THE DEVELOPMENT OF HIGH-DIMENSIONAL STATISTICS

Today the humanity developed facilities to create and analyze informational models of high complexity including models of biosystems, visual patterns, and of natural language. Modern computers easily treat information arrays that are comparable with the total life experience (near 10^{10} bits). Now objects of statistical investigation are often characterized by very large number of parameters whereas, in practice, sample data are rather restricted. For such statistical problems, values of separate parameters are usually of a small interest, and the purpose of investigation is displaced to finding optimal statistical decisions.

Some examples.

1. Statistical analysis of biological and economic objects.

These objects are characteristic by a great complexity and a considerable nuisance along with bounded samples. Their models depend on a great number of parameters, and the standard approach of mathematical statistics based on expansion in the inverse powers of sample size does not account for the problem specificity. In this situation, another approach proposed by A.N.Kolmogorov seems to be more appropriate. He introduced an asymptotics in which the dimension n tends to infinity along with sample size N allowing to analyze effects of inaccuracies accumulation in estimating a great number of parameters.

2. Pattern recognition.

Today we must acknowledge that the recognition of biological and economic objects requires not so much data accumulation, as the extraction of regularities and elements of structure against the noise background. These structure elements are then used as features for recognition. But the variety of possible structure elements is measured by combinatorial large numbers and the new mathematical problem arises of efficient discriminant analysis in space of high dimension.

3. Interface with computer using natural language.

This problem seems to become the central problem of our age. It is well known that the printed matter containing the main part of classical literature requires rather moderate computer resources. For example, a full collection of A.S.Pushkin's compositions occupies only 2–5 megabytes, while the main corpus of Russian literature can be written on one gigabyte disk. The principle problem to be solved is how to extract the meaning from the text. Identifying the meaning of texts with new information and measuring it with the Shannon measure, we can associate the sense of a phrase with the statistics of repeating words and phrases in the language experience of a human. This sets a problem of developing a technology of search for repeating fragments in texts of a large volume. A specific difficulty is that the number of repetitions may be far from numerous: indeed, the human mind would not miss even a single coincidence of phrases.

Traditionally, the statistical investigation is related to a cognition process, and according to the R.Fisher conception, the purpose of statistical analysis is to determine parameters of an object in the process of analyzing more and more data. This conception is formalized in the form of an asymptotics of sample size increasing indefinitely which lays in the foundation of well-developed theory of asymptotic methods of statistics. The most part of investigation in mathematical statistics deal with one-dimensional observations and fixed number of parameters under arbitrarily large sample sizes. The usual extension to many-dimensional case is reduced to the replacement of scalars by vectors and matrices and to studying formal relations with no insight to underlying phenomena.

The main problem of mathematical statistics today remains to study the consistency of estimators and their asymptotic properties under increasing sample size. Until recently no fruitful approach existed to the problem of quality estimation of the statistical procedures under fixed samples. It was only established that nearly all popular statistical methods allow improvement and must be classified as inadmissible. In many-dimensional statistics this conclusion is much more severe: nearly all consistent multivariate linear procedures may have infinitely large values of risk function. These estimators should be called "essentially inadmissible".

Meanwhile, we must acknowledge that today the state of methods of multivariate statistical analysis is far from satisfactory. Most popular linear procedures require the inversion of covariance matrix. True inverse covariance matrices are replaced by consistent estimators. But sample

covariance matrices (dependently on data) may be degenerate and their inversion can be impossible (even for the dimension 2). For large dimension the inversion of sample covariance becomes unstable and that leads to insignificant statistical inferences of no significance. If the dimension is larger than sample size, sample covariance matrices are surely degenerate and their inversion is impossible. As a consequence, standard consistent procedures of multivariate statistical analysis included into most of packages of statistical software do not guarantee neither stable, nor statistically significant results, and often prove to be inapplicable. Common researchers applying methods of multivariate statistical analysis to their concrete problems are left without theoretical support in front of their difficulties. The existing theory cannot recommend them nothing better as to ignore a part of data artificially reducing the dimension in hope that this would provide a plausible solution (see [3]).

This book presents the development of a new special branch of mathematical statistics applicable to the case when the number of unknown parameters is large. Fortunately, in case of a large number of boundedly dependent variables it proves to be possible to use specifically many-parametric regularities for the construction of improved procedures. These regularities include small variance of standard quality functions, the possibility to estimate them reliably from sample data, to compare statistical procedures by their efficiency and choose better ones. Mathematical theory developed in this book offers a number of more powerful versions of most usable statistical procedures providing solutions that are both reliable and approximately unimprovable in the situation when the dimension of data is comparable in magnitude with sample. The statistical analysis appropriate for this situation may be qualified as *the essentially multivariate analysis* [69]. The theory that takes into account effects produced by the estimation of a large number of unknown parameters may be called *the multiparametric statistics.*

The first discovery of the existence of specific phenomena arising in multiparametric statistical problems was the fact that standard sample mean estimator proves to be inadmissible, that is, its square risk can be diminished.

The Stein Effect

In 1956 C.Stein noticed that sample mean is not a minimum square risk estimator, and it can be improved by multiplying by a scalar decreasing the length of the estimation vector. This procedure was called "shrinkage", and such estimators were called "shrinkage estimators". The effect of improving estimators by shrinkage was called the "Stein effect". This effect was fruitfully exploited in applications (see [29], [33], [34]). Let us cite the well-known theorem by James and Stein.

Denote by $\mathbf{x}$ be an n-dimensional observation vector, and let $\bar{\mathbf{x}}$ denote sample mean calculated over a sample of size N. Denote (here and below) by I the identity matrix.

PROPOSITION 1. *For $n > 2$ and $\mathbf{x} \sim \mathbf{N}(\vec{\mu}, I)$, the estimator*

$$\hat{\mu}^{JS} = \left(1 - \frac{n-2}{N\bar{\mathbf{x}}^2}\right)\bar{\mathbf{x}} \tag{1}$$

has the quadratic risk

$$\mathbf{E}\ (\vec{\mu} - \hat{\mu}^{JS})^2 = \mathbf{E}\ (\vec{\mu} - \bar{\mathbf{x}})^2 - \left(\frac{n-2}{N}\right)^2\ \mathbf{E}\ \frac{1}{\bar{\mathbf{x}}^2}. \tag{2}$$

(*here and in the following, squares of vectors denote squares of their length*).

Proof. Indeed,

$$R^{JS} = \mathbf{E}\ (\vec{\mu} - \hat{\mu}^{JS})^2 = y + 2y_2\ \mathbf{E}\ \frac{(\vec{\mu} - \bar{\mathbf{x}})^T\bar{\mathbf{x}}}{\bar{\mathbf{x}}^2} + y_2^2\ \mathbf{E}\ \frac{1}{\bar{\mathbf{x}}^2}, \tag{3}$$

where $y = n/N$ and $y_2 = (n-2)/N$. Let f be the normal distribution density for $\mathbf{N}(\vec{\mu}, I/N)$. Then $\vec{\mu} - \bar{\mathbf{x}} = (Nf)^{-1}\nabla f$, where ∇ is the differentiation operator in components of $\bar{\mathbf{x}}$. Substitute this expression in the second addend of (3) and note that the expectation can be calculated by the integration in $f\ d\bar{\mathbf{x}}$. Integrating by parts we obtain (2). □

The James–Stein estimator is known as a "remarkable example of estimator inadmissibility".

This discovery produced the development of a new direction of investigation and a new trend in theoretical and applied statistics with hundreds of publications and effective applications [33].

In subsequent years other versions of estimators were offered that improved as standard sample mean estimator as the James–Stein estimator. The first improvement was offered in 1963 by Baranchik [10]. He proved that the quadratic risk of the James–Stein estimator can be decreased by excluding negative values of the shrinkage estimator ("positive-part shrinkage"). Other numerous shrinkage estimators were proposed subsequently decreasing the quadratic risk one after the other (see [31]).

However, the James–Stein estimator is singular for small $\bar{\mathbf{x}}^2$. In 1999 Das Gupta and Singh [30] offered a robust estimator

$$\hat{\mu}^G = \left(1 - \frac{n}{n + \bar{\mathbf{x}}^2}\right) \bar{\mathbf{x}},$$

that for $n \geq 4$ dominates $\hat{\mu} = \bar{\mathbf{x}}$ with respect to the quadratic risk and dominates $\hat{\mu}^{JS}$ with respect to the absolute risk $\mathbf{E}\,|\vec{\mu} - \hat{\mu}^G|$ (here and in the following the absolute value of a vector denote its length). In 1964 C.Stein suggested an improved estimator for $x \sim \mathbf{N}(\vec{\mu}, dI)$ with unknown $\vec{\mu}$ and d. Later a series of estimators were proposed subsequently improving his estimator (see [42]). A number of shrinkage estimators were proposed for the case $\mathbf{N}(\vec{\mu}, \Sigma)$ (see [49]).

The shrinkage was also applied in the interval estimation. Using shrinkage estimator, Cohen [15] has constructed confidence intervals that have the same length but are different by a uniformly greater probability of covering. Goutis and Gasela [27] proposed other confidence intervals improved with respect to the interval length and with respect to the covering probability as well. These results were extended to many-dimensional normal distributions with unknown expectations and unknown covariance matrix.

The Stein effect was discovered also for distributions different from normal. In 1979 Brandwein has shown that for spherically symmetrical distributions with the density $f(|\mathbf{x} - \theta|)$, where θ is a vector parameter, for $n > 3$ the estimator $\hat{\mu} = (1 - a/\bar{\mathbf{x}}^2)\ \bar{\mathbf{x}}$ dominates $\hat{\mu} = \bar{\mathbf{x}}$ for a such that $0 < a < a_{\max}$. For these distributions other improved shrinkage estimators were found (see the review by Brandwein and Strawdermann [13]).

For the Poisson distribution of independent integer-valued variables k_i, $i = 1, 2, \ldots, n$ with the vector parameter $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, the standard unbiased estimator is the vector of rates $(f_1, f_2, \ldots, f_n)$ of

events number $i = 1, 2, \ldots, n$ in the sample. Clevenson and Zedek [14] showed that the estimator $\hat{\lambda}$ of the form

$$\hat{\lambda}_i = \left(1 - \frac{a}{a+s}\right) f_i, \quad i = 1, 2, \ldots, n \qquad \text{where } s = \sum_{i=1}^{n} f_i \tag{4}$$

for $n > 2$ has the quadratic risk less than the vector $(f_1, f_2, \ldots, f_n)$, if $n - 1 \leq a \leq 2\,(n-1)$. A series of estimators were found improving the estimator (4).

Statistical meaning of shrinkage.

For the understanding of the mechanism of risk reduction it is useful to consider the expectation of the sample average square. For distributions with the variance d of all variables we have $\mathbf{E}\ \bar{\mathbf{x}}^2 = \vec{\mu}^2 + yd > \vec{\mu}^2$, where $y = n/N$, and, naturally, one can expect that shrinkage of sample average vectors may be useful. The shrinkage effect may be characterized by the magnitude of the ratio $\vec{\mu}^2/yd$. This ratio may be interpreted as the "signal-to-noise" ratio. The shrinkage is purposeful for sufficiently small $\vec{\mu}^2/yd$. For $d = 1$ and restricted dimension, $y \approx 1/N$, and shrinkage is useful only for the vector length less than $1/\sqrt{N}$. For essentially many-dimensional statistical problems with $y \approx 1$, the shrinkage can be useful only for bounded vector length when $\vec{\mu}^2 \approx 1$, and its components have the order of magnitude $1/\sqrt{N}$. This situation is characteristic for a number of statistical problems, in which vectors $\vec{\mu}$ are located in a bounded region. The important example of these is high-dimensional discriminant analysis in the case when the success can be achieved only by taking account of a large number of weakly discriminating variables.

Let it be known a priori that the vector $\vec{\mu}$ is such that $\vec{\mu}^2 \leq c$. In this case it is plausible to use the shrinkage estimator $\hat{\mu} = \alpha\bar{\mathbf{x}}$ with the shrinkage coefficient $\alpha = c/(c+y)$. The quadratic risk

$$R(a) = \mathbf{E}\ (\vec{\mu} - a\bar{\mathbf{x}})^2 = y\ \frac{\vec{\mu}^2 y + c^2}{(c+y)^2} \leq y\ \frac{c}{c+y} < R(1).$$

It is instructive to consider the shrinkage effect for simplest shrinkage with non-random shrinkage coefficients. Let $\hat{\mu} = \alpha\bar{\mathbf{x}}$, where non-random positive $\alpha < 1$. For $\mathbf{x} \sim \mathbf{N}(\vec{\mu}, I)$, the quadratic risk of this "a priori" estimator is

$$R = R(\alpha) = (1-\alpha)^2\vec{\mu}^2 + \alpha^2 y, \tag{5}$$

$y = n/N$. The minimum of $R(\alpha)$ is achieved for $\alpha = \alpha^0 = \vec{\mu}^2/(\vec{\mu}^2 + y)$ and is equal to

$$R^0 = R(\alpha^0) = y\vec{\mu}^2/(\vec{\mu}^2 + y). \tag{6}$$

Thus the standard quadratic risk y is multiplied by the factor $\vec{\mu}^2/(\vec{\mu}^2+y) < 1$. In traditional applications, if the dimension is not high, the ratio y is of order of magnitude $1/N$, and the shrinkage effect proves to be insignificant even for a priori bounded localization of parameters. However, if the accuracy of measurements is low and the variance of variables is so large that it is comparable with N, the shrinkage can considerably reduce the quadratic risk.

The shrinkage "pulls" estimators down to the coordinate origin; this means that the shrinkage estimators are not translation invariant. The question arises of their sensitivity to the choice of coordinate system and of the origin. In an abstract setting, it is quite not clear how to choose the coordinate center for "pulling" of estimators. The center of many-dimensional population may be located, generally speaking, at any faraway point of space and the shrinkage may be quite not efficient. However, in practical problems, some restrictions always exist on the region of parameter localization and there is some information on the central point. As a rule, the practical investigator knows in advance the region of the parameters localization. In view of this, it is quite obvious that the standard sample mean estimator must be improvable, and it may be improved, in particular, by shrinkage. Note that this reasoning has not attracted a worthy attention of researches as yet and this fact leads to the mass usage of the standard estimator in problems, where the quality of estimation could be obviously improved.

It is natural to expect that as much as the shrinkage coefficient in the James–Stein estimator is random, it can decrease the quadratic risk less efficiently than the best non-random shrinkage. Compare the quadratic risk R^{JS} of the James–Stein estimator (2) with the quadratic risk R^0 of the best a priori estimator (6).

PROPOSITION 2. *For n-dimensional observations $\mathbf{x} \sim \mathbf{N}(\vec{\mu}, I)$ with $n > 2$ we have*

$$R^{JS} \le R^0 + 4\,\frac{n-1}{N^2}\,\frac{1}{\vec{\mu}^2+y} \le R^0 + 4/N.$$

Proof. We start from Proposition 1. Denote $y_2 = (n-1)/N$. Using the properties of moments of inverse random values we find that

$$R^{JS} = y - y_2^2 \mathbf{E}\,(\bar{\mathbf{x}}^2)^{-1} \le y - y_2^2\,(\mathbf{E}\,\bar{\mathbf{x}}^2)^{-1} = R^0 + 4(n-1)N^{-2}/(\vec{\mu}^2+y).$$

□

Thus, for large N the James–Stein estimator practically is not worse than the unknown best a priori shrinkage estimator that may be constructed if the length of the vector is known.

Application in the regression analysis.
Consider the regression model

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathbf{N}(0, I),$$

where X is a non-random rectangular matrix of size $N \times n$ of a full rank, $\beta \in \mathbb{R}^n$, and I is the $n \times n$ identity matrix. The standard minimum square solution leads to the estimator $\hat{\beta}_0 = (X^T X)^{-1} X^T Y$, that is used in applied problems and included into most of applied statistics software. The effect of application of the James–Stein estimator for shrinking of vectors $\hat{\beta}_0$ was studied in [29]. Let the (known) plan matrix is such that $X^T X$ is the identity matrix. Then the problem of construction of regression model of best quality (in the meaning of minimum sum of residual squares) is reduced to estimation of the vector $\beta = \mathbf{E}\, X^T Y$ with the minimum square risk. The application of the Stein-type estimators allows to choose better versions of linear regression (see [33], [84]).

This short review shows that the fundamental problem of many-dimensional statistics - estimation of the position of the center of population is far from being ultimately solved. The possibility of improving estimators by shrinking attracts our attention to the improvement of solutions to other statistical problems.

Chapter 2 of this book presents an attempt of systematical advance in the theory of improving estimators of expectation vectors of large dimension.

In Section 2.1 the generalized Stein-type estimators are studied in which shrinkage coefficients are arbitrary functions of sample mean vector length. The boundaries of the quadratic risk decrease are found. In Section 2.2 it is established that in case when the dimension is large and comparable with sample size, shrinking of a wide class of unbiased estimators reduces the quadratic risk independently of distributions. In Section 2.3, the Stein effect is investigated for infinite-dimensional estimators. In Section 2.4 "component-wise" estimators are considered that are defined by arbitrary "estimation functions" presenting some functional transformation of each component of sample mean vector. The quadratic risk of this estimator is minimized with the accuracy to terms small for large dimension and sample size.

The Kolmogorov Asymptotics

In 1967 Andrei Nikolaevich Kolmogorov was interested in the dependence of errors of discrimination on sample size. He solved the following problem. Let $\mathbf{x}$ be a normal observation vector, and $\bar{\mathbf{x}}_\nu$ be sample averages calculated over samples from populations number $\nu = 1, 2$. Suppose the covariance matrix is the identity matrix. Consider a simplified discriminant function

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

and the classification rule $w(\mathbf{x}) > 0$ against $w(\mathbf{x}) \le 0$. This function leads to the probability of errors $\alpha_n = \Phi(-G/\sqrt{D})$, where G and D are quadratic functions of sample averages having a non-central χ^2 distribution. To isolate principal parts of G and D, Kolmogorov proposed to consider not one statistical problem, but a sequence of n-dimensional discriminant problems in which the dimension n increases along with sample sizes N_ν, so that $N_\nu \to \infty$ and $n/N_\nu \to \lambda_\nu > 0$, $\nu = 1, 2$. Under these assumptions he proved that the probability of error α_n converges in probability

$$\operatorname*{plim}_{n\to\infty} \alpha_n = \Phi\left(-\frac{J - \lambda_1 + \lambda_2}{2\sqrt{J + \lambda_1 + \lambda_2}}\right), \tag{7}$$

where J is the square of the Euclidean limit "Mahalanobis distance" between centers of populations. This expression is remarkable by that it explicitly shows the dependence of error probability on the dimension and sample sizes. This new asymptotic approach was called the "Kolmogorov asymptotics".

Later L.D.Meshalkin and the author of this book deduced formula (7) for a wide class of populations under the assumption that the variables are independent and populations approach each other in the parameter space (are contigual) [45], [46].

In 1970 Yu.N.Blagoveshchenskii and A.D.Deev studied the probability of errors for the standard sample Fisher–Andersen–Wald discriminant function for two populations with *unknown common* covariance matrix. A.D.Deev used the fact that the probability of error coincides with the distribution function of $g(\mathbf{x})$. He obtained an exact asymptotic expansion for the limit of the error probability α. The leading term of this expansion proved to be especially interesting. The limit probability of an error (of the first kind) proved to be

$$\alpha = \Phi\left(-\Theta\frac{J - \lambda_1 + \lambda_2}{2\sqrt{J + \lambda_1 + \lambda_2}}\right),$$

where the factor $\Theta = \sqrt{1-\lambda}$, with $\lambda = \lambda_1\lambda_2/(\lambda_1+\lambda_2)$ accounts for the accumulation of estimation inaccuracies in the process of the covariance matrix inversion. It was called "the Deev formula". This formula was thoroughly investigated numerically and a good coincidence was demonstrated even for not great n, N.

Note that starting from Deev's formulas one can easily see that the discrimination errors can be reduced if to use the rule $g(\mathbf{x}) > \theta$ against $g(\mathbf{x}) \leq \theta$ with $\theta = (\lambda_1 - \lambda_2)/2 \neq 0$. A.D.Deev also noticed [18] that the half-sum of discrimination errors can be further decreased by weighting of summands in the discriminant function.

After these investigations it became obvious that by keeping terms of the order of n/N one obtains a possibility of using specifically multidimensional effects for the construction of improved discriminant and other procedures of multivariate analysis. The most important conclusion was that traditional consistent methods of multivariate statistical analysis should be improvable, and a new progress of theoretical statistics is possible oriented to obtaining nearly optimal solutions for fixed samples.

The Kolmogorov asymptotics ("increasing dimension asymptotics", [3]) may be considered as a calculation tool for isolating leading terms in case of large dimension. But the principle role of the Kolmogorov asymptotics is that it reveals specific regularities produced by estimation of a large number of parameters. In a series of further publications, this asymptotics was used as a main tool of investigation of *essentially many-dimensional* phenomena characteristic for *high-dimensional statistical analysis.* The constant n/N became an acknowledged characteristics in many-dimensional statistics.

In Section 5.1 the Kolmogorov asymptotics is applied for the development of theory allowing to improve the discriminant analysis of vectors of large dimension with independent components. The improvement is achieved by introducing appropriate weights of contributions of independent variables in the discriminant function. These weights are used for the construction of asymptotically unimprovable discriminant procedure. Then the problem of selection of variables for discrimination is solved and the optimum selection threshold is found.

But the main success in the development of multiparametric solutions was achieved by combining the Kolmogorov asymptotics with the spectral theory of random matrices developed independently at the end of 20th century in another region.

Spectral Theory of Increasing Random Matrices

In 1955 the well-known physicist theoretician E. Wigner studied energy spectra of heavy nuclei and noticed that these spectra have a characteristic semicircle form with vertical derivatives at the edges. To explain this phenomenon he assumed that very complicated hamiltonians of these nuclei can be represented by random matrices of high dimension. He found the limit spectrum of symmetric random matrices of increasing dimension $n \to \infty$ with independent (over-diagonal) entries W_{ij}, zero expectation and the variance $\mathbf{E}\, W_{ii}^2 = 2v^2$, $\mathbf{E}\, W_{ij} = v^2$ for $i \neq j$ [88]. The empirical distribution function ("counting function") for eigenvalues λ_i of these matrices

$$F_n(u) = n^{-1} \sum_{i=1}^{n} \text{ ind } (\lambda_i \leq u)$$

proved to converge almost surely to the distribution function $F(u)$ with the density

$$F'(u) = (2\pi v^2)^{-1} \sqrt{4v^2 - u^2}, \qquad |u| \leq 2|v|$$

("limit spectral density"). This distribution was called Wigner's distribution.

In 1967 V.A.Marchenko and L.A.Pastur published the well-known paper [43] on the convergence of spectral functions of random symmetric Gram matrices of increasing dimension $n \to \infty$. They considered matrices of the form

$$B = A + N^{-1} \sum_{m=1}^{N} \mathbf{x}_m \mathbf{x}_m^T$$

where A are non-random symmetric matrices with converging counting functions $F_{An}(u) \to F_A(u)$, and $\mathbf{x}_m$ are independent random vectors with independent components x_{mi} such that $\mathbf{E}\, x_{mi} = 0$ and $\mathbf{E}\, x_{mi}^2 = 1$. They assumed that the ratio $n/N \to y > 0$, distribution are centrally symmetric and invariant with respect to components numeration, and the first four moments of x_{mi} satisfy some tensor relations. They established the convergence $F_{Bn}(u) \to F_B(u)$, where $F_{Bn}(u)$ are counting functions for eigenvalues of B, and derived a specific nonlinear relation between limit spectral functions of matrices A and B. In the simplest case when $A = I$ it reads

$$h(t) = \int (1 + ut)^{-1} dF_B(u) = (1 + ts(t))^{-1},$$

where $s(t) = 1 - t + yh(t)$. By the inverse Stilties transformation, they obtained the limit spectral density

$$F'_B(u) = \frac{1}{\sqrt{2\pi yu}} \sqrt{(u_2 - u)(u - u_1)}, \qquad u_1 \le u \le u_2.$$

where $u_2, u_1 = (1 \pm \sqrt{y})^2$. If $n > N$, the limit spectrum has a discrete component at $u = 0$ that equals $1 - N/n$.

In 1975–2001 V.L.Girko created an extended limit spectral theory of random matrices of increasing dimension that was published in a series of monographs (see [22]-[26]). Let us describe in general features some of his results. V.L.Girko studies various matrices formed by linear and quadratic transformations from initial random matrices $X = \{x_{mi}\}$ of increasing dimensions $N \times n$ with independent entries. The aim of his investigation is to establish the convergence of spectral functions of random matrices to some limit non-random functions $F(u)$ and then to establish the relation between $F(u)$ and limit spectral functions of non-random matrices. For example, for $B = A^T X X^T A$ the direct functional relation is established between limit spectra of non-random matrices A and random B. V.L.Girko calls such relations "stochastic canonical equations" (we prefer to call them "dispersion equations").

In the first (Russian) monograph "Random Matrices" published in 1975, V.L.Girko assumes that all variables are independent, spectral functions of non-random matrices converge, and the generalized Lindeberg condition holds: for any $\tau > 0$

$$\lim_{n\to\infty} N^{-1} \sum_{m=1}^{N} \; n^{-1} \sum_{i=1}^{n} \; x_{mi}^2 \text{ ind } (x_{mi}^2 \ge \tau) \xrightarrow{\mathbf{P}} 0.$$

The main result of his investigations in this monograph was a number of limit equations connecting spectral functions of different random matrices and underlying non-random matrices.

In monograph [25] (1995) V.L.Girko applied his theory specifically to sample covariance matrices. He refines his theory by withdrawing the assumption on the convergence of spectral functions of true covariance matrices. He postulates a priori some "canonical" equations, proves their solvability, and only then reveals their connection with limit spectra of random matrices. Then V.L.Girko imposes more restrictive requirements to moments (he assumes the existence of four uniformly bounded moments) and finds limit values of separate (ordered) eigenvalues.

Investigations of other authors in the theory of random Gram matrices of increasing dimension differ by more special settings and by less systematic results. However, it is necessary to cite paper by Q.Yin, Z.Bai, and P. Krishnaia (1984), who were first to establish the existence of limits for the least and the largest eigenvalues of Wishart matrices. In 1998, Z.Bai and J.Silverstein [9] discovered that eigenvalues of increasing random matrices stay within the boundaries of the limit spectrum with probability 1.

Spectral Functions of Sample Covariance Matrices.

Chapter 3 of this book presents the latest development in the spectral theory of sample covariance matrices of large dimension. Methods of spectral theory of random matrices were first applied to sample covariance matrices in paper [63] of the author of this monograph (1983). The straightforward functional relation was found between limit spectral functions of sample covariance matrices with limit spectral functions of unknown true covariance matrices. Let us cite this result since it is of a special importance for the multiparametric statistics. Spectra of true covariance matrices Σ of size $n \times n$ are characterized by the "counting" function

$$F_{0n}(u) = n^{-1} \sum_{m=1}^{n} \text{ ind } (\lambda_i \le u), \quad u \ge 0,$$

of eigenvalues $\lambda_i, \quad i = 1, 2, \dots, n$. Sample covariance matrices are calculated over samples $\mathfrak{X} = \{\mathbf{x}_m\}$ of size N, have the form

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

where $\bar{\mathbf{x}}$ are sample average vectors.

THEOREM 1. *If n-dimensional populations are normal $\mathbf{N}(0, \Sigma)$, $n \to \infty$, $n/N \to \lambda > 0$, and functions $F_{0n}(u) \to F_0(u)$, then for each $t \ge 0$ the limit exist*

$$h(t) = \lim_{n \to \infty} \mathbf{E}\, n^{-1} \text{tr } (I + tC)^{-1} = \int (1 + ts(t)u)^{-1} dF_0(u), \qquad (8)$$

$$\text{and} \quad \mathbf{E}\, (I + tC)^{-1} = (I + ts(t)\Sigma)^{-1} + \Omega_n,$$

where $s(t) = 1 - \lambda + \lambda h(t)$ and $\|\Omega_n\| \to 0$ (here the spectral norms of matrices are used).

In 1995 the author of this book proved that these relations remain valid for a wide class of populations restricted by the values of two specific

parameters: the maximum fourth central moment of a projection of $\mathbf{x}$ onto non-random axes (defined by vectors $\mathbf{e}$ of unit length)

$$M = \sup_{|\mathbf{e}|=1} \mathbf{E}\ (\mathbf{e}^T\mathbf{x})^4 > 0$$

and measures of dependence of variables

$$\nu = \sup_{\|\Omega\|=1} \text{var}\ (\mathbf{x}^T\Omega\mathbf{x}/n), \quad \text{and} \quad \gamma = \nu/M,$$

where Ω are non-random symmetric positive semidefinite matrices of unit spectral norm. Note that for independent components of $\mathbf{x}$, the parameter $\nu \leq M/n$. For normal distribution $\gamma \leq 2/3n$. The situation when the dimension n is large, sample size N is large, the ratio n/N is bounded, the maximum fourth moment M is bounded, and γ is small, may be called the situation of the *multiparametric statistics applicability.*

In Section 3.1 the latest achievements of spectral theory of large Gram matrices and sample covariance matrices are presented. Theorem 1 is proved under weakest assumptions for wide class of distributions. Analytical properties of $h(z)$ are investigated and finite location of limit spectra is established. In Section 3.2 the dispersion equations similar to (8) are derived for infinite-dimensional variables.

Note that the regularization of the inverse sample covariance matrix C^{-1} by an addition of a positive "ridge" parameter $\alpha > 0$ to the diagonal of C before inversion produces the resolvent of C involved in Theorem 1. Therefore, the ridge regularization of linear statistical procedures leads to functions admitting the application of our dispersion equations with remainder terms small in the situation of multiparametric statistics applicability. Theorems proved in Sections 3.1 allow to formulate the *Normal Evaluation Principle,* presented in Section 3.3. It states that limiting expressions of standard quality functions for regularized multivariate statistical procedures are determined by only two moments of variables and may be approximately evaluated under the assumption of populations normality.

We say that function $f(\mathbf{x})$ of variable $\mathbf{x}$ from population $\mathfrak{S}$ allows ε-*normal evaluation* in the square mean, if for $\mathfrak{S}$ some normal distribution exists $\mathbf{y} \sim \mathbf{N}(\vec{\mu}, \Sigma)$ with $\vec{\mu} = \mathbf{E}\ \mathbf{x}$ and $\Sigma = \text{cov}\ (\mathbf{x}, \mathbf{x})$ such that

$$\mathbf{E}\ (f(\mathbf{x}) - f(\mathbf{y}))^2 \leq \varepsilon$$

PROPOSITION 3. *Under conditions of multiparametric statistics applicability (large N, bounded n/N and M, and small γ), the principal parts of a number of standard quality functions of regularized linear procedures allow ε-normal evaluation with small $\varepsilon > 0$.*

This means that, in the multiparametric case, it is possible to develop (regularized) statistical procedures such that

1/ their standard quality functions have a small variance and allow reliable estimation from sample data;

2/ the quality of these procedures is only weakly depending on distributions.

Constructing Multiparametric Procedures

In Chapter 4 of this book the spectral theory of sample covariance matrices is used for systematical construction of practical approximately unimprovable procedures. Using dispersion equations one can calculate leading terms of quality functions in terms of parameters excluding dependence on random values, or, on the opposite, to express quality functions only in terms of observable data excluding unknown parameters. To choose an essentially multivariate statistical procedure of best quality, one may solve two alternative extremum problems:

a/ find an a priori best solution of statistical problem using the expression of quality function as a function only on parameters;

b/ find the best statistical rule using the quality function presented as a function of only observable data.

For $n \ll N$ all thus improved multiparametric solutions pass to standard consistent ones.

In case of large n and N the following practical recommendation may be offered.

1/ For multivariate data of any dimension it is desirable to apply always stable and not degenerate approximately optimal multiparametric solutions instead of traditional methods consistent only for fixed dimension.

2/ It is plausible to compare different multivariate procedures theoretically for large dimension and large sample size by quality function expressed in terms of first two moments of variables.

3/ Using the multiparametric technique it is possible to calculate principal parts of quality functions from sample data, compare different versions of procedures and choose better ones for treating concrete samples.

Let us describe the technology of construction of unimprovable multiparametric procedures.

1/ Standard multivariate procedure is regularized and a wide class of regularized solutions is chosen.

2/ The quality function is studied and its leading term is isolated. Then one of two tactics (a) or (b) is followed.

Tactics "a"

1. Using dispersion equation, the observable variables are excluded and the principal part of quality function is presented as a function only on parameters.

2. The extremum problem is solved and an a priori best solution is found.

3. The parameters in this solution are replaced by statistics (having small variance), and a consistent estimator of the best solution is constructed.

4. It is remains to prove that this estimator leads to a solution whose quality function approximates well the best quality function.

Tactics "b"

1. Using dispersion equations the unknown parameters are excluded and the principal part of quality function is expressed as a function only on statistics.

2. An extremum problem is solved and the approximately best solution is obtained depending only on observable data.

3. It is proved that this extremum solution provides the best quality with accuracy to remainder terms of the asymptotics.

In Chapter 4 this multiparametric technique is applied to construction of a number multivariate statistical procedures that are surely not degenerate and are approximately optimal independently of distributions. Among these are problems of optimal estimation of the inverse covariance matrices, optimal matrix shrinkage for sample mean vector, and minimizing quadratic risk of sample linear regression.

In 1983 the author of this book found [63] conditions providing the minimum of limit error probability in the discriminant analysis of large-dimensional normal vectors $\mathbf{x} \sim \mathbf{N}(\vec{\mu}_\nu, \Sigma)$, $\nu = 1, 2$, within a generalized class of linear discriminant function. The inverted sample covariance matrix C in the standard discriminant "plug-in" linear discriminant function is replaced by the matrix

$$\Gamma(C) = \int_{t>0} (I + tC)^{-1} d\eta(t)$$

where $\eta(t)$ are arbitrary functions of finite variation. In [63] the extremum problem is solved and the Stilties equation is derived for the unimprovable function $\eta(t) = \eta_0(t)$. This equation was used in [82] by V.S.Stepanov for treating some real (medical and economical) data. He found that it provides remarkably better results even for not great n, $N \approx 5 - 10$. In Section 5.2 this method is extended to a wide class of distributions.

Optimal Solution to Empirical Linear Equations

The sixth chapter of this book presents the development of statistical approach for finding minimum square pseudosolutions to large systems of linear empiric equations whose coefficients are random values with known distribution function. The standard solution to system of linear algebraic equations (SLAE) $A\mathbf{x} = \mathbf{b}$ using known empiric random matrix of coefficients R and empiric right-hand side vector $\mathbf{y}$ can be unstable or non-existing if the variance of coefficients is sufficiently large. The minimum square solution $\hat{\mathbf{x}} = (R^T R)^{-1} R^T \mathbf{y}$ with empiric matrix R and empiric right-hand sides also can be unstable or non-existing. These difficulties are produced by incorrect solution and the inconsistency of random system. The well-known Tikhonov regularization methods [83] are based on a rather artificial requirement of minimum complexity; they guarantee the existence of a pseudosolution, but minimize neither the quadratic risk, nor the residuals. Methods of the well known confluent analysis [44] lead to the estimator $\hat{\mathbf{x}} = (R^T R - \lambda I)^{-1} R^T \mathbf{y}$, where $\lambda \geq 0$, and I is the identity matrix. These estimators are even more unstable and surely do not exist when the standard minimum square solution does not exist (due to additional estimating of the coefficients matrices).

In Chapter 6 the extremum problem is solved. The quadratic risk of pseudosolutions is minimized within a class of arbitrary linear combinations of regularized pseudosolutions with different regularization parameters. First, in Section 6.1, an a priori best solution is obtained by averaging over all matrices A with fixed spectral norm and all vectors $\mathbf{b}$ of fixed length. Section 6.2 presents the theoretical development providing methods of the construction of asymptotically unimprovable solutions of unknown SLAE $A\mathbf{x} = \mathbf{b}$ from empiric coefficient matrix R and the right-hand side vector $\mathbf{y}$ under the assumption that all entries of the matrix R and components of the vector $\mathbf{b}$ are independent and normally distributed.

.