and matrices

$$\bar{\mathbf{x}} = N^{-1} \sum_{m=1}^{N} \mathbf{x}_m, \qquad C = N^{-1} \sum_{m=1}^{N} (\mathbf{x}_m - \bar{\mathbf{x}})(\mathbf{x}_m - \bar{\mathbf{x}})^T \tag{1}$$

along with

$$S = N^{-1} \sum_{m=1}^{N} \mathbf{x}_m \mathbf{x}_m^T \tag{2}$$

(matrices S can have the sense of sample covariance matrices if the expected values $\mathbf{x}$ are known a priori). We denote vectors by semi-boldface symbols and mark transposed column vectors by the upper symbol 'T'. Let the absolute value of a vector denote its length, and the square of a vector denote the square of its length. We only use the spectral norms of matrices. The indicator function $\operatorname{ind}(\cdot)$ will be also used in non-random relations. Let I denote the identity matrix.

Kolmogorov Asymptotics in Problems of Multivariate Analysis

The essentially multivariate approach in statistics was developed first in 1967–1988 in application to the discriminant analysis. First, let us briefly describe the progress achieved before 1980.

The discriminant analysis problem is set as follows.

Suppose two n-dimensional populations are given $\mathfrak{S}_\nu$, $\nu = 1, 2$, and we have samples $\mathfrak{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_1})$ and $\mathfrak{X}_2 = (\mathbf{x}_{N_1+1}, \ldots, \mathbf{x}_N)$, where $N = N_1 + N_2$, from $\mathfrak{S}_1$, and $\mathfrak{S}_2$, respectively. A sample discriminant function $w(\mathbf{x}) = w(\mathbf{x}, \mathfrak{X}_1, \mathfrak{X}_2)$ is constructed and a threshold c is fixed. The discrimination rule is of the form $w(\mathbf{x}) > c$ against $w(\mathbf{x}) \leqslant c$. Probabilities of errors (conditional under fixed samples) are

$$\alpha_1 = \mathbf{P}\big(w(\mathbf{x}) \leqslant c \mid \mathbf{x} \in \mathfrak{S}_1\big), \qquad \alpha_2 = \mathbf{P}\big(w(\mathbf{x}) > c \mid \mathbf{x} \in \mathfrak{S}_2\big). \quad (3)$$

For normal populations $\mathfrak{S}_\nu = \mathbf{N}(\mu_\nu, \Sigma)$, $\nu = 1, 2$, with a common non-degenerate known covariance matrix Σ, the minimum of $(\alpha_1 + \alpha_2)/2$ is provided (by virtue of the Neumann–Pearson Lemma) with the Anderson discriminant function

$$w^A(\mathbf{x}) = (\mu_1 - \mu_2)^T \Sigma^{-1} \big(\mathbf{x} - (\mu_1 + \mu_2)/2\big),$$

which presents the logarithm of the ratio of normal probability densities. The minimum attained with $c = 0$ is $\alpha_1 = \alpha_2 = \Phi(-\sqrt{J}/2)$, where $J = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$ is the square of the 'Mahalanobis distance'. Standard consistent multivariate procedures are usually constructed by a replacement of the parameters μ_1,, μ_1, and Σ by standard estimators ('plug-in procedure'). Thus the Fisher–Anderson–Wald sample discriminant function is constructed

$$w(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T C^{-1} \big(\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2\big), \quad (4)$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are sample mean vectors and C is the pooled sample covariance matrix of the form

$$(N-2)^{-1} \left[\sum_{m=1}^{N_1} (\mathbf{x}_m - \bar{\mathbf{x}}_1)(\mathbf{x}_m - \bar{\mathbf{x}}_1)^T + \sum_{m=N_1+1}^{N} (\mathbf{x}_m - \bar{\mathbf{x}}_2)(\mathbf{x}_m - \bar{\mathbf{x}}_2)^T \right] \quad (5)$$

which is an unbiased estimator of Σ. Wald [64] in 1944 proved the consistency of (4) for a non-degenerate matrix Σ as $N_1 \to \infty$ and $N_2 \to \infty$.

However, this procedure often fails in applications. The inverse sample covariance matrix is often ill-conditioned or does not exist in dependence on data. The degeneration can occur even for $n = 2$; for $n > N$, the inverse matrix C^{-1} certainly does not exist. Theoretical recommendations only advise us to reduce the dimension in the hope of obtaining a stable solution. In applied problems, some heuristic regularization methods are used. A simple regularization rule is to add a positive quantity to the diagonal of sample covariance matrices before the inversion (di Pillo, 1979). Such estimators of the inverse covariance matrices were called 'ridge'-estimators [1,2]. However, until recently the effect of such regularization was not investigated accurately.

In 1967 A.N. Kolmogorov was interested in the dependence of the probability of error on the sample sizes. He set and solved the following problem. Suppose the matrix Σ is the identity. Let us consider a simplified discriminant function

$$w^1(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \big(\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2\big).$$

This function is distributed normally, and this leads to the error probabilities of the form $\Phi(-G^2/D)$, where random G and D are quadratic functions of sample means and have a non-central χ^2-distribution. To single out the leading parts of G and D, A.N. Kolmogorov offered to consider not a single n-dimensional problem but a sequence $\mathfrak{P} = \{\mathfrak{P}_n\}$ of n-dimensional discriminant problems

$$\mathfrak{P}_n = \big(\mathfrak{S}_1, \mathfrak{S}_2, N_1, N_2, \mathfrak{X}_1, \mathfrak{X}_2, w^1(\mathbf{x}), \alpha_1, \alpha_2\big)_n, \qquad n = 1, 2, \dots,$$

(we do not write out the subscripts n for arguments of $\mathfrak{P}_n$) in which the observations $\mathbf{x}$ are classified with the discriminant function $w(\mathbf{x})$ calculated over samples $\mathfrak{X}_1$ and $\mathfrak{X}_2$ of size N_1 and N_2 from populations $\mathfrak{S}_1$ and $\mathfrak{S}_2$; α_1 and α_2 are probabilities of errors calculated for fixed samples. He assumed that ratios n/N_ν tend to some constants $\lambda_\nu > 0$ as $n \to \infty$. This asymptotic approach was called the 'Kolmogorov asymptotics', or the 'increasing dimension asymptotics' (see in [2]). Supposing that $\mathfrak{S}_\nu = \mathbf{N}(\mu_\nu, I)$, $\quad \nu = 1, 2, \quad w(\mathbf{x}) = w^1(\mathbf{x})$

with $c = 0$ for each n, and $(\mu_1 - \mu_2)^2 \to J_0 \geqslant 0$ as $n \to \infty$, he found that

$$\alpha_1 \to \Phi\left(-\frac{J_0}{2\sqrt{J_0 + \lambda_1 + \lambda_2}}\right) \tag{6}$$

in probability (the limit of α_2 is identical). This expression is remarkable by an explicit dependence of the error probability on the dimension and sample sizes.

In 1976 L.D. Meshalkin [28] deduced the same expression for the limit probabilities of discrimination error for populations different from normal ones under an assumption that the populations are approaching each other in a parametric space (the contiguity assumption) for independent components of the observation vector. In [27] this result was generalized to a wide class of densities given parametrically. In [71] it was shown that the same expression of the limit errors also remains valid for the dependent normal variables with some known structure of the dependence, when the inverse sample covariance matrix of a special structure is used.

In 1970 Yu.N. Blagoveschenskii and A.D. Deev [12] studied the error probabilities for the standard discriminant procedure using the increasing dimension asymptotics for two normal populations with identical unknown covariance matrices. A.D. Deev [14] used the fact that the probabilities of errors (3) coincide with the distribution functions of $w(\mathbf{x})$ for the argument c. He obtained the exact asymptotic expansion for the limit of the expected $\alpha_1 = \alpha_2$ with $c = 0$. The leading term of this expansion proved to be of a special interest. Let us cite it.

Suppose that in a sequence $\mathfrak{P} = \{\mathfrak{P}_n\}$ with the sample discriminant functions (4) the discrimination rule is $w(\mathbf{x}) > c$ against $w(\mathbf{x}) \leqslant c$.

THEOREM 1 (corollary of [14]). *Let $\mathfrak{P}$ satisfy the following conditions:*

(A) *For each n the sets are normal $\mathbf{N}(\mu_\nu, \Sigma)$, $\nu = 1, 2$, with a common non-degenerate covariance matrix Σ.*

(B) *The limit exists* $\lim\limits_{n\to\infty} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = J$.

(C) *In $\mathfrak{P}$, $n/N_\nu \to \lambda_\nu > 0$, $\nu = 1, 2$,*
and the quantity $\lambda \stackrel{\text{def}}{=} \lambda_1 \lambda_2 / (\lambda_1 + \lambda_2) < 1$.