¡html¿¡body¿¡h3¿
Russian Mathematical Surveys, vol. 54, 1999
English translation from *Uspekhi Mat. Nauk,* vol. 54, 2, pp. 85–112

THEORY OF ESSENTIALLY MULTIVARIATE STATISTICAL ANALYSIS

V. I. Serdobolskii

Introduction

It is well known that classical mathematical investigations in multivariate statistical analysis are reduced to calculation of some exact distributions and their functions under an assumption of the observation normality. Traditional asymptotic methods of statistics (see, for example, [1]) were developed for one-dimensional and small-dimensional problems. Their formal extrapolation to many-dimensional problems (by a replacement of scalars by vectors and matrices) not accounting of errors of a large number of parameters did not enrich the multivariate analysis neither with new methods, nor with new results interesting for applications. One can say that central problems of the multivariate analysis remain unsolved. Attempts to find non-improvable statistical procedures fail except for a few cases (see [2; Chapter 8]). The simplest problem of the estimation of the expectation value vector minimizing the quadratic risk is solved only for normal vectors with independent components. Standard linear methods of the multivariate analysis may lead to unstable solutions or (if sample covariance matrix is degenerate) to no solution at all. In the case when sample size is not much larger than the dimension of observations, traditional methods of multivariate analysis do not manifest their consistency.

A substantial progress was achieved in investigations [3], [4], [5], and [6] carried out by the initiative of A. N. Kolmogorov, where a new specific asymptotic approach was developed. Under this approach, a sequence of statistical problems of increasing dimension is considered, in which sample size increases along with the dimension in such a way that the ratio of the dimension to sample size tends to a constant. This constant became an additional parameter of the asymptotic theory. In contrast to the conventional asymptotic approach in mathematical statistics, this new approach was called the "increasing dimension asymptotics", "i.d.a." (see in [7, Chapter 2]). It was discovered that terms of magnitude of the ratio of the dimension to sample size are responsible for a number of essentially multivariate effects such as the accumulation of errors of estimation, appearance of finite asymptotic biases and multiples with vanishing variance and for effects related to the degeneration of

This investigation was supported by the Russian Foundation for Basic Research (grants No.96-01-01574 and 98-01-00781).

sample covariance matrices. The analysis of leading terms of this asymptotics started the development of a systematic theory of essentially multivariate analysis; its advance and achievements are presented below (citing only main publications).

In Section 1, first results are presented of the i.d.a. application to the investigation of the reliability of standard discriminant analysis of normal populations (A. N. Kolmogorov, Yu. N. Blagoveschenskii, A. D. Deev, L. V.Arkharov, L. D. Meshalkin .) In Section 2, the development of limit spectral theory of sample covariance matrices is described using methods of limit spectral theory of random matrices of increasing dimension (V. A. Marchenko and L. A. Pastur, V. L.Girko et al.) Starting form the publication [8] of 1983, this theory is a main tool for the development of a theory of the essentially multivariate analysis. Until recently, the i.d.a. was applied in a form of limit theorems, and only sometimes, rates of convergence were studied. In Sections 3–6, new asymptotic investigations are presented, distinguished by the isolation of i.d.a. principal terms for a fixed dimension and fixed sample sizes. The remainder terms are estimated from above with accuracy to absolute constants. They prove to be small under large samples and a large number of restrictively dependent variables. In Section 3, a refined theory of spectral properties of sample covariance matrices is developed. For a better understanding of the essence and of methods of this new approach, the central theorem on the relation between spectra of sample and true covariance matrices is first presented (under a simplest setting) with full proofs. In Section 4, a generalized linear regression with random predictors is studied and principle parts of the regression quadratic risk are singled out. In Section 5, a generalized class of linear discriminant functions is considered, and the problem of estimation of principle parts of the classification error is considered under i.d.a.. Section 6 is of a summarizing character: it is shown that principal parts of traditional quality functions of regularized multivariate procedures depend only on two moments of variables and can be evaluated (with an accuracy to the remainder terms of i.d.a.) under the assumption of variables normality. In Section 6, the upper estimates of the inaccuracy produced by the normality assumption are found. The theory developed in Sections 1–6 shows that, for a number of regularized versions of multivariate problems for sufficiently wide class of populations and for high dimension of variables, (1) quality functionals weakly depend on details of distributions, (2) reliable estimators of quality functions can be suggested, and thus, (3) a possibility is provided of the comparison of procedures and of search for approximately unimprovable solutions.

We introduce the necessary notations. Let $\mathfrak{S}$ denote an n-dimensional population. Vectors $\mathbf{x}$ from $\mathfrak{S}$ are called observations. Denote $\Sigma = \mathrm{cov}(\mathbf{x}, \mathbf{x})$. We consider samples $\mathfrak{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ from $\mathfrak{S}$ of size N and use sample means and matrices

$$\bar{\mathbf{x}} = N^{-1} \sum_{m=1}^{n} \mathbf{x}_m, \qquad \mathbf{C} = N^{-1} \sum_{m=1}^{N} (\mathbf{x}_m - \bar{\mathbf{x}})(\mathbf{x}_m - \bar{\mathbf{x}})^T \tag{0.1}$$

and

$$\mathbf{S} = N^{-1} \sum_{m=1}^{N} \mathbf{x}_m \mathbf{x}_m^T \tag{0.2}$$

(matrices $\mathbf{S}$ can have the sense of sample covariance matrices if the expected values of $\mathbf{x}$ are known a priori.) We denote the expectation value operator by $\mathbf{E}$ and the variance function by var(.). We use the indicator function ind(.) also for non-random inequalities. We denote vectors by semi-boldface symbols, the transposed vector-column by the upper symbol "T". The absolute value of a vector denote its length, and the square of a vector denotes the square of its length. We only use the spectral norms of matrices. Let $\mathbf{I}$ denote the identity matrix.

1. Method of increasing dimension in multivariate analysis problems

The essentially multivariate approach in statistics was developed first in 1968–1988 for the discriminant analysis. We describe the progress achieved by 1983. Let us set the discriminant problem as follows.

Two populations are considered $\mathfrak{S}_\nu$, $\nu = 1, 2$, and samples $\mathfrak{X}_\nu = (\mathbf{x}_1, \dots, \mathbf{x}_{N_\nu})$ from $\mathfrak{S}_\nu$, $\nu = 1, 2$. 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

$$\alpha_1 = \mathbf{P}\big(w(\mathbf{x}) \leqslant c \mid \mathbf{x} \in \mathfrak{S}_1\big), \qquad \alpha_2 = \mathbf{P}\big(w(\mathbf{x}) > c \mid \mathbf{x} \in \mathfrak{S}_2\big). \tag{1.1}$$

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 the Anderson discriminant function

$$w^0(\mathbf{x}) = (\mu_1 - \mu_2)^T \Sigma^{-1}\big(\mathbf{x} - (\mu_1 + \mu_2)/2\big) \quad \text{with the threshold} \quad c = 0.$$

The minimum is attained for $\alpha_1 = \alpha_2 = \Phi\big(-\sqrt{J}/2\big)$, where $J = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$ is the square of the "Mahalanobis distance". The standard discriminant procedure uses the "plug-in" Fisher–Anderson–Wald sample discriminant function

$$w(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{C}^{-1}\big(\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2\big), \tag{1.2}$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are sample mean vectors and

$$\mathbf{C} = (N-2)^{-1}\left[\sum_{m=1}^{N_1} (\mathbf{x}_m - \bar{\mathbf{x}}_1)(\mathbf{x}_m - \bar{\mathbf{x}}_1)^T + \sum_{m=N_1+1}^{N} (\mathbf{x}_m - \bar{\mathbf{x}}_2)(\mathbf{x}_m - \bar{\mathbf{x}}_2)\right]^T \tag{1.3}$$

is an unbiased estimator of Σ and a through numeration of sample vectors is used, i. e., vectors of the sample $\mathfrak{X}_1$ are numerated first and then vectors of $\mathfrak{X}_2$, $N = N_1 + N_2$. Wald [9] proved the consistency of this procedure for a non-degenerate matrix Σ as $N_1 \to \infty$ and $N_2 \to \infty$.

In view of a deficiency of this procedure (matrix $\mathbf{C}$ may be degenerate, and the inverse matrix certainly does not exist for $n > N$), A. N. Kolmogorov in 1968 was interested in an investigation of the dependence of probability errors on sample sizes. He solved the following problem. Suppose 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 have a non-central χ^2-distribution. To isolate principle 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 discriminant problems

$$\mathfrak{P}_n = \big(S_1, S_2, N_1, N_2, X_1, X_2, w(\mathbf{x}), \alpha_1, \alpha_2\big)_n, \qquad n = 1, 2, \dots \tag{1.4}$$

(we do not write out the subscripts n for arguments of $\mathfrak{P}_n$) of the analysis of observations $\mathbf{x} \in \mathbb{R}^n$, where the discriminant function $w(\mathbf{x})$ is constructed by samples $\mathfrak{X}_1$ and $\mathfrak{X}_2$ of size N_1 and N_2 from populations $\mathfrak{S}_1$ and $\mathfrak{S}_2$, and α_1 and α_2 are probabilities of errors. Supposing that $\mathfrak{S}_\nu = \mathbf{N}(\mu_\nu, \mathbf{I})$, $\nu = 1, 2$, and the discriminant function $w(\mathbf{x}) = w^1(\mathbf{x})$ with $c = 0$ for each n, $(\mu_1 - \mu_2)^2 \to J_0 \geqslant 0$ and $n/N_\nu \to \lambda_\nu > 0$, $\nu = 1, 2$, as $n \to \infty$, he found that $\alpha_1 \to \Phi\big(-J_0/2\sqrt{J_0 + \lambda_1 + \lambda_2}\big)$ in probability (the limit of α_2 is identical.)

In [6] L.D.Meshalkin deduced the same expression of the limit error of discrimination for populations different from normal ones under an assumption that the populations are approaching each other in the parameter space (the contiguity assumption) and the observation vector components are independent. In [10] this result was generalized. In [11] it was shown that the same expression of the limit errors also remains valid for dependent normal variables if the inverted sample covariance matrix is used but this matrix has a special structure and this structure is known a priori. Under the setting of [10], a number of investigations were carried out with the purpose to improve the discrimination procedure for contigual populations with increasing number of blocks of independent variables (see [12], [14], [15].) These papers present a theory of weighting and selection of independent variables in the discrimination problem.

In 1970 Yu. N.Blagoveschenskii and A. D.Deev investigated probability errors of the standard sample discriminant procedure under i.d.a. for two normal populations with coinciding unknown covariance matrices. In [3] and [4], an asymptotic expansion of the function $\mathbf{EP}\big(w(\mathbf{x}) < c, \mathbf{x} \in \mathfrak{S}_1\big)$ was found for $w(\mathbf{x})$ of the form (1.2). The main result is as follows. Suppose that a sequence $\mathfrak{P} = \{\mathfrak{P}_n\}$ of problems (1.4) is given, in which sample discriminant function $w(\mathbf{x})$ is calculated and the discrimination rule $w(\mathbf{x}) > c$ against $w(\mathbf{x}) \leqslant c$ is used.

Theorem 1.1 [3]. *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) *As $n \to \infty$ the limit exists* $\lim (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = J$.
(C) *Let $N_\nu \to \infty$ as $n \to \infty$, $\nu = 1, 2$, in such a way that $n/N_\nu \to \lambda_\nu > 0$, $\nu = 1, 2$, and $\lambda \stackrel{\text{def}}{=} \lambda_1\lambda_2/(\lambda_1 + \lambda_2) < 1$.*

Then

$$\alpha_1 \to \Phi\big(-\sqrt{1-\lambda}\,(J - \lambda_1 + \lambda_2 - 2c)/2\sqrt{J + \lambda_1 + \lambda_2}\,\big)$$

in probability (the limit value of $\mathbf{E}\alpha_2$ is symmetric.)

It is easy to see that minimum of the limit value $(\alpha_1 + \alpha_2)/2$ is attained for the threshold $c = (\lambda_1 - \lambda_2)/2$, i. e., in the classification with a preference of the lesser sample. It was obvious that by taking into account terms of the order of n/N_ν,

$\nu = 1, 2$, we have a possibility to construct improved discriminant (and other) procedures.

If populations are normal, matrix (1.3) is the Wishart matrix. For these, the entries distribution density as well as the eigenvalue density are well known [16], [17], and can be written in the form of analytical expressions. Unfortunately, efforts to use these expressions under i.d.a. were unsuccessful. In [5] for the Wishart matrices $\mathbf{W}$ with $\Sigma = \mathbf{I}$ some recurrent relations were found (by an explicit evaluation) defining the limit momenta $M_k = \operatorname{plim}_{n\to\infty} n^{-1} \operatorname{tr} \mathbf{W}^k$, $k = 1, 2, \dots$; and the existence of a limit distribution function

$$F(u) = \operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \operatorname{ind}(\lambda_i \leqslant u),$$

was proved, where $\lambda_1, \dots, \lambda_n$ are eigenvalues of $\mathbf{W}$. Moreover, in [5] an attempt was made to numerically recover the function $F(u)$ using the momenta $\{M_k\}$. It is noteworthy that this function was found in an analytical form earlier (as a consequence of Theorem 1 from [18], see below). For $\mathbf{x} \sim \mathbf{N}(0, \Sigma)$, the moments M_k were calculated under i.d.a. in [8] (by differentiation with respect to parameters.) For $y > 0$, the formula was deduced $M_k = (\mathbf{L}^k)_{11}/y$, $k = 1, 2, \dots$, where an infinite matrix $\mathbf{L}$ has entries $\mathbf{L}_{ij}$ that are equal to zero for $j < i-1$; to 1 for $j = i-1$; and to $y\Lambda_{j-i+1}$ for $j > i-1$, $\quad i, j = 1, 2, \dots$, where by definition $\Lambda_k = \lim_{n\to\infty} n^{-1} \operatorname{tr} \Sigma^k$, $k = 1, 2, \dots$.

In [8] the principle parts of functions $\mathbf{E} n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{C})^{-1}$ were found for normal $\mathbf{x} \sim \mathbf{N}(0, \Sigma)$ under i.d.a., where $\mathbf{C}$ of the form (1.3) are Wishart matrices. Let us cite the main result.

Let $\mathfrak{P} = \{\mathfrak{P}_n\}$ be a sequence of statistical problems

$$\mathfrak{P}_n = (\mathfrak{S}, \Sigma, N, \mathfrak{X}, \mathbf{C})_n, \qquad n = 1, 2, \dots, \tag{1.5}$$

in which sample covariance matrices $\mathbf{C}$ of the form (1.3) are calculated over samples $\mathfrak{X}$ of size N from populations $\mathfrak{S}$ with $\operatorname{cov}(\mathbf{x}, \mathbf{x}) = \Sigma$, and the limit spectrum of $\mathbf{C}$ is studied (we do not write out the subscripts n of arguments in $\mathfrak{P}_n$.)

Theorem 1.2 [8]**.** *Suppose the sequence $\mathfrak{P}$ satisfies the following conditions* :

(A) *for each n in $\mathfrak{P}_n$, the sets $\mathfrak{S}$ are normal $(0, \Sigma)$;*
(B) *for each n all eigenvalues of Σ lie on a segment $[c_1, c_2]$, where $c_1 > 0$ and c_2 do not depend on n;*
(C) *for any $t \geqslant 0$ as $n \to \infty$ in $\mathfrak{P}$ $n^{-1} \operatorname{tr}(\mathbf{I} + t\Sigma)^{-1} \to \eta(t)$;*
(D) *the limit exists $y = \lim_{n\to\infty} n/N > 0$.*

Then for any $t \geqslant 0$ the limit in probability exists

$$h(t) = \operatorname*{plim}_{n\to\infty} n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{C})^{-1},$$

and the equation is satisfied

$$h(t) = \eta\big(ts(t)\big), \quad \text{where} \quad s(t) = 1 - y + yh(t). \tag{1.6}$$

The importance of the equation (1.6) is that it allows to connect limit spectral functions of sample covariance matrices with spectral functions of unknown true covariance matrices.

We note that moments M_k can be evaluated by differentiating the function $h(t)$. For $\Sigma = \mathbf{I}$, $n = 1, 2, \dots$ the equation (1.6) is reduced to the quadratic equation that was found earlier in [18] (see below, Section 2).

2. Limit spectral theory of sample covariance matrices of increasing dimension

This theory is a development of the theory of random matrices that was created first for some applications in theoretical physics. We present the progress achieved in investigations by V. A. Marchenko and L. A. Pastur, V. L. Girko and the author of this paper in 1967–1995. In 1947 E. Wigner discovered the convergence of spectral functions of random self-adjoint operators represented by random Gram matrices of the increasing dimension of the form $\mathbf{S}$ defined by (0.2) that have the properties of standard sample covariance matrices for normal distributions and obtained a limit spectral density $f(u)$ proportional to $\sqrt{(u_2 - u)(u - u_1)}$, $0 \leqslant u_1 \leqslant u_2$ ("semicircle law"). In 1967 V. A. Marchenko and L. A. Pastur [18] investigated limit spectra of self-adjoint operators of the form of a sum $\mathbf{A} + \mathbf{S}$, where $\mathbf{A}$ are non-random Hermitian matrices of increasing dimension n with convergent spectral functions, and $\mathbf{S}$ are random matrices (0.2) for $\Sigma = \mathbf{I}$, and obtained an equation connecting limit spectral functions of matrices $\mathbf{A}$ and $\mathbf{A} + \mathbf{S}$. We cite one of their results.

Suppose a sequence $\mathfrak{P} = \{\mathfrak{P}_n\}$ of problems is considered $\mathfrak{P}_n = (\mathfrak{S}, \Sigma, N, \mathfrak{X}, \mathbf{S})_n$, $n = 1, 2, \dots$, of the analysis of spectra of true covariance matrices $\Sigma = \operatorname{cov}(\mathbf{x}, \mathbf{x})$ by sample covariance matrices $\mathbf{S}$ of the form (0.2) that are calculated over a sample $\mathfrak{X}$ of size N from $\mathfrak{S}$ (we do not write out the subscripts n for matrices.) Denote

$$h_n(t) = n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{S})^{-1}, \qquad F_n(u) = n^{-1} \sum_{i=1}^{n} \operatorname{ind}(\lambda_i \leqslant u),$$

where $\lambda_1, \dots, \lambda_n$ are eigenvalues of $\mathbf{S}$.

Theorem 2.1 (a special case of Theorem 1 from [18]). *Suppose $\mathfrak{P}$ satisfies the following assumptions.*

(A) *For each n the observation vectors $\mathbf{x}$ from $\mathfrak{S}$ are such that $\mathbf{E}\mathbf{x} = 0$, $\Sigma = \mathbf{I}$, and all fourth moments of all components of the vector $\mathbf{x}$ exist and are uniformly bounded in $\mathfrak{P}$.*

(B) *The distribution of $\mathbf{x}$ is symmetric with respect to a permutation of the components of $\mathbf{x}$ and invariant with respect to the replacement of $\mathbf{x}$ by $-\mathbf{x}$. This means that the components of $\mathbf{x} = (x_1, \dots, x_n)$ satisfy the relation*

$$\mathbf{E}x_i x_j x_k x_l = a_n(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + (b_n - 3a_n)\delta_{ij}\delta_{jk}\delta_{kl},$$

$i, j, k, l = 1, \dots, n$ (here δ is the Kronecker symbol), $a_n = \mathbf{E}x_1^2 x_2^2$, $b_n = \mathbf{E}x_1^4$.

(C) *The limits exist $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$.*

(D) *The limit exists $\lim_{n\to\infty} n/N = y > 0$.*

Then

1° *for any* $t \geqslant 0$, *the limit exists* $\operatorname{plim}_{n\to\infty} h_n(t) = h(t)$;

2° *for* $u \geqslant 0$, $F_n(u) \xrightarrow{\mathbf{P}} F(u)$ *almost everywhere;*

3° *for* $\operatorname{Re} z < 0$ *and for* $\operatorname{Im} z \neq 0$, $h(z) = \int (1 - zu)^{-1}\, dF(u)$;

4° *the equality is valid* $h(t)\big(1 + ts(t)\big) = 1$, *where* $s(t) \stackrel{\text{def}}{=} 1 - y + yh(t)$, $t \geqslant 0$.

Remark 1. The conditions (B) are satisfied for isotropic distributions, and, in particular, for normal distributions $\mathbf{N}(0, \mathbf{I})$. In this case, the limit spectrum density of matrices $\mathbf{S}$ equals $dF(u)/du = (2\pi yu)^{-1}\sqrt{(u_2 - u)(u - u_1)}$ for $u_1 \leqslant u \leqslant u_2$, where $u_1 = (1 - \sqrt{y})^2$, $u_2 = (1 + \sqrt{y})^2$, and equals 0 for $0 < u < u_1$ and for $u > u_2$; for $y > 1$ the function $F(u)$ has a jump at the point $u = 0$ that is equal to $1 - y^{-1}$.

Remark 2. Suppose the observation vectors are normal $(0_n, \mathbf{I}_n)$, $n = 1, 2, \ldots$. Then, a linear transformation of variables exists transforming matrices $\mathbf{C}$ of the form (0.1) to matrices $\mathbf{S}$ of the form (0.2) (with N lesser by unit); and Theorem 1 from [18] is also valid for matrices $\mathbf{C}$ of the form (0.1).

The subsequent success of the i.d.a. application in the theory of multivariate analysis was provided by the spectral function method that was developed in the limit spectrum theory of random matrices of increasing dimension by V. L. Girko (monographs [19] and [20].) In a series of investigations [19]–[22], etc., limit spectral properties of matrices $\mathbf{S}$ of the form (0.2) were studied. It was proved that the normed trace of the resolvent of matrices $\mathbf{S}$ converges almost surely; the convergence rate was investigated, limit (semicircle) spectra were obtained, and it was shown that these spectra stay within finite boundaries with the probability 1. This class of matrices and these results can be applied in some problems of theoretical physics and theory of neuron nets dynamics (A. Khorunzhy, G. Rodgers, A. Boutet de Monvel, V. Vasiliuk et al.) We note that matrices of the form $\mathbf{S}$ can have the sense of sample covariance matrices under a special setting of statistical problems if expectation values of variables are known a priori (or equal 0.)

The methods of [19]–[22] were applied to standard sample covariance matrices $\mathbf{C}$ of the form (0.1) in [23]–[29]. We illustrate the use of spectral function method for the investigation of sample covariance matrices spectra in the proof of Theorem 3.1 below.

The general approach is as follows. First, for some functionals depending on the resolvent of matrices $\mathbf{S}$, $\mathbf{C}$ and vectors $\overline{\mathbf{x}}$, the decrease of variances is established. This fact is proved using some martingale lemmas (see in Sections 3 and in [20].) We use the following simple statement.

Lemma 2.1 (a consequence of the Burkholder inequality, [30; Chapter 7]). *Let* $f(\mathfrak{X})$ *be a function of a sample* $\mathfrak{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, *and let* $f^m(\mathfrak{X})$ *be not depending on* $\mathbf{x}_m$ *functionally,* $m = 1, \ldots, N$. *Suppose two first moments of these functions exist.*

Then $\operatorname{var} f(\mathfrak{X}) \leqslant \sum_{m=1}^{N} \mathbf{E}\big(f(\mathfrak{X}) - f^m(\mathfrak{X})\big)^2$.

One of independent vectors is excluded, and expectation values of functionals involving the resolvent of sample covariance matrices are connected by a functional relation with expectation values of functionals dependent on Σ. Then, in the same

way as in Theorem 3.1, the basic spectral equation is derived relating limit spectra of matrices $\mathbf{S}$ or $\mathbf{C}$ to limit spectra of Σ. The transition to limit spectra is performed by using statements of the following type (Theorem 3.2.4 in [20]).

Lemma 2.2 [20]. *Let us consider a sequence of non-negatively definite symmetric random matrices $\mathbf{S}_n$ of the form (0.2) and of functions $h_n(t) = n^{-1}\operatorname{tr}(\mathbf{I}+t\mathbf{S}_n)^{-1}$ and $F_n(u) = n^{-1}\sum_{i=1}^{n}\operatorname{ind}(\lambda_i \leqslant u)$, where λ_i are eigenvalues of $\mathbf{S}_n$, $i = 1,\dots,n$, $n = 1,2,\dots$.*

The convergence $h_n(t) \to h(t)$ for each $t \geqslant 0$ in probability is necessary and sufficient for the weak convergence $F_n(u) \to F(u)$, $u > 0$, and for the validity of the relation $\displaystyle\int (1+tu)^{-1}\,dF(u) = h(t)$, $t \geqslant 0$.

We present the result of the investigation of limit spectral functions of sample covariance matrices obtained in [23].

Let $\mathfrak{P} = \{\mathfrak{P}_n\}$ be a sequence of statistical problems

$$\mathfrak{P}_n = (\mathfrak{S}, \Sigma, N, \mathfrak{X}, \mathbf{C})_n, \qquad n = 1,2,\dots, \tag{2.1}$$

of the analysis of spectra of matrices $\Sigma = \operatorname{cov}(\mathbf{x},\mathbf{x})$ by sample matrices $\mathbf{C}$ of the form (0.1) calculated over samples $\mathfrak{X}$ of size N (we do not write out the subscripts n for arguments of (2.1))

Theorem 2.2 [23]. *Suppose $\mathfrak{P}$ satisfies the following requirements (with account of [28]; in [23], the assumptions are more complicated.)*

(A) *For each n, $\mathbf{E}\mathbf{x} = 0$, and there exist all fourth moments of projections of $\mathbf{x}$ onto arbitrary non-random axes uniformly bounded in $\mathfrak{P}$. For each n, all eigenvalues of matrices Σ lie on a segment $[c_1, c_2]$, where the magnitudes $0 < c_1 \leqslant c_2$ do not depend on n.*

(B) *The values $\sup_{\|\Omega\|=1}\operatorname{var}(\mathbf{x}^T\Omega\mathbf{x}/n) \to 0$, where Ω are non-random symmetric non-negatively definite matrices with unit (spectral) norm.*

(C) $\lim_{n\to\infty} n/N = y > 0$.

(D) *For $u \geqslant 0$ almost everywhere*

$$n^{-1}\sum_{i=1}^{n}\operatorname{ind}(\lambda_i \leqslant u) \to F_0(u)$$

uniformly with respect to u, where $\lambda_1,\dots,\lambda_n$ are eigenvalues of Σ.

Then

1° *for $t \geqslant 0$ uniformly in the square mean*

$$n^{-1}\operatorname{tr}(\mathbf{I}+t\mathbf{S})^{-1} \to h(t) \quad \text{and} \quad n^{-1}\operatorname{tr}(\mathbf{I}+t\mathbf{C})^{-1} \to h(t);$$

2° *for $u > 0$ almost everywhere*

$$n^{-1}\sum_{i=1}^{n}\operatorname{ind}(\lambda_i \leqslant u) \to F(u)$$

uniformly in u, where $\lambda_1,\dots,\lambda_n$ are eigenvalues of $\mathbf{S}$ or $\mathbf{C}$;

3° *for* $t \geqslant 0$

$$h(t) = \int (1+tu)^{-1}\, dF(u) = \int \bigl(1+ts(t)u\bigr)^{-1}\, dF_0(u), \tag{2.2}$$

where $s(t) = 1 - y + yh(t)$;

4° *the analytical continuation* $h(z)$ *of* $h(t)$ *to the plane of complex* z *satisfies the Gölder condition*

$$\bigl|h(z) - h(z')\bigr| < c|z - z'|^{\zeta}, \quad \textit{where} \quad c > 0 \quad \textit{and} \quad \zeta > 0;$$

5° *for each* $u > 0$ *as* $\varepsilon \to +0$ *there exists the limit density*

$$\frac{dF(u)}{du} = (\pi u)^{-1} \lim \operatorname{Im} h(-u^{-1} - i\varepsilon);$$

6° *for* $0 < u < c_1(1 - \sqrt{y})^2$ *and for* $u > c_2(1 + \sqrt{y})^2$, *the density* $\dfrac{dF(u)}{du} = 0$;

7° *for* $y > 1$ *the function* $F(u)$ *has a jump at the point* $u = 0$ *equal to* $1 - 1/y$.

The limit spectral equations coincide for matrices $\mathbf{S}$ and $\mathbf{C}$, but the rate of convergence may be different. Also, the principle parts of functions depending on vectors $\overline{\mathbf{x}}$ are different for these matrices.

The equations of the type (2.2) relating limit spectrum distribution of random matrices to spectra of their expectation values are called "canonic" in [20] and [26]. They demonstrate a fundamental property of spectra of large random matrices: under i.d.a., their smooth limit spectral functions prove to be insensitive to moments of order higher than 2 and, consequently, can be evaluated under the assumption of population normality.

In investigations by V. L. Girko [24]–[26], methods of spectral theory of random matrices of increasing dimension [19], [20] were applied to the investigation of spectral properties of standard sample covariance matrices $\mathbf{C}$. The main results are obtained under an assumption of the independence of observation vector components (this condition is weakened and replaced by requirements to parameters in the monograph [26], 1995.) In particular, theorems were proved establishing the convergence of spectral functions of matrices $\mathbf{C}$ under i.d.a. and limit equations were derived for matrices Σ and $\mathbf{C}$; estimators were constructed for the normed trace of the resolvent of matrices Σ and their asymptotic normality proved; finite boundaries of limit spectra of sample covariance matrices were studied. Let us cite three statements proved in [26], 1995.

Let $\mathfrak{P} = \{\mathfrak{P}_n\}$ be a sequence of statistical problems (2.1) of the investigation of spectral functions of matrices Σ of increasing dimension by the observed matrices $\mathbf{C}$ of the form (0.1). We assume the following.

(A) For each n in $\mathfrak{P}$ $\mathbf{Ex} = 0$, matrices Σ are non-degenerate, and vectors $\xi = \Sigma^{-1/2}\mathbf{x}$ in $\mathfrak{S}$ have independent components.

(B1) In $\mathfrak{P}$, components ξ_i of vectors ξ satisfy the Lindeberg condition: for each $\tau > 0$ as $n \to \infty$

$$n^{-1} \sum_{i=1}^{n} \mathbf{E}\xi_i^2 \operatorname{ind}\bigl(|\xi_i| > \sqrt{n}\,\tau\bigr) \to 0.$$

(B2) For each n in $\mathfrak{P}$, for all components $\mathbf{x}$, there exist moments of the order higher than the fourth uniformly bounded for $n = 1, 2, \dots$; all eigenvalues of Σ lie on a segment $[c_1, c_2]$, where $c_1 > 0$ and c_2 do not depend on n.

(C) In $\mathfrak{P}$ as $n \to \infty$, the ratios $y_n \stackrel{\text{def}}{=} n/N \to y > 0$.

Theorem 2.3. *Suppose conditions* (A), (B1) *and* (C) *are satisfied in* $\mathfrak{P} = \{\mathfrak{P}_n\}$. *Then*

1° *for each* n *and each* $t \geqslant 0$ *the system of equations*

$$h_n(t) = n^{-1} \operatorname{tr}\bigl(\mathbf{I} + t s_n(t)\Sigma\bigr)^{-1}, \qquad s_n(t) = 1 - y_n + y_n h_n(t)$$

is uniquely solvable;

2° *for each* $t \geqslant 0$ *as* $n \to \infty$ $h_n(t) - \widehat{h}_n(t) \to 0$ *in probability, where* $\widehat{h}_n(t) = n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{C})^{-1}$;

3° *for* $u > 0$ *in probability*

$$n^{-1} \sum_{i=1}^{n} \operatorname{ind}(\lambda_i \leqslant u) - F_n(u) \to 0,$$

where λ_i, $i = 1, \dots, n$, *are eigenvalues of* $\mathbf{C}$, *and the non-random distribution function* $F_n(u)$ *satisfies the equation*

$$\int (1 + tu)^{-1}\, dF_n(u) = h_n(t), \qquad t \geqslant 0.$$

Theorem 2.4. *Suppose conditions* (A), (B2) *and* (C) *are satisfied in* $\mathfrak{P} = \{\mathfrak{P}_n\}$.

Then for any $t \geqslant 0$ *duly centered and normed values* $\widehat{h}_n(t)$ *have a distribution function that converges as* $n \to \infty$ *to the standard normal law distribution function.*

Theorem 2.5. *Suppose conditions* (A), (B2) *and (C) are satisfied in* $\mathfrak{P} = \{\mathfrak{P}_n\}$.

Then the minimum and maximum eigenvalues λ_1 *and* λ_2 *of matrices* $\mathbf{C}$ *in* $\mathfrak{P}$ *are such that as* $n \to \infty$ *the differences* $\lambda_\nu - \alpha_\nu \to 0$ *in probability,* $\nu = 1, 2$, *where*

$$\alpha_\nu = z_\nu^{-1}\bigl[1 + z_\nu N^{-1} \operatorname{tr} \Sigma(\mathbf{I} - z_\nu \Sigma)^{-1}\bigr], \qquad \nu = 1, 2,$$

and $z_1 > 0$, $z_2 > 0$ *are the maximum and minimum roots of the equation*

$$z^2 N^{-1} \operatorname{tr} \Sigma^2 (\mathbf{I} - z\Sigma)^{-2} = 1.$$

For $\Sigma = \mathbf{I}$ from this equation, one gets the boundaries that are mentioned in a remark to Theorem 2.1.

3. Principal parts of the resolvent and spectral functions of sample covariance matrices

In this section, refining results of [27] and [28], we pass from limit formulas of the asymptotic spectral theory of sample covariance matrices developed before 1988 ([23]–[26]), to relations between principal parts of spectral functions that are valid

for any fixed dimension and any fixed sample size. We will show that it is possible to isolate principal parts of spectral functions with an accuracy that is defined by two parameters depending on the first four moments of variables. We will obtain upper estimates of the remainder terms accurate to absolute constants.

It is well-known that limit formulas of the theory of spectral properties of sample covariance matrices of increasing dimension are only valid for a restricted dependence between components of the observation vectors $\mathbf{x}$. In papers [18] and [23], these restrictions have a form of tensor relations for moments of variables. The theory by V. L. Girko [19]–[20] is based on an assumption of the independence of components of $\mathbf{x}$. In the author's paper [27] of 1994, it was found that, for $\mathbf{E}\mathbf{x} = 0$, the restricted dependence conditions can be reduced to restrictions on two parameters: on the maximal fourth moment of a projection of $\mathbf{x}$ onto non-random axes (defined by vectors $\mathbf{e}$ of unit length)

$$M = \sup_{|\mathbf{e}|=1} \mathbf{E}(\mathbf{e}^T\mathbf{x})^4 \tag{3.1}$$

and a special measure of the quadratic from variance

$$\nu = \sup_{\|\Omega\|=1} \operatorname{var}(\mathbf{x}^T\Omega\mathbf{x}/n), \tag{3.2}$$

where Ω are non-random symmetric non-negatively definite matrices of the unit spectral norm. In [28], the theory of spectral properties of sample covariance matrices was developed for finite n and N. Inequalities of paper [28] are suffice to deduce limit spectral equations of the theory [19]–[26] as $N \to \infty$ if $M < \infty$ and $\nu \to 0$.

Method of the isolation of i.d.a. principal terms. To illustrate basic ideas of the theory and its mathematical tools, we first present the essentially multivariate approach with all calculations for an example of the study (i. e., proof of Theorem 3.1, see below) of the spectral function $h_o(t) = \mathbf{E}n^{-1}\operatorname{tr}\mathbf{H}_o(t)$, where $\mathbf{H}_o(t) = (\mathbf{I} + t\mathbf{S})^{-1}$. This function is fundamental for [19]–[26]. We consider it for more simple covariance matrices $\mathbf{S}$ of the form (0.2) calculated over samples $\mathfrak{X} = \{x_m\}$, $m = 1, \dots, N$. We restrict populations $\mathfrak{S}$ with only two assumptions: that $\mathbf{E}\mathbf{x} = 0$ and that all fourth moments of the vector $\mathbf{x}$ exist. For simplicity, assume that $M > 0$. Then we can denote $\gamma = \nu/M$. Also, we denote $y = n/N$, $s_o(t) = 1 - y + yh_o(t)$ and $\tau = \sqrt{M}\,t$.

Similiarly to [22] and [28], we use the method of one-by-one exclusion of independent vectors. Let us single out the vector $\mathbf{x}_m$ from $\mathfrak{X}$. Define

$$\mathbf{S}^m = \mathbf{S} - N^{-1}\mathbf{x}_m\mathbf{x}_m^T, \qquad \mathbf{H}_o^m = (\mathbf{I} + t\mathbf{S}^m)^{-1},$$
$$\varphi_m = \mathbf{x}_m^T\mathbf{H}_o^m\mathbf{x}_m/N, \qquad \psi_m = \mathbf{x}_m^T\mathbf{H}_o\mathbf{x}_m/N,$$

where $m = 1, \dots, N$. It is easy to verify the identities

$$\mathbf{H}_o = \mathbf{H}_o^m - t\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m^T\mathbf{H}_o/N, \qquad \mathbf{H}_o\mathbf{x}_m = (1 - t\psi_m)\mathbf{H}_o^m\mathbf{x}_m, \tag{3.3}$$
$$(1 + t\varphi_m)(1 - t\psi_m) = 1,$$

$m = 1, \dots, N$. Obviously,

$$|\mathbf{e}^T \mathbf{H}_o \mathbf{x}_m| \leqslant |\mathbf{e}^T \mathbf{H}_o^m \mathbf{x}_m| \quad \text{and} \quad 0 \leqslant t\psi_m \leqslant 1, \quad m = 1, \dots, N. \tag{3.4}$$

We also verify that

$$1 - t\mathbf{E}\psi_m = 1 - t\mathbf{E}\,\mathrm{tr}\,\mathbf{H}_o\mathbf{S}/N = 1 - y\,\mathrm{tr}(\mathbf{I} - \mathbf{H}_o)/n = s_o(t),$$

$m = 1, \dots, N$, and that $s_o(t) \geqslant (1 + \sqrt{M}\,ty)^{-1}$.

Lemma 3.1. *For $t \geqslant 0$ the variance* $\mathrm{var}(\mathbf{e}^T \mathbf{H}_o \mathbf{e}) \leqslant \tau^2/N$.

Proof. We estimate the variance using Lemma 2.1. Excluding the dependence on $\mathbf{x}_m$ with identities (3.3) and (3.4), we find that

$$\mathrm{var}(\mathbf{e}^T \mathbf{H}_o \mathbf{e}) \leqslant t^2 N^{-2} \sum_{m=1}^{N} \mathbf{E}(\mathbf{e}^T \mathbf{H}_o^m \mathbf{x}_m)^2 (\mathbf{x}_m^T \mathbf{H}_o^m \mathbf{e})^2 \leqslant t^2 N^{-1} \mathbf{E}(\mathbf{e}^T \mathbf{H}_o^m \mathbf{x}_m)^4.$$

In view of the independence of $\mathbf{H}_o^m$ from $\mathbf{x}_m$, the right-hand side is not greater than $t^2 N^{-1} M = \tau^2/N$. The lemma is proved.

Denote $\delta = 2\tau^2 y^2 (\gamma + \tau^2/N)$.

Lemma 3.2. *For $t \geqslant 0$ and $N > 1$* $\mathrm{var}(t\psi_m) \leqslant \delta$, $m = 1, \dots, N$.

Proof. We single out one of sample vectors $\mathbf{x}_m$ from $\mathbf{H}_o$ using the first of the identities (3.3). Denote $\Delta\varphi_m = \varphi_m - \mathbf{E}\varphi_m$, $\Delta\psi_m = \psi_m - \mathbf{E}\psi_m$. From the last of identities (3.3), it follows

$$(1 + t\varphi_m) t\Delta\psi_m = (1 - t\mathbf{E}\psi_m) t\Delta\varphi_m + t^2 \mathbf{E}\Delta\varphi_m \Delta\psi_m.$$

We square these expressions, calculate expectation values, and obtain

$$\mathrm{var}(t\psi_m) \leqslant \mathrm{var}(t\varphi_m) + \mathrm{var}(t\varphi_m)\,\mathrm{var}(t\psi_m),$$

where $t\psi_m \leqslant 1$ and, consequently, $\mathrm{var}(t\psi_m) \leqslant 2\,\mathrm{var}(t\varphi_m)$. In order to apply Lemma 2.1, we introduce the parameter $t' = (1 - N^{-1})t$. Then $\mathbf{H}_o^m = (\mathbf{I} + t'\mathbf{S}')^{-1}$, where $\mathbf{S}'$ is a matrix of the form $\mathbf{S}$ for a sample of size lesser by unit (without $\mathbf{x}_m$). In view of the independence of $\mathbf{S}'$ from $\mathbf{x}_m$, we obtain

$$\mathrm{var}(t\varphi_m) = \mathrm{var}\big[t\mathbf{x}_m^T (\mathbf{E}\mathbf{H}_o^m)\mathbf{x}_m/N\big] + t^2 \mathbf{E}(\mathbf{x}_m^2/N)^2\,\mathrm{var}(\mathbf{e}^T \mathbf{H}_o^m \mathbf{e}),$$

where $\mathbf{e}$ is a unit vector directed along $\mathbf{x}_m$, and the second variance in the right-hand side is conditional under fixed $\mathbf{e}$. Here the first summand is not greater than $t^2 y^2 \nu = \tau^2 \gamma$. In the second summand, the variance is not greater $Mt'^2/(N-1) \leqslant Mt^2/N = \tau^2/N$ by Lemma 2.1, and $\mathbf{E}(\mathbf{x}_m^2/N)^2 \leqslant My^2$. Thus, the second summand is not greater than $\tau^4 y^2/N$. This proves the lemma.

Theorem 3.1. *For $t \geqslant 0$*

$$h_o(t) = n^{-1} \operatorname{tr}\bigl(\mathbf{I} + ts_o(t)\Sigma\bigr)^{-1} + \omega, \tag{3.5}$$

where $|\omega| \leqslant \tau(\sqrt{\delta} + \tau/N)$.

Proof. We choose a vector $\mathbf{x}_m \in \mathfrak{X}$. In view of (3.3)

$$t\mathbf{H}_o\mathbf{x}_m\mathbf{x}_m^T = t(1 - t\psi_m)\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m^T.$$

Here, the expectation value of the left-hand side is $t\mathbf{E}\mathbf{H}_o\mathbf{S} = \mathbf{I} - \mathbf{E}\mathbf{H}_o$. In the right-hand side, $1 - t\psi_m = s_o(t) - \Delta_m$, where Δ_m is the deviation of $t\psi_m$ from the expected value, and $\mathbf{E}\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m = \mathbf{E}\mathbf{H}_o^m\Sigma$. We find that

$$\mathbf{I} - \mathbf{E}\mathbf{H}_o = ts_o(t)\mathbf{E}\mathbf{H}_o^m\Sigma - t\mathbf{E}\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m^T\Delta_m.$$

We substitute the expression for $\mathbf{H}_o^m$ in terms of $\mathbf{H}_o$ from (3.3). The equation can be rewritten as follows

$$\mathbf{I} = \mathbf{E}\mathbf{H}_o\bigl(\mathbf{I} + ts_o(t)\Sigma\bigr) + \Omega,$$

where $\Omega = t^2 s(t)\mathbf{E}\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m^T\mathbf{H}_o\Sigma/N - t\mathbf{E}\mathbf{H}_o^m\mathbf{x}_m\mathbf{x}_m^T\Delta_m$. Multiplying from the left by $\mathbf{R} = \bigl(\mathbf{I} + ts_o(t)\Sigma\bigr)^{-1}$, we calculate the trace and divide by n. We find that $n^{-1}\operatorname{tr}\mathbf{R} = h(t) + \omega$, where

$$\omega = t^2 s_o(t)\mathbf{E}|\mathbf{x}_m^T\mathbf{H}_o\Sigma\mathbf{R}\mathbf{H}_o^m\mathbf{x}_m/n|/N + t\mathbf{E}|\mathbf{x}_m^T\mathbf{R}\mathbf{H}_o^m\mathbf{x}_m\Delta_m/n|.$$

In the first summand, $0 \leqslant s_o(t) \leqslant 1$. We estimate the matrix expressions by norm, apply the Schwarz inequality and obtain

$$|\omega| \leqslant t^2\|\Sigma\|\mathbf{E}\mathbf{x}_m^2/nN + t\bigl[\mathbf{E}(\mathbf{x}_m^2/n)^2\mathbf{E}\Delta_m^2\bigr]^{1/2} \leqslant \tau^2/N + \tau\sqrt{\operatorname{var} t\psi_m}.$$

The statement of the theorem follows.

The principle parts of (3.5) connect spectra of sample covariance matrices with spectra of true covariance matrices and lead to "canonic" equations of the limit spectral theory of sample covariance matrices (see [23]–[26]).

Example. Let $\Sigma = \mathbf{I}$. The principle parts of (3.5) produce the quadratic equation $h_o(t)\bigl(1 + ts_o(t)\bigr) = 1$ for $h_o(t) = \mathbf{E}n^{-1}\operatorname{tr}(\mathbf{I} + t\mathbf{S})^{-1}$. Using analytical continuation of the function $h_o(t)$, one can calculate the spectral function $F(u) = n^{-1}\sum \mathbf{P}(\lambda_i \leqslant u)$, where the sum is extended over all eigenvalues λ_i, $i = 1, \ldots, n$, of the matrix $\mathbf{S}$. In the limit form, the equation (3.5) was first obtained in [18] (see Section 2.)

Restricted dependence condition. We note that the boundedness of moments M essentially restricts the dependence of variables. Indeed, let Σ be a correlation matrix with the Bayes distribution of correlation coefficients uniform on the segment $[-1, 1]$. Then the Bayes mean $\mathbf{E}M \geqslant \mathbf{E}n^{-1}\operatorname{tr}\Sigma^2 \geqslant (n + 2)/3$. In case of $N(0, \Sigma)$ with matrix Σ of entries 1, the value $M = 3n^2$. Let us show that the equation (3.5) can be established with accuracy to terms, in which the moments are restricted only in a set.

We denote

$$\Lambda_k = n^{-1}\operatorname{tr}\Sigma^k, \qquad Q_k = \mathbf{E}(\mathbf{x}^2/n)^k, \qquad W = n^{-2}\sup_{\|\Omega\|=1}\mathbf{E}(\mathbf{x}^T\Omega\mathbf{x}')^4, \tag{3.6}$$

where $t, k \geqslant 0$, $\mathbf{x}$ and $\mathbf{x}'$ are independent vectors from $\mathfrak{S}$, and Ω are non-random symmetric non-negatively definite matrices of unit spectral norm.

Theorem 3.2. *For $t \geqslant 0$*

$$h_o(t) = n^{-1} \operatorname{tr}\bigl(\mathbf{I} + ts_0(t)\Sigma\bigr)^{-1} + \omega, \tag{3.7}$$

where $\omega^2/2 \leqslant \bigl[Q_2 y^2(\nu + Wt^2/N) + W/N^2\bigr]t^4$.

Example. Let $\mathbf{x} \sim \mathbf{N}(0, \Sigma)$. Denote $\Lambda_k = n^{-1}\ tr\Sigma^k$, $k = 1, 2, \dots$. For normal $\mathbf{x}$

$$M = 3\|\Sigma\|^2, \quad Q_2 = \Lambda_1^2 + 2\Lambda_2/n, \quad W = 3(\Lambda_2^2 + 2\Lambda_4/n), \quad \nu = 2\Lambda_2/n.$$

Now, let $\Sigma = \mathbf{I} + \rho\mathfrak{E}$, where $\mathfrak{E}$ is matrix of entries 1, $0 \leqslant \rho \leqslant 1$. Then $M = 3(1 + n\rho)^2$, $\Lambda_1 = 1 + \rho$, $\Lambda_2 = 1 + 2\rho + n\rho^2$, $\Lambda_k \leqslant a_k + b_k\rho^k n^{k-1}$, where a_k and b_k are positive numbers independent on n, and all $Q_k < c$, $k = 1, 2, \dots$. For $\rho = n^{-3/4}$ and $n \to \infty$, the values $M \to \infty$, whereas Λ_3, Λ_4 and Q_3 remain bounded, $\nu = O(n^{-1})$, and the condition $\omega \to 0$ is fulfilled.

Main results. We present some statements proved in [27] and [28]. In these papers, the principle parts are singled out not only for the function $h_o(t) = n^{-1} \operatorname{tr} \mathbf{H}_o$ but also for separate matrix elements of the resolvent $\mathbf{H}_o$ and $\mathbf{H}$ of matrices $\mathbf{S}$ and $\mathbf{C}$.

To simplify notations of the remainder terms, we denote

$$\varepsilon = \sqrt{\gamma + 1/N}, \qquad c_{lm} = c_{lm}(t) = a \max(1, \tau^l) \max(1, \lambda^m), \tag{3.8}$$

$l, m = 1, \dots, 9$, where a are absolute constants (for brevity, we omit the parenthesis in $c_{lm}(t)$ denoting the dependence on t.)

Theorem 3.3. *For $t \geqslant 0$*

$$\mathbf{E}\mathbf{H}_o = \bigl(\mathbf{I} + ts_o(t)\Sigma\bigr)^{-1} + \Omega_o, \tag{3.9}$$

where $\|\Omega_o\|^2 \leqslant c_{62}\varepsilon^2$ and $\operatorname{var} \mathbf{e}^T\mathbf{H}_o\mathbf{e} \leqslant \tau^2/N$.

In [28] it was shown that the difference between resolvent $\mathbf{H}_o$ and $\mathbf{H}$ effects only remainder terms. We obtain the following statement.

Theorem 3.4. *For $t \geqslant 0$*

$$\mathbf{E}\mathbf{H} = \bigl(\mathbf{I} + ts(t)\Sigma\bigr)^{-1} + \Omega, \tag{3.10}$$

where $\|\Omega\|^2 \leqslant c_{63}\varepsilon^2$, and $\operatorname{var}(\mathbf{e}^T\mathbf{H}\mathbf{e}) \leqslant a\tau^2/N$, where a is an absolute constant.

The statistics $\widehat{h}_o(t) = n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{S})^{-1}$, $\widehat{s}_o(t) = 1 - y + y\widehat{h}_o(t)$, and $\widehat{h}(t) = n^{-1} \operatorname{tr}(\mathbf{I} + t\mathbf{C})^{-1}$ can be offered as estimators of functions $h_o(t)$, $s_o(t)$, and $h(t)$.

Theorem 3.5. *For $t \geqslant 0$*

(1) $\mathbf{E}\bigl|\widehat{h}_o(t) + h_o(t)\bigr|^2 \leqslant \tau/(nN)$;
(2) $\mathbf{E}\bigl|\widehat{s}_o(t) - s_o(t)\bigr|^2 \leqslant c_{12}/(nN)$;
(3) $\mathbf{E}\bigl|\widehat{h}(t) - h(t)\bigr|^2 \leqslant c_{20}\varepsilon^2$.

As an alternative estimator of the function $s_o(t)$, the statistic $\Psi(t) = \overline{\mathbf{x}}^T\mathbf{H}(t)\overline{\mathbf{x}}$ can be offered:

Theorem 3.6. *For $t \geqslant 0$ the statistic $\tilde{s}_o(t) = \big(1 + t\Psi(t)\big)^{-1}$ is an estimator of $s_o(t)$ such that*

$$\mathbf{E}\tilde{s}_o(t) = s_o(t) + o, \quad \textit{where} \quad |o| \leqslant c_{42}\varepsilon, \quad \textit{and} \quad \operatorname{var}\tilde{s}_o(t) \leqslant c_{64}/N.$$

Theorem 3.4 is a central point of the theory of essentially multivariate analysis. Using it, one can prove and strengthen theorems on "canonic" equations of the theory by V. L. Girko [24]–[26] establishing the relation between spectra of sample and true covariance matrices (by tending γ to zero), approximately calculate spectra of matrices $\mathbf{C}$ in terms of the parameters, and construct "G-estimators" of spectral functions of matrices Σ with guaranteed bounds of inaccuracy. From this theorem, the main results of [8] and [27]–[29] follow. The significance of Theorem 3.4 is that it presents a basement for the construction of improved methods of multivariate analysis (see Sections 4, 5, and 6.)

4. Reduction of the quadratic risk for linear regression with a large number of random predictors

In this section we present some new results that are obtained by the application of methods developed in Sections 2 and 3:

1° we consider a class of generalized regularized sample regression procedures depending on an arbitrary function;

2° using i.d.a., we isolate the principle part of the quadratic risk for this class of regressions and construct its estimator;

3° we obtain upper estimates of the inaccuracy produced by using the i.d.a.

Suppose an $(n+1)$-dimensional population $\mathfrak{S}$ is given, in which observations are pairs $(\mathbf{x}, y)$, where $\mathbf{x} = (x_1, \dots, x_n)$ is a vector of predictors and the scalar y is a response.

We restrict the populations $\mathfrak{P}$ with the following requirements only: the expectation values $\mathbf{E}x = 0$, $\mathbf{E}y = 0$, and there exist all fourth moments of components of $\mathbf{x}$, the fourth moment of y, and all fourth moments of their products. Let $\mathbf{E}\mathbf{x}^2 > 0$ (non-degenerate case.) In this section, we denote $M_4 = \sup_{\mathbf{e}} \mathbf{E}(\mathbf{e}^T\mathbf{x})^4$, $M_8 = \mathbf{E}(\mathbf{x}^2/n)^2y^4$ and introduce two parameters

$$M = \max(M_4, \sqrt{M_8}, \mathbf{E}y^4), \tag{4.1}$$

$$\gamma = \sup_{\Omega} \operatorname{var}(\mathbf{x}^T\Omega\mathbf{x}/n)/M, \tag{4.2}$$

where (and in the following) $\mathbf{e}$ are non-random vectors of unit length, and Ω are symmetrical non-negatively definite matrices with unit spectral norm. We consider linear regression $y = \mathbf{k}^T\mathbf{x} + l + \Delta$, where $\mathbf{k} \in \mathbb{R}^n$ and $l \in \mathbb{R}^1$. The problem is to minimize the quadratic risk $R = \mathbf{E}\Delta^2$ by the best choice of $\mathbf{k}$ and l calculated over a sample $\big\{(\mathbf{x}_m, y_m),\ m = 1, \dots, N\big\}$ from $\mathfrak{S}$.

Denote $\lambda = n/N$, $\mathbf{a} = \mathbf{E}\mathbf{x}$, $a_0 = \mathbf{E}y$, $\Sigma = \operatorname{cov}(\mathbf{x}, \mathbf{x})$, $\sigma^2 = \operatorname{var} y$, $\mathbf{g} = \operatorname{cov}(\mathbf{x}, y)$.

If $\sigma > 0$ and the matrix Σ is non-degenerate, then the a priori coefficients $\mathbf{k} = \Sigma^{-1}\mathbf{g}$ and $l = a_0 - \mathbf{k}^T\mathbf{a}$ provide the minimum of R equal to $R = R^o = \sigma^2 - \mathbf{g}^T\Sigma^{-1}\mathbf{g} = \sigma^2(1 - r^2)$, where r is the multiple correlation coefficient.

We start from the statistics

$$\overline{\mathbf{x}} = N^{-1}\sum_{m=1}^{N}\mathbf{x}_m, \qquad \overline{y} = \sum_{m=1}^{N} y_m,$$
$$\widehat{\sigma}^2 = N^{-1}\sum_{m=1}^{N}(y_m-\overline{y})^2 \qquad \mathbf{S} = N^{-1}\sum_{m=1}^{N}\mathbf{x}_m\mathbf{x}_m^T,$$
$$\mathbf{C} = N^{-1}\sum_{m=1}^{N}(\mathbf{x}_m-\overline{\mathbf{x}})(\mathbf{x}_m-\overline{\mathbf{x}})^T \quad \text{and} \qquad \widehat{\mathbf{g}} = N^{-1}\sum_{m=1}^{N}(\mathbf{x}_m-\overline{\mathbf{x}})(y_m-\overline{y}).$$

The standard "plug-in" procedure with $\widehat{\mathbf{k}} = \mathbf{C}^{-1}\widehat{\mathbf{g}}$ and $\widehat{l} = \overline{y} - \widehat{\mathbf{k}}^T\overline{\mathbf{x}}$ has well-known deficiencies: this procedure does not guarantee the risk minimum, is degenerate in case of multi-collinear data, (when matrix $\mathbf{C}$ is degenerate) and is consistent not uniformly in dimension [31].

We consider the regression $y = \widehat{\mathbf{k}}^T\mathbf{x} + \widehat{l} + \Delta$, where $\widehat{\mathbf{k}}$ and $\widehat{l}$ are calculated over a sample with the "plug-in" constant term $\widehat{l} = \overline{y} - \widehat{\mathbf{k}}^T\overline{\mathbf{x}}$. Its quadratic risk is

$$R = \mathbf{E}\Delta^2 = R^1 + \mathbf{E}(\overset{\circ}{y} - \widehat{\mathbf{k}}^T\overset{\circ}{\mathbf{x}})^2 = (1+1/N)R^1, \tag{4.3}$$

where

$$R^1 \stackrel{\text{def}}{=} \mathbf{E}(\sigma^2 - 2\widehat{\mathbf{k}}^T\mathbf{g} + \widehat{\mathbf{k}}^T\Sigma\widehat{\mathbf{k}}), \tag{4.4}$$

$\overset{\circ}{y} = y - a_0$ and $\overset{\circ}{\mathbf{x}} = \overline{\mathbf{x}} - \mathbf{a}$.

Let us calculate and minimize R^1. We consider the following class of generalized and regularized regressions. Let $\mathbf{H}_o = (\mathbf{I}+t\mathbf{S})^{-1}$ and $\mathbf{H} = (\mathbf{I}+t\mathbf{C})^{-1}$ be resolvents of matrices $\mathbf{S}$ and $\mathbf{C}$, where (as above) $\mathbf{I}$ is the identity matrix. We choose the coefficients $\widehat{\mathbf{k}}$ (everywhere in the following) from a class $\mathfrak{K}$ of statistics of the form $\widehat{\mathbf{k}} = \Gamma\widehat{\mathbf{g}}$, where matrices $\Gamma = \int t\mathbf{H}(t)\,d\eta(t)$, and $\eta(t)$ are functions of variation not greater that 1 on $[0,\infty)$ having a sufficient number of moments

$$\eta_k \stackrel{\text{def}}{=} \int t^k\,|d\eta(t)|, \qquad k = 1,2,\dots . \tag{4.5}$$

We note that the step-like function $\eta(t)$ with a step 1 corresponds to the "ridge regression", see [31]. The regression equation with coefficients $\widehat{\mathbf{k}} \in \mathfrak{K}$ can be called a generalized ridge regression. The value (4.4) depends on $\eta(t)$: $R^1 = R^1(\eta)$, and for $\widehat{\mathbf{k}} \in \mathfrak{K}$,

$$R^1(\eta) = \sigma^2 - 2\mathbf{E}\int t\mathbf{g}^T\mathbf{H}(t)\widehat{\mathbf{g}}\,d\eta(t) + \mathbf{E}\iint D(t,u)\,d\eta(t)\,d\eta(u), \tag{4.6}$$

where

$$D(t,u) \stackrel{\text{def}}{=} tu\widehat{\mathbf{g}}^T\mathbf{H}(t)\Sigma\mathbf{H}(u)\widehat{\mathbf{g}}.$$

Since all arguments in $R^1(\eta)$ are invariant with respect to a displacement of the coordinate origin, we set $\mathbf{a} = \mathbf{E}\mathbf{x} = 0$ and $a_0 = \mathbf{E}y = 0$. Similiarly to [28], we isolate principle parts of functionals under i.d.a. and obtain upper estimates of remainder terms with an accuracy to absolute constants. To be more concise in estimates of the remainder terms, we use the notations

$$\tau = \sqrt{M}t, \qquad \varepsilon = \sqrt{\gamma + 1/N},$$
$$c_{lm} = c_{lm}(t) = \alpha \max(1, \tau^l) \max(1, \lambda^m), \qquad \alpha, l, m \geqslant 0,$$

where α, l and m are numbers (in the following, we do not write out the parenthesis with the dependence on t in coefficients $c_{lm}(t)$.) It is easy to verify that

$$\mathbf{E}(\mathbf{x}^2)^2 \leqslant M, \qquad \mathbf{E}(\overline{\mathbf{x}}^2)^2 \leqslant M\lambda^2, \qquad \|\Sigma\|^2 \leqslant M, \qquad \mathbf{g}^2 \leqslant M,$$
$$\mathbf{E}(\widehat{\mathbf{g}}^2)^2 \leqslant 2M^2(1+\lambda)^2, \qquad \mathbf{E}(\overline{\mathbf{x}}^T\widehat{\mathbf{g}})^2 \leqslant 3M^{3/2}(1+\lambda)^2.$$

Similiarly to Section 3, we first study properties of the Gram matrices $\mathbf{S}$ and its resolvent $\mathbf{H}_o$. We consider functions

$$h_o(t) = n^{-1}\mathbf{E}\,\mathrm{tr}\,\mathbf{H}_o(t) \quad \text{and} \quad s_o(t) = 1 - \lambda + \lambda h_o(t).$$

These functions have the property

$$1 - s_o(t) = tN^{-1}\mathbf{E}\,\mathrm{tr}\,\mathbf{H}_o(t)\mathbf{S} \quad \text{and} \quad (1+\tau\lambda)^{-1} \leqslant s_o(t) \leqslant 1.$$

By virtue of Theorem 3.3

$$\mathbf{E}\mathbf{H}_o(t) = \big(\mathbf{I} + ts_o(t)\Sigma\big)^{-1} + \Omega_o,$$

where $\|\Omega_o\| \leqslant c_{31}\varepsilon$ and $\mathrm{var}\big(\mathbf{e}^T\mathbf{H}_o(t)\mathbf{e}\big) \leqslant \tau^2/N$.

Using the variable exclusion method, we isolate principle parts of functionals that will be needed below.

Lemma 4.1. *For $t \geqslant 0$*

1° $\big|t\mathbf{E}\overline{\mathbf{x}}^T\mathbf{H}_o(t)\widehat{\mathbf{g}}\big| \leqslant M^{1/4}c_{32}\varepsilon$;
2° $\mathrm{var}\big(t\overline{\mathbf{x}}^T\mathbf{H}_o(t)\widehat{\mathbf{g}}\big) \leqslant \sqrt{M}\,c_{42}/N$.

Lemma 4.2. *For $t \geqslant 0$*

1° $t\mathbf{E}\mathbf{e}^T\mathbf{H}_o(t)\widehat{\mathbf{g}} = ts_o(t)\mathbf{E}\mathbf{e}^T\mathbf{H}_o(t)\mathbf{g} + o$, *where* $|o| \leqslant c_{31}\varepsilon$;
2° $t\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\widehat{\mathbf{g}} = \sigma^2\big(1 - s_o(t)\big) + ts_o(t)\mathbf{E}\mathbf{g}^T\mathbf{H}_o(t)\widehat{\mathbf{g}} + o_1$
$= \sigma^2\big(1 - s_o(t)\big) + ts_o^2(t)\mathbf{E}\mathbf{g}^T\mathbf{H}_o(t)\mathbf{g} + o_2$,
where $|o_1| \leqslant \sqrt{M}\,c_{32}\varepsilon$ *and* $|o_2| \leqslant \sqrt{M}\,c_{32}\varepsilon$.

Lemma 4.3. *For $t \geqslant u \geqslant 0$*

$$\begin{aligned} tu\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\mathbf{S}\mathbf{H}_o(u)\widehat{\mathbf{g}} &= ts_o(t)us_o(u)\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\Sigma\mathbf{H}_o(u)\widehat{\mathbf{g}} \\ &\quad + \big(1 - s_o(u)\big)t\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\widehat{\mathbf{g}} + \big(1 - s_o(t)\big)u\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(u)\widehat{\mathbf{g}} \\ &\quad + \sigma^2\big(1 - s_o(t)\big)\big(1 - s_o(u)\big) + o, \end{aligned}$$

where $|o| \leqslant \sqrt{M}\,c_{42}\varepsilon$.

To pass to $\mathbf{C}$, $\mathbf{H}$, and $\widehat{\mathbf{g}}$, we use the identities $\mathbf{C} = \mathbf{S} - \overline{\mathbf{x}}\;\overline{\mathbf{x}}^T$ and $\mathbf{H}(t) = \mathbf{H}_o(t) + t\mathbf{H}_o(t)\overline{\mathbf{x}}\;\overline{\mathbf{x}}^T\mathbf{H}(t)$. Denote $U(t) = \mathbf{e}^T\mathbf{H}(t)\overline{\mathbf{x}}$, $\Psi(t) = \overline{\mathbf{x}}^T\mathbf{H}(t)\overline{\mathbf{x}}$.

Lemma 4.4.

1° $U(t) = V(t) + tU(t)\Phi(t), \quad \big(1 + t\Psi(t)\big)\big(1 - t\Phi(t)\big) = 1;$

2° $ts_o^2(t)\big(\mathbf{E}U(t)\big)^2 \leqslant c_{63}\varepsilon^2;$

3° $ts_o(t)\Psi(t) = 1 - s_o(t) + o$, *where* $o^2 \leqslant c_{74}\varepsilon^2$.

Denote

$$\widehat{h}(t) = n^{-1}\operatorname{tr}\mathbf{H}(t), \qquad h(t) = \mathbf{E}\widehat{h}(t), \qquad s(t) = 1 - \lambda + \lambda h(t).$$

It is easy to verify that for $t \geqslant 0 \quad 0 < s(t) \leqslant (1 + \tau\lambda)^{-1}$.

Lemma 4.5.

1° $\|\mathbf{E}\mathbf{H}(t) - \mathbf{E}\mathbf{H}_o(t)\| \leqslant \min(c_{74}\varepsilon^2, c_{32}\varepsilon);$

2° $|s(t) - s_o(t)| \leqslant c_{11}/N$.

Denote

$$\phi(t) = t\mathbf{g}^T(\mathbf{I} + t\Sigma)^{-1}\mathbf{g}, \qquad \kappa(t) = s(t)\phi\big(ts(t)\big) + \sigma^2\big(1 - s(t)\big).$$

Lemma 4.6.

1° $t\mathbf{E}\mathbf{g}^T\mathbf{H}(t)\widehat{\mathbf{g}} = ts(t)\mathbf{E}\mathbf{g}^T\mathbf{H}(t)\mathbf{g} + o$, *where* $|o| \leqslant \sqrt{M}\, c_{42}\varepsilon;$

2° $t\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(t)\widehat{\mathbf{g}} = \sigma^2\big(1 - s(t)\big) + ts(t)\mathbf{E}\mathbf{g}^T\mathbf{H}(t)\widehat{\mathbf{g}} + o_1 = \kappa(t) + o_2,$
where $|o_1| \leqslant \sqrt{M}\, c_{42}\varepsilon$ *and* $|o_2| \leqslant \sqrt{M}\, c_{42}\varepsilon$.

Lemma 4.7. *For* $t \geqslant u \geqslant 0$

1° $tu\big|\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(t)\Sigma\mathbf{H}(u)\widehat{\mathbf{g}} - \mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\Sigma\mathbf{H}_o(u)\widehat{\mathbf{g}}\big| \leqslant \sqrt{M}\, c_{63}\varepsilon;$

2° $tu\big|\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(t)\mathbf{C}\mathbf{H}(u)\widehat{\mathbf{g}} - \mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}_o(t)\mathbf{S}\mathbf{H}_o(u)\widehat{\mathbf{g}}\big| \leqslant \sqrt{M}\, c_{43}\varepsilon.$

Theorem 4.1. *For* $t \geqslant u \geqslant 0$

$$tu\big|\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(t)\mathbf{C}\mathbf{H}(u)\widehat{\mathbf{g}}\big| = tus(t)s(u)\mathbf{E}D(t,u) + \big(1 - s(u)\big)t\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(t)\widehat{\mathbf{g}} + \big(1 - s(t)\big)u\mathbf{E}\widehat{\mathbf{g}}^T\mathbf{H}(u)\widehat{\mathbf{g}} + \sigma^2\big(1 - s(t)\big)\big(1 - s(u)\big) + o,$$

where $D(t,u)$ *is defined by* (4.6) *and* $|o| \leqslant \sqrt{M}\, c_{63}\varepsilon$.

We isolate the principle part of the quadratic risk R^1, first, in terms of sample characteristics, i. e., in form of a function of $\mathbf{C}$ and $\widehat{\mathbf{g}}$. The principle parts of $\mathbf{E}\widehat{\mathbf{k}}^T\mathbf{g}$ and $\mathbf{E}\widehat{\mathbf{k}}^T\Sigma\widehat{\mathbf{k}}$ are integrals of sample functions (with respect to the measure $\eta(t)$.) Expressions for these are prepared by 4.6 and Theorem 4.1. We consider the statistics

$$\widehat{s}(t) = 1 - \lambda + N^{-1}\operatorname{tr}(\mathbf{I} + t\mathbf{C})^{-1}, \qquad \widehat{\kappa}(t) = t\widehat{\mathbf{g}}^T\mathbf{H}(t)\widehat{\mathbf{g}},$$

$$\widehat{K}(t,u) \stackrel{\text{def}}{=} tu\widehat{\mathbf{g}}^T\mathbf{H}(t)\mathbf{C}\mathbf{H}(u)\widehat{\mathbf{g}} = \frac{t\widehat{\kappa}(u) - u\widehat{\kappa}(t)}{t - u}, \qquad t, u \geqslant 0,$$

$$\widehat{\Delta}(t,u) = \widehat{K}(t,u) - \big(1 - \widehat{s}(t)\big)\widehat{\kappa}(u) - \big(1 - \widehat{s}(u)\big)\widehat{\kappa}(t) + \widehat{\sigma}^2\big(1 - \widehat{s}(t)\big)\big(1 - \widehat{s}(u)\big),$$

where $\widehat{K}(t,u)$ is continuously extended for $t = u$.

Lemma 4.8. *For $t \geqslant u \geqslant 0$*

$$s(t)s(u)\mathbf{E}D(t,u) = \mathbf{E}\widehat{\Delta}(t,u) + o, \quad \textit{where} \quad \mathbf{E}|o| \leqslant \sqrt{M}\, c_{63}\varepsilon.$$

It is convenient to replace the dependence of functionals on $\eta(t)$ by the dependence on function $\rho(t)$ of the form $\rho(t) \stackrel{\text{def}}{=} \int_{0\leqslant x\leqslant t} \frac{1}{s(x)}\, d\eta(x)$. We note that the function $t^k\rho(t)$ has a variation on $[0,\infty)$ that does not exceed $\sqrt{M}\,\eta_{k+1}\lambda$. Let us consider the quadratic risk (4.3) as a function of $\rho(t)$: $R = \mathbf{E}\Delta^2 = R(\eta) = R(\rho)$.

Theorem 4.2. *The statistic*

$$\widehat{R} = \widehat{R}(\rho) = \widehat{\sigma}^2 - 2\int \big[\widehat{\kappa}(t) - \widehat{\sigma}^2\big(1-\widehat{s}(t)\big)\big]\, d\rho(t) + \iint \widehat{\Delta}(t,u)\, d\rho(t)\, d\rho(u)$$

is an estimator of $R = R(\rho)$ for which $\mathbf{E}\widehat{R}(\rho) = R(\rho) + o$, where $|o| \leqslant \sqrt{M}\,\eta_8 c_{05}\varepsilon$.

This theorem presents an estimator approximating the quadratic risk with such accuracy.

Now we find a non-random principle part of the quadratic risk. Denote

$$K(t,u) = \frac{t\kappa(u) - u\kappa(t)}{t-u},$$

$$\Delta(t,u) = K(t,u) - \big(1-s(t)\big)\kappa(u) - \big(1-s(u)\big)\kappa(t) + \sigma^2\big(1-s(t)\big)\big(1-s(u)\big),$$

where the function $K(t,u)$ is continuously extended for $t = u$.

Theorem 4.3. *The quadratic risk $R = R(\rho) = R_o(\rho) + o$, where*

$$R_o(\rho) \stackrel{\text{def}}{=} \sigma^2 - 2\int s(t)\phi\big(ts(t)\big)\, d\rho(t) + \iint \Delta(t,u)\, d\rho(t)\, d\rho(u),$$

and $|o| \leqslant \sqrt{M}\,\eta_6 c_{05}\sqrt{\varepsilon}$.

In a special case, let $\eta(x) = \alpha\ \operatorname{ind}(x \geqslant t)$ for $\alpha, t \geqslant 0$, and $R_o(\rho) = R_o(t,\alpha)$ ("shrinkage ridge regression".) We pass to the limit as $N \to \infty$ and $n \to \infty$, supposing matrices Σ to be non-degenerate for all n and

$$\lambda = n/N \to \lambda_* < 1, \quad \gamma \to 0, \quad \varepsilon \to 0, \quad \sigma^2 \to \sigma_*^2, \qquad r^2 = g^T\Sigma^{-1}g/\sigma^2 \to r_*^2.$$

Example 1. Under these conditions, let $\alpha = 1$ and $t \to \infty$ (we pass to the standard non-regularized regression under i.d.a.) Then

$$s(t) \to 1-\lambda_*, \qquad s'(t) \to 0, \qquad \phi\big(ts(t)\big) \to \sigma_*^2 r_*^2,$$

$$\kappa(t) \to \kappa(\infty) \stackrel{\text{def}}{=} \sigma_*^2 r_*^2(1-\lambda_*) + \sigma_*^2\lambda_*, \qquad t\kappa'(t) \to 0.$$

The quadratic risk (4.3) tends to a limit R^* such that

$$\lim_{t\to\infty} \overline{\lim_{\substack{N\to\infty\\ \varepsilon\to 0}}} \big|R_o(1,t) - R_*\big| = 0$$

and

$$R_* \stackrel{\text{def}}{=} \sigma_*^2(1-\lambda_*)^{-1}(1-r_*^2).$$

For normal distributions, this limit expression was obtained by I. S. Eniukov (see in [31]). It explicitly shows the dependence of the standard regression quality on the dimension of observations and sample size.

Example 2. Under the same conditions for a fixed t, let us choose the parameter α in an optimal way: $\alpha = \alpha^0(t) = s^2(t)\phi\big(ts(t)\big)/\Delta(t,t)$. Then, let us tend t to infinity. We have

$$R(\alpha^0,t) \to R_\infty \stackrel{\text{def}}{=} \sigma_*^2(1-r_*^2)\big[\lambda_* + (1-\lambda_*)r_*^2\big]\big/\big[\lambda_*(1-r_*^2) + (1-\lambda_*)r_*^2\big].$$

The value $R_\infty \leqslant \sigma_*^2(1-r_*^2)/(1-\lambda_*)$. As $\lambda_* \to 1$, the weight coefficient $\alpha^0 \to 0$ in such a way that the quadratic risk remains finite (it tends to σ_*^2) in spite of no regularization.

5. Improvement of the discriminant analysis

In [8], the limit theory of generalized linear discriminant procedures was developed (using a superposition of "ridge estimators" of the inverse covariance matrix) for the analysis of n-dimensional observations as $n \to \infty$ and $n/N \to \lambda > 0$, where N are sample sizes, for two normal populations with coinciding unknown covariance matrices. In the same paper, an extremum limit solution was found. Here we describe a generalization of this theory that is valid for a fixed dimension of observations and for fixed sample sizes.

We consider two n-dimensional populations $\mathfrak{S}_\nu = \mathbf{N}(\mathbf{a}_\nu, \Sigma)$, $\nu = 1,2$, with a common covariance matrix Σ. Let $\mathfrak{X}_\nu = \{\mathbf{x}_m\}$, $\nu = 1,2$, be samples (we mean a through numeration of sample vectors) of size $N_\nu > n \geqslant 1$ from populations $\mathfrak{S}_\nu$, $\nu = 1,2$, $N = N_1 + N_2$, $\overline{\mathbf{x}}_\nu$ are sample means of the form (0.1), $\nu = 1,2$, and $\mathbf{C}$ is a pooled sample covariance matrix (1.3). We denote $\mathbf{H} = \mathbf{H}(t) = (\mathbf{I} + t\mathbf{C})^{-1}$, $t \geqslant 0$.

We study a class $\mathfrak{K}$ of linear discriminant functions, introduced in [8], of the form

$$w(\mathbf{x}) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T\Gamma(\mathbf{C})\big(\mathbf{x} - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2\big), \tag{5.1}$$

where $\Gamma(\mathbf{C}) = \int t(\mathbf{I} + t\mathbf{C})^{-1}\,d\eta(t)$ is a matrix diagonalizing along with $\mathbf{C}$ with eigenvalues that are integrals $\int t(1+t\lambda)^{-1}\,d\eta(t)$, where λ are eigenvalues of $\mathbf{C}$. We assume that $\eta(t)$ is a function of bounded variation that has a sufficient number of moments $\eta_k = \displaystyle\int t^k\,|d\eta(t)|$, $k = 1,2\ldots$.

Let the discriminant rule be of the form $w(\mathbf{x}) > \theta$ against $w(\mathbf{x}) \leqslant \theta$, where θ is a threshold of classification. Probabilities of errors (1.1) depend on samples and on the parameters $\eta(t)$ and θ; we denote

$$\alpha(\eta) \stackrel{\text{def}}{=} \min_\theta(\alpha_1 + \alpha_2)/2 = \Phi(-\sqrt{J}/2), \tag{5.2}$$

where $J = J(\eta) = G^2/D$, $G = G_1 - G_2$ and, conditional under fixed samples, moments G_ν, $\nu = 1,2$, and variance D are equal to

$$G_\nu = \mathbf{E}\big(w(\mathbf{x}) \mid \mathfrak{X}_1, \mathfrak{X}_2, \mathbf{x} \in \mathfrak{S}_1\big) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T\Gamma(\mathbf{C})\big(\mathbf{a}_\nu - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2\big),$$
$$D = \operatorname{var}\big(w(\mathbf{x}) \mid \mathfrak{X}_1, \mathfrak{X}_2, \mathbf{x} \in \mathfrak{S}_\nu\big) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T\Gamma(\mathbf{C})\Sigma\Gamma(\mathbf{C})(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2), \qquad \nu = 1,2.$$

We denote

$$\overline{\mathbf{x}} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2, \qquad k(t) = t\overline{\mathbf{x}}^T\mathbf{H}(t)\overline{\mathbf{x}},$$
$$g_\nu(t) = t(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T\mathbf{H}(t)\big(\mathbf{a}_\nu - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2\big), \qquad \nu = 1,2,$$
$$d(t,u) = tu\overline{\mathbf{x}}^T\mathbf{H}(t)\Sigma\mathbf{H}(u)\overline{\mathbf{x}}.$$

By definition,

$$G_\nu = \int g_\nu(t)\,d\eta(t), \quad \nu = 1,2, \qquad D = \iint d(t,u)\,d\eta(t)\,d\eta(u).$$

It is convenient to introduce a single scale parameter

$$L = \max\bigl(3\|\Sigma_1\|, 3\|\Sigma_2\|, \mathbf{a}^2\bigr), \quad \text{where the vector} \quad \mathbf{a} = \mathbf{a}_1 - \mathbf{a}_2. \tag{5.3}$$

To be concise in estimates of the remainder terms, we denote $c_j = a\max(1, L^j t^j)$, $j = 1,2,\dots,$ where a are absolute constants, $n_0 = \min(N_1 - 1, N_2 - 1)$, and $\varepsilon = \sqrt{\gamma + 1/N}$.

We define $y = n/N$, $y_\nu = n/N_\nu$, $\nu = 1,2$, $h(t) = n^{-1}\mathbf{E}\operatorname{tr}\mathbf{H}(t)$, and $\mathbf{R}(t) = \bigl(\mathbf{I} + ts(t)\Sigma\bigr)^{-1}$.

We construct the following statistics to approximate functions $h(t)$, $s(t)$, $g_1(t)$ (estimators of $g_1(t)$ and $g_2(t)$ are symmetrical) and $d(t,u)$:

$$\widehat{h}(t) = n^{-1}\operatorname{tr}\mathbf{H}(t), \qquad \widehat{s}(t) = 1 - t\ \operatorname{tr}\bigl(\mathbf{H}(t)\mathbf{C}\bigr)/N,$$
$$\widehat{g}_1(t) = k(t)/2 - \bigl(1 - \widehat{s}(t)\bigr)/\widehat{s}(t),$$
$$\widehat{d}(t,u) = tu\overline{\mathbf{x}}^T\mathbf{H}(t)\mathbf{C}\mathbf{H}(u)\overline{\mathbf{x}}.$$

Theorem 5.1. *For $0 \leqslant y \leqslant 1$ and $0 \leqslant u \leqslant t$*

1° $\mathbf{E}\bigl(\widehat{h}(t) - h(t)\bigr)^2 \leqslant \tau^2/N$, $\mathbf{E}\bigl(\widehat{s}(t) - s(t)\bigr)^2 \leqslant c_2/N$;

2° $\mathbf{E}k(t) = t\mathbf{a}^T\bigl(\mathbf{I} + ts(t)\Sigma\bigr)^{-1}\mathbf{a} + (y_1 + y_2)\bigl(1 - h(t)\bigr)/s(t) + o$, *where* $o^2 \leqslant c_7\varepsilon^2$;

3° $(1-y)^2\mathbf{E}\bigl(\widehat{g}_\nu(t) - g_\nu(t)\bigr)^2 \leqslant c_{12}\varepsilon^2$, $\nu = 1,2$;

4° $(1-y)^2\mathbf{E}\bigl[\widehat{d}(t,u)/\bigl(\widehat{s}(t)\widehat{s}(u)\bigr) - d(t,u)\bigr]^2 \leqslant c_{12}\varepsilon^2$.

The problem of estimating of two moments of the discriminant function can be solved as follows. Denote

$$\widehat{G}_\nu = \widehat{G}_\nu(\eta) = \int \widehat{g}_\nu(t)\,d\eta(t), \qquad \nu = 1,2,$$
$$\widehat{D} = \widehat{D}(\eta) = \iint \frac{\widehat{d}(t,u)}{\widehat{s}(t)\widehat{s}(u)}\,d\eta(t)\,d\eta(u).$$

Theorem 5.2. *For $0 \leqslant y \leqslant 1$*

1° $(1-y)^2(\mathbf{E}\widehat{G}_\nu - \mathbf{E}G_\nu)^2 \leqslant a\eta_{10}\varepsilon^2$, $\nu = 1,2$;

2° $\operatorname{var} G_\nu \leqslant a\eta_4/n_0$, $\quad (1-y)^2\operatorname{var}\widehat{G}_\nu \leqslant a\eta_4/n_0$, $\nu = 1,2$;

3° $(1-y)^4(\mathbf{E}\widehat{D} - \mathbf{E}D)^2 \leqslant a\eta_{12}\varepsilon^2$;

4° $\operatorname{var} D \leqslant a\eta_6/n_0$, $\quad \operatorname{var}\widehat{D} \leqslant a\eta_4/n_0$, $\nu = 1,2$,

where a are absolute constants.

We construct the statistic

$$\widehat{J}(\eta) = \frac{\left[\int \bigl(k(t) - (y_1 + y_2)\bigl(1 - \widehat{h}(t)\bigr)/\widehat{s}(t)\bigr)\,d\eta(t)\right]^2}{\iint \bigl[uk(t) - tk(u)\bigr]/\bigl[\widehat{s}(t)\widehat{s}(u)(u-t)\bigr]\,d\eta(t)\,d\eta(u)}.$$

Theorem 5.3. *For $0 < y < 1$ the inequality holds $(1-y)^2 \mathbf{E} D\widehat{D}\big|\widehat{J}(\eta) - J(\eta)\big| \leqslant \eta_9 \varepsilon$.*

This theorem makes it possible to estimate the probability of error $\alpha(\eta)$ by sample with a small bias and a small variance under i.d.a.. Passing to the limit, we obtain limit formulas of paper [8]. The functional $\widehat{J}(\eta)$ is a ratio of two quadratic in $\eta(t)$ expressions, and an obvious minimization is possible. In [8] this minimization is carried out for limit formulas under some additional assumptions, and limit extremum condition are found. Let us formulate this result in a form of two theorems.

Under A. N. Kolmogorov's approach, we consider a sequence (1.4) of problems $\mathfrak{P}_n$ of discriminant analysis of observations from populations $\mathfrak{S}_\nu = \mathbf{N}(a_\nu, \Sigma)$, $\nu = 1, 2$, with the discriminant function (5.1).

Theorem 5.4. *Suppose that in $\{\mathfrak{P}_n\}$*

(A) *conditions of Theorem* 1.2 *are satisfied for the covariance matrices Σ;*
(B) *the ratios $y = n/N \to y^* > 0$, $y_\nu = n/N_\nu \to y_\nu^*$, $\nu = 1, 2$;*
(C) *functions $t\mathbf{a}^T(I + t\Sigma)^{-1}\mathbf{a} \to \phi^*(t)$ uniformly in $t \geqslant 0$.*

Then

1° *uniformly in $t \geqslant 0$ $h(t) \to h^*(t)$, $s(t) \to s^*(t)$, and in probability $k(t) \to k^*(t)$;*
2° *in probability $G_\nu \to G_\nu^*$, $\nu = 1, 2$, and $D \to D^*$;*
3° *if $D^* > 0$, then the sample dependent probability of error* (5.2) *tends to $\alpha^* = \alpha^*(\eta) = \Phi\big(-\sqrt{J^*}/2\big)$ in probability, where $J^* = (G_1^* - G_2^*)^2/D^*$.*

Theorem 5.5. *Suppose that in $\{\mathfrak{P}_n\}$ the conditions* (A)–(C) *of Theorem* 5.4 *are satisfied; moreover, for each n in a system of coordinates, where the matrix Σ is diagonal, the inequality $\max_i a_i^2/\lambda_i < c$ is valid, where a_i are components of the vector $\mathbf{a} = \mathbf{a}_1 - \mathbf{a}_2$ corresponding to the eigenvalues λ_i of the matrix Σ, $i = 1, \ldots, n$, and c does not depend on n.*

Then the analytical continuations of functions $s^*(z)$, $\phi^*(z)$ and $k^*(z)$ satisfy the Hölder condition on the plane of complex z.

If, in addition, the integral equation

$$\int (z+t)^{-1}\, d\eta(t) = \operatorname{Im}\big(s^*(-z)\phi^*(-z)\big)/\operatorname{Im} k^*(-z)$$

is solvable for all $z > 0$, where $\operatorname{Im} k^*(-z) > 0$, and its solution $\eta(t) = \eta^o(t)$ is a function of bounded variation on $[0, \infty)$, then $\alpha^*(\eta^o) = \inf_\eta \alpha^*(\eta)$.

Example 1. Suppose that $\eta(t)$ is a step-wise function with a unit step at the point t, and $t \to \infty$. This corresponds to a vanishing "ridge" regularization and a transition to the standard procedure. Then

$$h^*(t) \to 0, \qquad s^*(t) \to 1, \qquad G_1^* \to (J^*/2 - y_1^*)/(1 - y^*),$$
$$D^* \to (J^* + y_1^* + y_2^*)/(1 - y^*)^3$$

and

$$J^* \to J_o^{*2}(1 - y^*)/(J_o^* + y_1^* + y_2^*),$$

where $J_o^* = \lim_{n\to\infty} \mathbf{a}^T\Sigma^{-1}\mathbf{a}$ in agreement with the Deev formula (Theorem 1.1.)

Example 2. Let a limit distribution of eigenvalues of matrices Σ exist such that it is described by the "ρ-model" (see [8]) of limit spectra depending on two parameters $0 \leqslant \rho < 1$ and $\sigma \geqslant 0$. In this case, the limit spectral equation (1.6) has an explicit analytical solution. Suppose that $\eta(x) = \operatorname{ind}(x \leqslant t),\ t \geqslant 0$ ("ridge regularization" of the discriminant function.) Let for each n for each i, the ratios a_i^2/λ_i be equal to each other and equal to J_n/n, where a_i are components of the vector $\mathbf{a}$ in a system of coordinates, where Σ is diagonal, corresponding to the eigenvalues λ_i of the matrix Σ, $i = 1, \ldots, n$ ("equal contribution model"), and $J_n = \mathbf{a}^T\Sigma^{-1}\mathbf{a}$. Then as $n \to \infty$ the limit exists $J_o^* = \lim J_n$ and the limit value of $J = J(t)$ defined by (5.2) is of the form

$$J^*(t) = J_o^{*2}\big(1 - y^* + 2y^*h^*(t) - (\rho + y^*)h^{*2}(t)\big)/(J_o^* + y_1^* + y_2^*).$$

The maximum is attained for $t = t^o = \rho y^*/(\sigma^2(1-\rho^2))$ and equals to

$$J^*(t^o) = \max_t J^*(t) = J_o^{*2}\big(1 - \rho y^*/(\rho + y^*)\big)/(J_o^* + y_1^* + y_2^*).$$

6. Normalization in estimating the multivariate procedures quality

In Sections 3–5, we saw that a number of functionals measuring the quality of multivariate procedures can be reliably estimated in i.d.a. under the hypothesis of observation normality. In this section, following [29], we establish this property for five classes of rotation invariant functionals of the quality function type for (regularized) multivariate procedures including most often used ones.

As before, we restrict populations $\mathfrak{S}$ by an assumption that vectors $\mathbf{x}$ from $\mathfrak{S}$ have the expectation value $\mathbf{E}\mathbf{x} = 0$, and fourth moments of all components exist.

We introduce a measure of the "functionals normalizability". We say that function $f\colon \mathbb{R}^n \to \mathbb{R}^1$ of a random argument $\mathbf{x}$ is ε-normalizable (everywhere here in the square mean) for a class of populations $\mathfrak{K}$ if for any population $\mathfrak{S} \in \mathfrak{K}$ there exists a normal random value $\mathbf{y} \sim \mathbf{N}(a, \Sigma)$ with the same moments $\mathbf{a} = \mathbf{E}\mathbf{x}$ and $\Sigma = \operatorname{cov}(\mathbf{x}, \mathbf{x})$ as in $\mathfrak{S}$ and such one that $\mathbf{E}\big|f(\mathbf{x}) - f(\mathbf{y})\big|^2 \leqslant \varepsilon$.

Example 1. Let $n = 1$, $\xi \sim \mathbf{N}(0, 1)$, function $f(\mathbf{x}) = \mathbf{x}$. For a population given by random $\mathbf{x} = \xi^3/15$, this function is ε-normalizable (by $\mathbf{y} = \xi$) with $\varepsilon = 0.18$.

Example 2. Let function $f(t)$ has the first and the second derivative for $t \geqslant 0$ that are bounded in absolute value by b_1 and b_2. Then for populations, in which all variables have all the first and fourth moments and the parameter (3.1) exists, the function $f(\overline{\mathbf{x}}^2)$ with the argument defined by (0.1) is ε-normalizable with $\varepsilon = 4b_1^2d + 16b_2^2d^2$, where $d = \operatorname{var}(\overline{\mathbf{x}}^2) \leqslant My(2+y)/N,\ \ y = n/N$.

Example 3. Let all variables in a population have all first and fourth moments and the parameters (3.1) and (3.2) exist. Then by Theorem 3.4, the matrix elements of the resolvent $\mathbf{H} = (\mathbf{I} + t\mathbf{C})^{-1}$ are ε-normalizable with $\varepsilon = c_{63}(\gamma + 1/N)$ for $t \geqslant 0$, where c_{63} is defined by (3.8).

We investigate the normalizability of functionals of the quality function type involving resolvents of sample covariance matrices $\mathbf{S}$ and $\mathbf{C}$. Let $\mathfrak{S}_1, \ldots, \mathfrak{S}_k$ be

n-dimensional populations with expectation values $\mathbf{E}_i\mathbf{x} = \mathbf{a}_i$ and $\operatorname{cov}_i(\mathbf{x},\mathbf{x}) = \Sigma_i$ and moments M_i of the form (3.1) for $\mathbf{x}$ from $\mathfrak{S}_i$, $i = 1,\dots,k$. Let $\mathfrak{X}_i = \{\mathbf{x}_m\}$ be a sample from $\mathfrak{S}_i$ of size N_i; denote $y_i = n/N_i$; let $\overline{\mathbf{x}}_i$ be sample mean and

$$\mathbf{S}_i = N^{-1}\sum_{m=1}^{N_i}(\mathbf{x}_m - \mathbf{a}_i)(\mathbf{x}_m - \mathbf{a}_i)^T,$$
$$\mathbf{C}_i = N^{-1}\sum_{m=1}^{N_i}(\mathbf{x}_m - \overline{\mathbf{x}}_i)(\mathbf{x}_m - \overline{\mathbf{x}}_i)^T, \qquad i = 1,\dots,k.$$

We consider the following classes of functionals:

The class $\mathfrak{L}_1 = \{\Phi_1(t,\mathbf{A})\}$ of functionals of the form $n^{-1}\operatorname{tr}\Gamma$ and of the form $\overline{\mathbf{x}}_1^T\Gamma\overline{\mathbf{x}}_1$, where $\Gamma = (\mathbf{I}+\mathbf{A}+t\mathbf{S}_1)^{-1}$ or $\Gamma = (\mathbf{I}+\mathbf{A}+t\mathbf{C}_1)^{-1}$, $0 \leqslant t < c_1$, and $\mathbf{A}$ are non-random symmetric non-negatively definite matrices.

The class $\mathfrak{L}_2 = \{\Phi_2(t_0,t_1,\dots,t_k)\}$ of functionals of the form

$$n^{-1}\operatorname{tr}\Gamma \quad\text{and}\quad \overline{\mathbf{x}}_i^T\Gamma\overline{\mathbf{x}}_i, \quad i = 1,\dots,k, \quad\text{for}\quad 0 \leqslant t_i < c_2, \quad i = 0,1,\dots,k,$$

where $\Gamma = (\mathbf{I}+t_0\mathbf{A}+t_1\mathbf{S}_1+\dots+t_k\mathbf{S}_k)^{-1}$ or $\Gamma = (\mathbf{I}+t_0\mathbf{A}+t_1\mathbf{C}_1+\dots+t_k\mathbf{C}_k)^{-1}$ and $\mathbf{A}$ are non-random symmetric non-negatively definite matrices $n\times n$.

The class $\mathfrak{L}_3 = \{\Phi_3(t_0,t_1,\dots,t_k)\}$ of functionals that are all possible (and mixed) partial derivatives $\Phi_2(t_0,t_1,\dots,t_k)$ with respect to arguments $0 < t_0,t_1,\dots,t_k < c_3$.

The class $\mathfrak{L}_4 = \{\Phi_4(\rho_0,\rho_1,\dots,\rho_k)\}$ of functionals $n^{-1}\operatorname{tr}\Gamma$, where

$$\Gamma = \int \Phi_3(t_0,t_1,\dots,t_k)\,d\rho_0(t_0)\,d\rho_1(t_1)\cdots d\rho_k(t_k),$$

and functions $\rho_i(t)$, $i = 1,\dots,k$, are defined and have a bounded variation on $[0,c_4]$.

The class $\mathfrak{L}_5 = \{\Phi_5(z_1,\dots,z_k)\}$, where Φ_5 are differentiable functions of $z_1,\dots,z_m$ with derivatives bounded by a constant c_5 in absolute value and with arguments that are equal to functionals from $\mathfrak{L}_4$.

We note that the class $\mathfrak{L}_3$ includes well-known rotation invariant functionals $(\mathbf{a}-\alpha\overline{\mathbf{x}})^2$ (where $\mathbf{a} = \mathbf{E}\mathbf{x}$, α is a non-random scalar), $(\overline{\mathbf{x}}-\mathbf{a})^T\Sigma^{-1}(\overline{\mathbf{x}}-\mathbf{a})$, $(\overline{\mathbf{x}}-\mathbf{a})^T\mathbf{S}_\alpha^{-1}(\overline{\mathbf{x}}-\mathbf{a})$ and $(\overline{\mathbf{x}}-\mathbf{a})^T\mathbf{C}_\alpha^{-1}(\overline{\mathbf{x}}-\mathbf{a})$, where (and in the following) $\mathbf{S}_\alpha = \mathbf{S}+\alpha\mathbf{I}$, $\mathbf{C}_\alpha = \mathbf{C}+\alpha\mathbf{I}$, and $\alpha > 0$ are the regularization parameters, as well as the functionals

$$n^{-1}\operatorname{tr}(\mathbf{S}_\alpha^{-1}-\Sigma^{-1})^2, \qquad n^{-1}\operatorname{tr}(\mathbf{C}_\alpha^{-1}-\Sigma^{-1})^2, \qquad n^{-1}\operatorname{tr}(\mathbf{I}-\mathbf{S}_\alpha^{-1}\Sigma)^2,$$
$$n^{-1}\operatorname{tr}(\mathbf{I}-\mathbf{C}_\alpha^{-1}\Sigma)^2, \qquad (\overline{\mathbf{x}}-\mathbf{a})^T\mathbf{C}_\alpha^{-1}\Sigma\mathbf{C}_\alpha^{-1}(\overline{\mathbf{x}}-\mathbf{a}), \qquad (\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)^T\mathbf{C}_{12}^{-1}(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2),$$

where $\mathbf{C}_{12} = \alpha_0\mathbf{I}+\alpha_1\mathbf{C}_1+\alpha_2\mathbf{C}_2$, $\alpha_0,\alpha_1,\alpha_2 > 0$ (the latter functionals are used in the "ridge" linear discriminant analysis.) The class $\mathfrak{L}_4$ includes functionals measuring the quality of a number of procedures with a "generalized ridge estimator" of the inverse covariance matrix (see [7; Chapter 2]).

Theorem 6.1. *Let* $\mathfrak{P} = \{\mathfrak{P}_n\}$ *be a sequence of problems*

$$\mathfrak{P}_n = (\mathfrak{S}_\nu, \mathbf{a}_\nu, M_\nu, N_\nu, \nu = 1, \dots, k;\ \Phi_i, i = 1, \dots, 5)_n, \qquad n = 1, 2, \dots,$$

of the analysis of observations $\mathbf{x} \in \mathbb{R}^n$ *from populations* $\mathfrak{S}_\nu$ *with* $\mathbf{E}\mathbf{x} = \mathbf{a}_\nu$ *and moments* M_ν *of the form* (3.1), $\nu = 1, \dots, k$. *Let the constants* $c_1 - c_5$ *and* k *do not depend on* n.

Suppose that for all $\nu = 1, ..., k$ *for* $n = 1, 2, \dots$

(A) *the values* M_ν *and* $\mathbf{a}_\nu^2$ *are uniformly bounded;*
(B) *for populations* $\mathfrak{S}_\nu$, *values* (3.2) *vanish as* $n \to \infty$*;*
(C) *the ratios* n/N_ν *are uniformly bounded;*
(D) *the variation of functions* $\rho_\nu(t)$ *is uniformly bounded on* $[0, c_4]$.

Then functionals from the classes $\mathfrak{L}_1$–$\mathfrak{L}_5$ *are* ε*-normalizable with* $\varepsilon \to 0$.

Discussion

Thus, in recent years, a new method of asymptotic investigation was developed in mathematical statistics called the increasing dimension asymptotics (i.d.a.) that takes into account specific effects produced by the estimation of a large number of parameters, in which an essential role plays the ratio of the observation dimension to sample size. It can be called a theory of essentially multivariate, or, more precisely, of essentially multi-parametric phenomena. These phenomena are caused by an additional averaging ("self-averaging", "mixing") over a large number of weakly dependent variables. The additional averaging makes functions of large number of variables insensitive to details of distributions so that principal terms of i.d.a. depend only on two first moments of variables. In this sense, the essentially multivariate problems prove to be "normalizable". This approach can be called a theory of statistical analysis by only two first moments of variables (sample and true ones) under conditions of bounded dependence of a large number of variables.

The bounded dependence conditions sufficient for application of our theorems are defined by two parameters: the maximum fourth moment M of the observation vectors projection onto non-random axes (3.1) and a measure γ of the quadratic form variance (3.2). For normal distributions $(0_n, \Sigma_n)$, the moment $M = 3\|\Sigma_n\|^2$, and $\gamma = n^{-2} 2 \operatorname{tr} \Sigma_n^2 / 3\|\Sigma_n\|^2$ ($\|\Sigma_n\|$ is the spectral norm.) For independent components of $\mathbf{x}$, the value $\gamma = O(n^{-1})$.

In contrast to asymptotic approach of [3]–[6], [8], [10]–[15] and [18]–[26], the theory developed in Sections 3–6 makes it possible to isolate principle parts of functions for any fixed dimension of observations n and for any fixed sample sizes N. The remainder terms are estimated from above with accuracy to absolute constants. Their small value is guaranteed under bounded moments M, bounded ratios n/N, large N, and small γ. The results of previous investigations in multivariate analysis carried out under an assumption of normality (in particular, [3]–[6], [11]–[13], [16], [32] and [33]) can be extended to distributions of wider classes. The inaccuracy of such extension of the domain of applicability can be estimated by methods of Section 6.

The problem of stability of estimates to an extension of the class of populations (to "contamination" of samples) was eagerly discussed after well-known papers by

Tukey (1960) concerned with the stability of scale estimators. The discussion stimulated a series of investigations on the construction of various "robust" estimators. We note that, until now, only one class of stable multivariate estimators is known, namely, the class of exponentially weighted estimators [34]; [7], Chapter 8. However, these estimators rapidly lose their effectivity with the increase of the dimension, and that can be explained by an absence of assumptions restricting properties of contaminations. Now we are able to enrich well-known robust methods for problems of bounded dimension with essentially multivariate methods. In those cases, when the class of contaminations is characterized by a priori bounded moments M and small γ, the obtained above essentially multivariate solutions prove to be certainly robust. Their stability is due to the regularization of procedures and the insensitivity to higher moments.

Dealing mainly with one-dimensional problems until recently, asymptotic methods of statistics obtain an additional mathematical tool for the account of essentially multivariate phenomena. Its characteristic feature is an introduction of a multiple description of variables in terms of empirical distribution functions, which represents a large number of boundedly dependent variables. Functionals constructed using empirical distributions of a large number of restrictively dependent variables prove to be approximately normalizable and, in this sense, distribution free. These functionals include standard quality functions of regularized modifications of mostly used procedures. As a result, we have a possibility to construct ε-unimprovable multivariate procedures free from hypotheses on distributions.

The following general approach can be offered for the solution of essentially multivariate statistical problems.

1. Equations are derived connecting principal parts of functionals dependent on parameters and functionals dependent on estimators under i.d.a. (spectral equations, in particular) and upper estimates of their variance are obtained. These equations are used to investigate spectral properties of unknown covariance matrices (Section 3), for sharpening of estimates and for the stabilization and refinement of multivariate procedures (Sections 4 and 5.) Small variance of functionals measuring the quality of statistical procedures under i.d.a. provides the possibility to reliably estimate their quality and to guarantee the improvement.

2. A class of generalized multivariate procedures is introduced depending on a priori parameters and functions.

3. Principle parts of quality functions are singled out under i.d.a. and are expressed, first, in terms of parameters of the populations (that is of theoretical importance), and, second, in terms of functions of statistics (for applications.)

4. An extremum problem is solved for principle parts of quality functions, and conditions of the extremum in a class are found.

5. Inaccuracy of the extremum solution is estimated.

However, a question arises whether the accuracy of asymptotic expressions and quality estimates is sufficient for applications. The remainder terms of the i.d.a. have an order of magnitude of $\sqrt{\gamma + 1/N}$, where in the best case $\gamma = O(n^{-1})$. Efforts to improve solutions lead to a problem of the recovery of the parameter distribution function by the distribution of estimates, and here the accuracy decreases. Here, in all cases, an additional regularization is necessary: in solution of ill-conditioned inverse problems, in using derivatives of empirical step-wise func-

tions, and in solution of the Fredholm integral equations of the first kind. The nature of these difficulties is the same as in a classical problem of a distribution function estimation. The interval of averaging of empirical functions must be sufficiently small to single out the regularity, and at the same time, sufficiently large for reliable judgments. As a result, the inaccuracy of modified procedures can increase to $O(n^{-\alpha})$, where $\alpha > 0$ is small (in [35] $\alpha = 1/4$).

Summing up, we can formulate a general conclusion: the approach developed above makes it possible to suggest reliable distribution free estimators of quality functions for a number of multivariate procedures and open a way, where one can search for improved and, possibly, approximately unimprovable solutions. In this direction, only first steps are made.

I would like to express my sincere gratitude to L..D. Meshalkin for his attention to this development and for suggestions that promoted to improve the manuscript in essence. Also, I am thankful to V. M. Bukhshtaber for a discussion of results of the discriminant problem solution and for his invariable support.

Moscow State Institute of Electronics and Mathematics