SPRINGER BRIEFS IN STATISTICS
JSS RESEARCH SERIES IN STATISTICS

Yasunori Fujikoshi
Vladimir V. Ulyanov

Non-Asymptotic Analysis of Approximations for Multivariate Statistics

Japan Statistical Society
since 1931 日本統計学会

Springer

SpringerBriefs in Statistics

JSS Research Series in Statistics

The current research of statistics in Japan has expanded in several directions in line with recent trends in academic activities in the area of statistics and statistical sciences over the globe. The core of these research activities in statistics in Japan has been the Japan Statistical Society (JSS). This society, the oldest and largest academic organization for statistics in Japan, was founded in 1931 by a handful of pioneer statisticians and economists and now has a history of about 80 years. Many distinguished scholars have been members, including the influential statistician Hirotugu Akaike, who was a past president of JSS, and the notable mathematician Kiyosi Itô, who was an earlier member of the Institute of Statistical Mathematics (ISM), which has been a closely related organization since the establishment of ISM. The society has two academic journals: the Journal of the Japan Statistical Society (English Series) and the Journal of the Japan Statistical Society (Japanese Series). The membership of JSS consists of researchers, teachers, and professional statisticians in many different fields including mathematics, statistics, engineering, medical sciences, government statistics, economics, business, psychology, education, and many other natural, biological, and social sciences. The JSS Series of Statistics aims to publish recent results of current research activities in the areas of statistics and statistical sciences in Japan that otherwise would not be available in English; they are complementary to the two JSS academic journals, both English and Japanese. Because the scope of a research paper in academic journals inevitably has become narrowly focused and condensed in recent years, this series is intended to fill the gap between academic research activities and the form of a single academic paper. The series will be of great interest to a wide audience of researchers, teachers, professional statisticians, and graduate students in many countries who are interested in statistics and statistical sciences, in statistical theory, and in various areas of statistical applications.

More information about this subseries at http://www.springer.com/series/13497

Yasunori Fujikoshi · Vladimir V. Ulyanov

Non-Asymptotic Analysis of Approximations for Multivariate Statistics

Springer

Yasunori Fujikoshi
Hiroshima University
Higashi-Hiroshima, Japan

Vladimir V. Ulyanov
National Research University
Higher School of Economics
Moscow State University
Moscow, Russia

ISSN 2191-544X ISSN 2191-5458 (electronic)
SpringerBriefs in Statistics
ISSN 2364-0057 ISSN 2364-0065 (electronic)
JSS Research Series in Statistics
ISBN 978-981-13-2615-8 ISBN 978-981-13-2616-5 (eBook)
https://doi.org/10.1007/978-981-13-2616-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This book provides readers with recent non-asymptotic results for approximations in multivariate statistical analysis. There are many traditional multivariate methods based on large-sample approximations. Furthermore, in recent years more high-dimensional multivariate methods have been proposed and utilized for cases where the dimension p of observations is comparable with the sample size n or even exceeds it. Related to this, there are also many approximations under high-dimensional frameworks when $p/n \to c \in (0, 1)$ or $(0, \infty)$.

An important problem related to multivariate approximations concerns their errors. Most results contain only so-called order estimates. However, such error estimates do not provide information on actual errors for given values of n, p, and other parameters. Ideally, we need non-asymptotic or computable error bounds that relate to these actual errors, in addition to order estimates. In non-asymptotic bounds, the pair (n, p), as well as other problem parameters, are viewed as fixed, and statistical statements such as tail or concentration probabilities of test statistics and estimators are constructed as a function of them. In other words, these results are applied for actual values of (n, p). In general, non-asymptotic error bounds involve an absolute constant. If the absolute constant is known, then such an error bound is called the computable error bound.

Our book focuses on non-asymptotic bounds for high-dimensional and large-sample approximations. A brief explanation of non-asymptotic bounds is given in Chap. 1. Some commonly used notations are also explained in Chap. 1. Chapters 2–6 deal with computable error bounds. In Chap. 2, the authors consider computable error bounds on scale-mixed variables. The results can be applied to asymptotic approximations of t- and F-distributions, and to various estimators. In Chap. 3, error bounds for MANOVA tests are given based on large-sample results for multivariate scale mixtures. High-dimensional results are also given. In Chap. 4, the focus is on linear and quadratic discriminant contexts, with error bounds for location and scale mixture variables. In Chaps. 5 and 6, computable error bounds for Cornish–Fisher expansions and Λ-statistics are considered, respectively.

Next, in Chaps. 7–11, new directions of research on non-asymptotic bounds are discussed. In Chap. 7, the focus is on high-dimensional approximations for bootstrap procedures in principal component analysis. Then, in Chap. 8 we consider the Kolmogorov distance between the probabilities of two Gaussian elements to hit a ball in Hilbert space. In Chap. 9, the focus is on approximations of statistics based on observations with random sample sizes. In Chap. 10, the topic is large-sample approximations of power-divergence statistics including the Pearson chi-squared statistic, the Freeman–Tukey statistics, and the log-likelihood ratio statistic. Finally, Chap. 11 proposes a general approach for constructing non-asymptotic estimates and provides relevant examples for several complex statistics.

This book is intended to be used as a reference for researchers interested in asymptotic approximations in multivariate statistical analysis contexts. It will also be useful for instructors and students of graduate-level courses as it covers important foundations and methods of multivariate analysis.

For many approximations, detailed derivations would require a lot of space. For the sake of brevity and presentation, we therefore mainly give their outline. We believe and hope that the book will be useful for stimulating future developments in non-asymptotic analysis of multivariate approximations.

We are very grateful to our colleagues R. Shimizu, F. Götze, G. Christoph, V. Spokoiny, H. Wakaki and A. Naumov to be our co-authors for years. Our joint works are widely used in the book. We express our sincere gratitude to Prof. Naoto Kunitomo, Meiji University, Tokyo, and Dr. Tetsuro Sakurai, Suwa University of Science, Nagano, for their valuable comments on various aspects of the content of this book. We are also grateful to Mr. Y. Hirachi for his assistance in the preparation of this book.

The research was partially supported by the Ministry of Education, Science, Sports, and Culture through a Grant-in-Aid for Scientific Research (C), 16K00047, 2016–2018, and results of Chapters 8 and 9 have been obtained under support of the RSF Grant No. 18-11-00132. The study was done within the framework of the Moscow Center for Fundamental and Applied Mathematics, Moscow State University and HSE University Basic Research Programs.

Hiroshima, Japan Yasunori Fujikoshi
Moscow, Russia Vladimir V. Ulyanov
April 2020

Contents

Chapter 1
Non-Asymptotic Bounds

Abstract Most asymptotic errors in statistical inference are based on error estimates when the sample size n and the dimension p of observations are large. More precisely, such statistical statements are evaluated when n and/or p tend to infinity. On the other hand, "non-asymptotic" results are derived under the condition that n, p, and the parameters involved are fixed. In this chapter, we explain non-asymptotic error bounds, while giving the Edgeworth expansion, Berry–Essen bounds, and high-dimensional approximations for the linear discriminant function.

1.1 Errors and Non-Asymptotic Bounds

In mathematical statistics and probability theory, asymptotics are used in analysis of long-run or large-sample and high-dimensional behavior of random variables related to various statistical inference contexts. Most asymptotic theory is based on results when the sample size n and the dimension p of observations tend to infinity. Often, such results do not provide information that is relevant in situations where the sample size and dimension are given finite values. To appreciate this more precisely, consider Edgeworth-type expansion. Let $\{F_n(x)\}$ be a set of distribution functions indexed by a parameter n, typically, sample size. For example, suppose $F_n(x)$ is approximated by the first k terms of asymptotic expansion of the Edgeworth type:

$$G_{k,n}(x) = G(x) + \sum_{j=1}^{k-1} n^{-j/2} p_j(x) g(x), \tag{1.1}$$

where $G(x)$ is the limiting distribution function of $F_n(x)$, $g(x)$ is the pdf of $G(x)$, and the $p_j(x)$'s are suitable polynomials. For a wide class of statistics, it is known (see, e.g., [1]) that the error $R_{k,n}(x) = F_n(x) - G_{k,n}(x)$ satisfies the order condition

$$R_{k,n}(x) = \mathrm{O}(n^{-k/2}), \quad \text{uniformly in } x. \tag{1.2}$$

Y. Fujikoshi and V. V. Ulyanov, *Non-Asymptotic Analysis of Approximations for Multivariate Statistics*, JSS Research Series in Statistics,
https://doi.org/10.1007/978-981-13-2616-5_1

Note that this statement is rather implicit because it does not give any information about the largest possible value of $R_{k,n}(x)$ for a given n. More precisely, condition (1.2) means that there exists a positive constant C_k and a positive number N_k such that the inequality

$$\sup_x |R_{k,n}(x)| \le C_k \big/ n^{k/2} \quad \text{for all } n \ge N_k \tag{1.3}$$

holds. However, the values of C_k and N_k are unknown. In some cases, the situation becomes even worse: F_n may depend on a "nuisance" parameter, say, $\theta \in \Theta$. Then C_k and N_k depend on θ as well, i.e., $C_k = C_k(\theta)$ and $N_k = N_k(\theta)$, which are, again, totally unknown except that they are finite, or at best, that they are monotone in θ.

Our knowledge about the error bound will improve if we can fix a number C_k for a given N_k, possibly depending on θ, in such a way that inequality (1.3) is true. We call this kind of error bound a computable error bound or a non-asymptotic error bound. In general, the term "computable error bound" means that the error can be numerically computed when (n, p) and the values of parameters are given.

One of the main problems associated with approximations in multivariate statistical analysis concerns their errors. Most results only supply so-called order estimates. Such estimates do not provide information on actual errors for given values of sample size n, dimension p, and other parameters. It is desirable to develop non-asymptotic results whose statements also hold for fixed values of (n, p) and other problem parameters. Therefore, ideally, we want to have non-asymptotic bounds or computable error bounds, in addition to order estimates.

The most well-known computable error bound is the Berry–Esseen theorem, or the bound in the central limit theorem. For expository purposes we consider the case when $X_1, X_2, \ldots$ are i.i.d. random variables. Without loss of generality, we assume that $\mathrm{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) = 1$. Let $\beta_3 = \mathrm{E}(|X_1|^3)$ be finite and let $F_n(x)$ and $\Phi(x)$ be the distribution functions of the normalized sum $(X_1 + \cdots + X_n)/\sqrt{n}$ and the standard normal distribution, respectively. Then, it is known that there exists a positive constant C such that for all n

$$\sup_x \left| F_n(x) - \Phi(x) \right| \le C\beta_3/\sqrt{n}. \tag{1.4}$$

There are many works on seeking a value of C. It is known [5] that $C \le 0.4748$ for all $n \ge 1$. In terms of error bounds for its first-order expansion, see [2], for example. However, these topics are beyond the scope of this book.

As another non-asymptotic result, consider misclassification errors in two-group discriminant analysis. Suppose we are interested in classifying a $p \times 1$ observation vector $\boldsymbol{X}$ as coming from one of two populations Π_1 or Π_2. Let Π_i be a two p-variate normal population with $\mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_1 \ne \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ is positive definite. When the values of the parameters are unknown, assume that random samples of sizes n_1 and n_2 are available from Π_1 and Π_2, respectively. Let $\bar{\boldsymbol{X}}_1$, $\bar{\boldsymbol{X}}_2$, and $\mathbf{S}$ be the sample mean vectors and the sample covariance matrix. Then, a well-known linear discriminant function is defined by $W = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)'\mathbf{S}^{-1}\left\{\boldsymbol{X} - \frac{1}{2}(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2)\right\}$. The observation

$\boldsymbol{X}$ may be classified to Π_1 or Π_2 according to $W \geq 0$ or $W < 0$. We are interested in the misclassification probabilities given by $e_W(2|1) = \mathrm{P}(W \leq 0 \mid \boldsymbol{X} \in \Pi_1)$ and $e_W(1|2) = \mathrm{P}(W \leq 0 \mid \boldsymbol{X} \in \Pi_2)$. Under a high-dimensional asymptotic framework $n_1 \to \infty$, $n_2 \to \infty$, $m = n_1 + n_2 - p \to \infty$, it is known (see, e.g., [3, 4]) that $e_W(2|1)$ converges to $\Phi(\gamma)$, and

$$|e_W(2|1) - \Phi(\gamma)| \leq B(p, n_1, n_2, \Delta^2), \tag{1.5}$$

where $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the square of the Mahalanobis distance. For closed forms of γ and $B(p, n_1, n_2, \Delta^2)$, see [4]. Similar results have been obtained for $e_W(1|2)$. Note that result (1.5) holds for any given (p, n_1, n_2, Δ^2) such that $m > 7$; thus it is a non-asymptotic result. Further, $B(p, n_1, n_2, \Delta^2)$ is a computable error bound of order $\mathrm{O}(m^{-1})$. Similar results are given for the quadratic discriminant function. In Chap. 4, we show such results by extending non-asymptotic theory for a location and scale mixture.

In general, approximation errors can be described either asymptotically as an order of a remainder term with respect to a sample size n and/or a dimension p, or non-asymptotically as a bound for a remainder term with clearly expressed dependence on n, moment characteristics, and p. In this book, we deal with the latter non-asymptotic results and try to determine the correct structure of the bound for a remainder term in relation to all parameters involved, with the exception of the absolute constants. If we can give the values of absolute constants as well, then the corresponding bounds are called computable. Such computable bounds were considered for some multivariate statistics in the earlier book by [4]. In this new book, first we consider computable error bounds on scale-mixed variables, MANOVA tests, discriminant functions, Cornish–Fisher expansions, and some other statistics. These considerations are based on the recent findings, where applicable.

After the book by Fujikoshi et al. [4] was published by Wiley in 2010, a series of new non-asymptotic results appeared. Research continued in new directions, in particular, for bootstrap procedures, for approximations of statistics based on observations with random sample sizes, and for power-divergence statistics. These results are presented in this book. Further, in Chap. 11, we also suggest a general approach for constructing non-asymptotic bounds and provide corresponding examples for relatively complex statistics.

In statistics, there are many asymptotic results including asymptotic expansions as n and/or p tend to infinity. Some of them are given without rigorous proofs for their error terms. Non-asymptotic results will also be useful as rigorous proofs for the order estimates of such approximation errors.

Here, we note that some commonly used notations are used in this book without defining them in detail. For example, the transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}'$ or $\mathbf{A}^\mathsf{T}$. For a squared matrix $\mathbf{A}$, its determinant and trace are denoted by $|\mathbf{A}|$ and $\mathrm{tr}\mathbf{A}$, respectively. The usual norm $(\boldsymbol{a}'\boldsymbol{a})^{1/2}$ of a vector $\boldsymbol{a}$ is denoted by $\|\boldsymbol{a}\|$ or $\|\boldsymbol{a}\|_2$. For a random vector $\boldsymbol{X}$, its mean is denoted by $\mathrm{E}(\boldsymbol{X})$. The covariance matrix is denoted by $\mathrm{Var}(\boldsymbol{X})$ or $\mathrm{Cov}(\boldsymbol{X})$.

References

1. Bhattacharya, R. N., & Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *The Annals of Statistics*, *6*, 434–451. (corrigendum ibid., **8**, 1980).
2. Dobrić, V., & Ghosh, B. K. (1996). Some analogs of the Berry-Esseen bound for first-order Chebyshev–Edgeworth expansions. *Statistics & Decision*, *14*, 383–404.
3. Fujikoshi, Y. (2000). Error bounds for asymptotic approximations of the linear discriminant function when the sample size and dimensionality are large. *Journal of Multivariate Analysis*, *73*, 1–17.
4. Fujikoshi, Y., Ulyanov, V. V., & Shimizu, R. (2010). *Multivariate Analysis: High-Dimensional and Large-Sample Approximations*. Hoboken: Wiley.
5. Shevtsova, I. G. (2011). On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands. arXiv: 1111.6554.