Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis 89 (2004) 1–35
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

Results in statistical discriminant analysis: a review of the former Soviet Union literature

Šarūnas Raudys[a] and Dean M. Young[b,*]

[a] *Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania*
[b] *Institute of Statistics, Baylor University, P.O. Box 98005, Waco, TX 76798-8005, USA*

Received 9 February 1999

Abstract

Much work in discriminant analysis and statistical pattern recognition has been performed in the former Soviet Union. However, most results derived by former Soviet Union researchers are unknown to statisticians and statistical pattern recognition researchers in the West. We attempt to give a succinct overview of important contributions by Soviet Block researchers to several topics in the discriminant analysis literature concerning the small training-sample size problem. We also include a partial review of corresponding work done in the West.

AMS 1991 subject classifications: 62H30

Keywords: Plug-in statistical classifiers; Asymptotic error-rate approximations; Regularized discriminant analysis; Nonparametric statistical classifiers; Nonparametric error rate estimation; Feature subset selection

1. Introduction

Largely unknown to discriminant analysis (DA) and statistical pattern recognition (SPR) researchers in the Western Hemisphere, a tremendous amount of research on many varied topics in DA and SPR has been conducted and published in the former Soviet Union. The initial motivation behind these research efforts was the late A.N. Kolmogorov and his colleagues at his statistical methods laboratory at Moscow University.

*Corresponding author.
E-mail addresses: raudys@das.mii.lt (Š. Raudys), dean_young@baylor.edu (D.M. Young).

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0047-259X(02)00021-0

Former Soviet Union DA and SPR researchers have also held several special conferences on SPR. Bi-annual statistics conferences organized by S.A. Aivazyan were held in Estonia and Armenia. Also, two specialized workshops concerning the small training-sample problem were held in Druskininkai, Lithuania, in 1974 and 1984. Unfortunately, most of the SPR results presented at these conferences were published only in Russian. Three exceptions are the review papers by Raudys and Pikelis [128] and Raudys and Jain [126], and the monographs by Vapnik [152] and Raudys [123]. Therefore, most of these results derived by Soviet Union DA and SPR researchers are essentially unknown to all but a few Western DA and SPR researchers.[1]

The goal of this review paper is to succinctly present an overview of interesting and important results relating to small sample DA and SPR. These results have been formulated and published not only by mathematicians and mathematical statisticians, but also by engineers, physicists, computer scientists, and geologists. Thus, not all results discussed here have been derived and stated in a totally rigorous manner. We hope that many Western DA and SPR researchers will become cognizant of the excellent breadth and depth of work that has been done and is ongoing in the former Soviet Union. We include a partial review of corresponding work on these topics published in the West.

This review paper is organized as follows. In Section 2 we provide preliminary definitions used throughout the paper. In Section 3 we present some important results concerning the error rates of several parametric-based statistical discriminant functions. Section 4 contains some interesting results concerning regularized statistical discriminant functions. Section 5 is devoted to results for various nonparametric discriminant functions and to the misallocation robustness of the *sample linear discriminant function* (SLDF). Section 6 deals with alternative linear classifiers. In Section 7 we present important results on error-rate estimation, and in Section 8 we give several little-known results on feature-subset selection for statistical discrimination. We conclude with brief comments in Section 9.

2. Preliminaries

Allocatory DA and SPR techniques are powerful tools for designing statistical classification algorithms and are frequently applied in the areas of data analysis and pattern recognition. The paradigm for DA and SPR is as follows. Assume that each member of the union of m distinct populations possesses a finite set of common characteristics or features which we denote by $\mathbf{f} = (f_1, f_2, \ldots, f_p)$. Also, the observed feature values are denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ such that x_k is the observed value of the feature f_k, $k = 1, 2, \ldots, p$. Let Π_i, $i = 1, 2, \ldots, m$, denote m distinct populations

[1] In the Lithuanian journal *Statistical Problems of Control*, many papers written by a plethora of Soviet authors were published in Russian on the topic of small sample problems in statistical discrimination. For instance, the issues 5, 11, 18, 27, 38, 50, 58, 66, 74, 82, and 93, edited by S. Raudys, contain many such papers. Aivazyan et al. [1] also contains an excellent review of the topic.

of observation vectors having multivariate probability density functions $f_i(\boldsymbol{x})$, $i = 1, 2, \ldots, m$, with means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, m$. Also, let α_i, $i = 1, 2, \ldots, m$, be the known a priori probabilities that an observation is selected from population Π_i, $i = 1, 2, \ldots, m$, respectively. Then, given a $p \times 1$ observation vector $\mathbf{x}$ selected randomly from the union of the populations Π_i, $i = 1, 2, \ldots, m$, the allocatory DA problem is to formulate a decision rule that optimizes some classifier performance criterion. Throughout the remainder of this paper we shall consider the probability of misclassification (PMC) as the classifier performance criterion, and we consider only the two-population case, i.e. $m = 2$.

For the two-population case, assuming all parameters are known, the optimal statistical classification rule which minimizes the PMC is the Bayes rule. This rule is formulated as follows: assign an observation $\mathbf{x}$ to Π_1 if

$$g_{\text{Bayes}}(\mathbf{x}) = \ln \frac{\alpha_1 f_1(\mathbf{x}|\theta_1)}{\alpha_2 f_2(\mathbf{x}|\theta_2)} > 0 \tag{2.1}$$

and to Π_2, otherwise.

There are two main parametric approaches for designing sample-based statistical discriminant functions. The first, the "plug-in approach," assumes the class densities $f_i(\mathbf{x}) = f_i(\mathbf{x}|\theta_i)$, $i = 1, 2$, are known with the exception of the unknown vector of parameters θ_i. Inserting sample estimates $\hat{\theta}_i$ into $f_i(\mathbf{x}|\theta_i)$ for θ_i and then applying (2.1) yields a plug-in statistical discriminant function. The second method of designing sample-based statistical discriminant functions is a Bayesian approach in which a prior density $p_{\text{prior}}(\theta_i)$ on the parameter vector θ_i is assumed to be known. Then, one utilizes the ratio of the a posteriori densities $f(\theta_i|\mathbf{x})$, $i = 1, 2$, to construct a statistical classifier. This approach to classifier design, known as predictive discrimination, has been utilized by Geisser [42], Keehn [60], Kovalevskij [72], and Pugachev [103] to formulate statistical discriminant functions.

The performance of trained, or sample-based, classifiers depends on the particular set of training vectors used in the classifier. Therefore, training-sample-based classification rules differ from optimal classifiers in that the trained classifiers will always have a larger PMC, or error rate, than optimal classifiers.

There are three types of nonoptimal error rates or nonoptimal PMCs: the conditional PMC, the expected or unconditional PMC, and the asymptotic PMC. The *conditional* PMC is the error rate of the classifier for a single training sample from each population of interest. The *expected* PMC is the average conditional error rate of the classifier trained or configured on arbitrary training sets of size N_1 and N_2. The expected PMC, denoted by $EP^{\alpha}_{N_1 N_2}$, depends on the type of classifier α, the training-sample sizes N_1 and N_2, and the classifier-training method. As the training-sample sizes N_1 and N_2 increase, i.e., when $N_1, N_2 \to \infty$, the parameters of the classifier will be estimated with increasing precision. Thus, EP^{α}_{N} will tend to its limiting value P^{α}_{∞}. Here and below we use notations $N_1 = N_2 = N$ and $P^{\alpha}_{\infty} = \lim_{N \to \infty} EP^{\alpha}_{N}$. We refer to P^{α}_{∞} as the *asymptotic* PMC of the classifier α. When the pattern-class density models are correct, then $P^{\alpha}_{\infty} = P^{\text{B}}$, where P^{B} is the optimal, or

Bayes, error rate. The Bayes error rate is the PMC for a classifier when the data model and all parameters are known.

For a finite number of training samples, $EP_N^\alpha - P_\infty^\alpha > 0$ usually obtains. This difference in the expected and asymptotic PMCs depends on the type, complexity, and capacity of the classification rule used, on the training-sample sizes N_1 and N_2, and on the pattern-class characteristics (the dimensionality of the feature vector, the configuration of the class distributions, etc.).

The difference $EP_N^\alpha - P_\infty^\alpha$ and the ratio $\kappa = EP_N^\alpha / P_\infty^\alpha$ are very important characteristics of any classifier. One can use these entities as criteria for selecting the proper classifier complexity, which is a function of the number of parameters in the classifier to be estimated, for determining the optimum number of the features, deriving sufficiently large training-sample sizes, and estimating the PMC of interest. Therefore, many papers have been written concerning the difference $EP_N^\alpha - P_\infty^\alpha$ (or the ratio $EP_N^\alpha / P_\infty^\alpha$) for various statistical classification rules. Most of the error-rate results derived by Western DA and SPR researchers are reviewed in [22,28,39,56,89].

3. The expected PMC of some parametric-based sample discriminant functions

The first attempt to estimate the difference between the expected and asymptotic error rates was made at the Institute for Numerical Analysis at the University of California in Los Angeles using Monte Carlo simulation (see references in [143]). Sitgreaves [140] derived the first exact expression for the expected error rate of Fisher's sample linear discriminant function (SLDF). Her closed-form expression is a five-fold infinite sum of products of certain hypergeometric functions. Her derivation was based on Bowker's [10] representation of the SLDF in terms of independent standard-normal random variables. Estes [33] succeeded in calculating this sum, but the accuracy attained was poor. John [57] represented the SLDF as the difference of two independent chi-square variables and expressed the expected error rate in a closed form as an infinite sum. Pikelis [99] improved the calculation accuracy obtained by Estes and presented a table for the expected error rate as a function of the Mahalanobis distance (see [128], and references therein).

Okamoto (1963, [96]) first derived an asymptotic expansion for the expected error rate of the SLDF. The expansion is obtained under the scenario where the training-sample size $n = N_1 + N_2 \to \infty$ and all other parameters, including the data dimensionality, are fixed. Okamoto's expected error-rate approximation often yields rather inaccurate expected error-rate values if the dimensionality p is large relative to N, i.e., if $N/p < 5$. Later, several similar asymptotic expansions appeared both in the Western countries and in the USSR. These include asymptotic expansions by Anderson [6], Efron [29], Smith [142], McLachlan [85–87], Fukunaga [39], Raudys [111], Raudys and Skurikhina [131], Zarudskij [160], Meshalkin [91], Meshalkin and Serdobolskij [93], Serdobolskij [137], Barsov [7,8], Kharin [63–65], Kharin and Duchinskas [68], and Golcov and Troitsky [44].

As mentioned above, for Western asymptotic approximations of expected error rates, the convention has been to fix p and let $N_i \to \infty$, $i = 1, 2$. This approach differs from that used by most Soviet Union DA and SPR researchers. Their approach has been to simultaneously let $N_i \to \infty$, $i = 1, 2$, and $p \to \infty$ at a rate such that $p/N_i \to \lambda_i$, where λ_i, $i = 1, 2$, are positive constants. The resulting asymptotic error-rate formulations often yield very good approximations even for the case when the training-sample sizes are small relative to the observation dimension [159]. During the last decade this "double asymptotic approach" to expected error-rate approximation has become popular in the analysis of the generalization error, or expected error rate, of artificial neural network (ANN) classifiers [52,53,76,90,98]. In the ANN literature this approach is known as the *thermodynamic limit approach*. We now review a portion of the many important error-rate results for parametric-based statistical classifiers developed in the former Soviet Union.

We now discuss some results concerning the linear discriminant function. First, consider the plug-in linear discriminant function with known covariance matrix and unknown mean vectors

$$g_{\mathrm{F}^*}(\mathbf{x}) = \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{3.1}$$

where $\bar{\mathbf{x}}, \bar{\mathbf{x}}_2$ are the sample mean vectors calculated from the training data and $\boldsymbol{\Sigma}$ is the known covariance matrix. Apparently, the first paper using the double asymptotic approach to error-rate approximation, where the ratio of the learning sample size N and the dimensionality p of the feature space, are increasing at a constant rate, was published by Raudys [108] for (3.1) when $\boldsymbol{\Sigma} = \mathbf{I}$. Raudys [108] assumed the dimensionality p and the training sample sizes $N_1 = N_{2'} = N$ are large, in which case the classifier (3.1) is approximately distributed as a Gaussian random variable. Under these assumptions Raudys [108] derived the error-rate expression

$$EP_N^{\mathrm{F}^*} \approx \Phi\left(-\frac{\delta}{2\sqrt{T_\mu}}\right), \tag{3.2}$$

where $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the squared Euclidean distance between the multivariate Gaussian populations Π_1 and Π_2, $T_\mu = 1 + \frac{1}{N}(1 + \frac{2p}{\delta^2}) + \frac{p}{N^2\delta^2}$, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and the notation $a \approx b$ means $a - b \to 0$. This type of asymptotic analysis was also used to obtain the approximate error-rate expressions (3.4), (3.7),(3.10), (3.14), (3.16), (7.2), and (7.3) below. For a sufficiently large common training-sample size N, expression (3.2) reduces to

$$EP_N^{\mathrm{F}^*} \approx \Phi\left(-\frac{\delta}{2}\left[1 + \frac{2p}{N\delta^2}\right]^{-1/2}\right). \tag{3.3}$$

The expected PMC approximation (3.3) reveals that the dimensionality p and the common training-sample size N are linearly related.

If one assumes covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ in expression (3.1), the statistical discriminant function (3.1) is known as the *sample Euclidean-distance classifier*

(SEDC), which is of the form

$$g_{\mathrm{E}}(\mathbf{x}) = \mathbf{x}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Raudys [108] also analyzed the case when $f_i(\mathbf{x}|\theta_i) = f_i(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, where $\boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}$, and derived an expected PMC approximation for the SEDC. His expected error-rate approximation is

$$EP_N^{\mathrm{E}} \approx \Phi\left(-\frac{\delta}{2}\left[1 + \frac{2p^*}{N^2(\delta^*)^2}\right]^{-1/2}\right), \tag{3.4}$$

where $\delta^* = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^{1/2}}$ is a modified squared Mahalanobis distance and

$$p^* = \frac{((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 \operatorname{tr}\boldsymbol{\Sigma}^2}{((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2} \tag{3.5}$$

is a modified dimensionality measure. When $\boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}$, $P_\infty^{\mathrm{E}} = \Phi(-\delta^*/2) \geqslant P_\infty^{\mathrm{F}} = \Phi(-\delta/2)$.

Conditions on the population covariance structures such that $P_\infty^{\mathrm{E}} = P_\infty^{\mathrm{F}}$, assuming the population means are fixed, are given in [81]. One can easily see from expression (3.5) that $1 < p^* < \infty$. Therefore, in the extreme case when $p^* = 1$, the SEDC is only slightly sensitive to the common training-sample size N. In the case when p^* is extremely large, the SEDC is very sensitive to the common training-sample size N.

Expression (3.3) approximates the expected PMC of a piecewise-linear classifier based on the Euclidean distances of the vector $\mathbf{x}$ to the $2m$ spherical cluster centers of the training-data sets of two separate pattern classes. In this case one should use a common training-sample size of $N_m = N/m$ instead of N in expression (3.3) [58].

Estimating and incorporating the common population covariance matrix $\boldsymbol{\Sigma}$ from the two training samples yields the traditional SLDF,

$$g_{\mathrm{F}}(\mathbf{x}) = \mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{3.6}$$

that is regularly referred to as the standard Fisher's linear discriminant function, or Anderson's statistic (referring to T.W. Anderson). Assuming multivariate Gaussian pattern classes, the SLDF can be derived as a plug-in-based classifier. A.D. Deev, from the A.N. Kolmogorov Laboratory of Statistical Methods at Moscow State University, formalized the double asymptotic approach in a strictly mathematical way: he formally required $N \to \infty$, $p \to \infty$, $p/N \to$ constant, and Mahalanobis distance $\delta =$ constant. Under this approach several subsequent asymptotic expansions were obtained for Gaussian and nonGaussian models. Two simple formulae for the approximate expected error for the standard Fisher linear DF were obtained in [19,20,111]. For the SLDF with $N_1 = N_2 = N$ and assuming Gaussian populations, an approximate expected PMC expression for

large values of N and p is

$$EP_N^{\mathrm{F}} \approx \Phi\left\{-\frac{\delta}{2}\left[\frac{(2N-3)(2N-p-3)}{(2N-p-2)(2N-p-5)}\left(1+\frac{1}{N}\left(1+\frac{2p}{\delta^2}\right)+\frac{p}{N^2\delta^2}\right)+\frac{\delta^2}{2(2N-p-5)}\right]^{-1/2}\right\}$$
$$\to \Phi\left(-\frac{\delta}{2}\left[\left(1+\frac{2p}{N\delta^2}\right)\frac{2N}{2N-p}\right]^{-1/2}\right), \tag{3.7}$$

where $N>p/2$ and $\delta^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ is the squared Mahalanobis distance between the populations Π_1 and Π_2.

Notice that in the last expression in (3.7), the term $\frac{2N}{2N-p}$ reflects the increase in the expected PMC due to estimation of the common covariance matrix $\boldsymbol{\Sigma}$.[2] The term $1+\frac{2p}{N\delta^2}$ reflects the increase in the expected PMC due to estimation of the two p-variate mean vectors μ_1 and μ_2. We also note that (3.7) allows one to understand the asymptotic relationship between the sample size and the dimensionality.

If the training-sample sizes differ so that $N_1\neq N_2$, then Deev [19,20], using the double asymptotic approach, showed that the expected PMC of the SLDF is approximately

$$EP_{N_1N_2}^{\mathrm{F}} \approx \alpha_1\Phi\left(-\frac{\delta^2-\lambda_1+\lambda_2}{2\sqrt{(\delta^2+\lambda_1+\lambda_2)\frac{N_1+N_2}{N_1+N_2-p}}}\right)$$
$$+\ \alpha_2\Phi\left(-\frac{\delta^2+\lambda_1-\lambda_2}{2\sqrt{(\delta^2+\lambda_1+\lambda_2)\frac{N_1+N_2}{N_1+N_2-p}}}\right), \tag{3.8}$$

where $p/N_i\to\lambda_i$, $i=1,2$. Note that expression (3.8) is a generalization of (3.7) and that (3.8) reduces to (3.7) when $N_1=N_2$.

In view of expression (3.8), we see that differing signs of the terms $(\lambda_1-\lambda_2)$ in both of the above summands can cause an increase in the classification error when $N_1\neq N_2$.[3] The evaluation of the approximation to $EP_{N_1,N_2}^{\mathrm{F}}$ given in (3.8) indicates that the plug-in approach to linear classifier construction is not optimal. To partially remedy this property of the SLDF, Deev [20] proposed the utilization of a threshold

[2] Pikelis [100] used an exact formula in the form of an infinite sum of certain functions to calculate the expected error of the SLDF. Wyman et al. [159] used a simulation study to compare the small sample efficacy of several asymptotic expansions for the SLDF and found that the expansions based on the approach that $p\to\infty$ and $N\to\infty$ such that $p/N\to\lambda$ are superior to the asymptotic expansions where p is held constant and $N\to\infty$.

[3] Deev, in fact, proposed using additional higher order terms than those in the error-rate expansion (3.8). The complete asymptotic error-rate expansion is very complex but has been verified by Pikelis [100] to be highly accurate.

constant yielding a linear classifier of the form

$$g_{\mathrm{Deev}}(\mathbf{x}) = g_{\mathrm{F}}(\mathbf{x}) + c,$$

where c is chosen to minimize the expected PMC expression

$$EP_{N_1N_2}^{\mathrm{Deev}}\alpha_1 \approx \Phi\left(\frac{2c-(\delta^2-\lambda_1+\lambda_2)}{2\sqrt{(\delta^2+\lambda_1+\lambda_2)\frac{N_1+N_2}{N_1+N_2-p}}}\right) + \alpha_2\Phi\left(-\frac{2c+(\delta^2+\lambda_1-\lambda_2)}{2\sqrt{(\delta^2+\lambda_1+\lambda_2)\frac{N_1+N_2}{N_1+N_2-p}}}\right).$$

One can easily derive the optimal value of c that minimizes $EP_{N_1,N_2}^{\mathrm{Deev}}$, which is

$$c_{\mathrm{opt}} = -\tfrac{1}{2}(\lambda_1-\lambda_2).$$

We note that when $N_1 = N_2$, then $\lambda_1 = \lambda_2$ and therefore $c_{\mathrm{opt}} = 0$. There are a number of asymptotic expansions derived where only the sample size tends to infinity. To derive such an expansion, Efron [29] represented an increase in the conditional classification error as a sum of two chi-square random variables and obtained

$$EP_N^{\mathrm{F}} \approx \Phi\left(-\frac{\delta}{2}\right) + \phi\left(\frac{\delta}{2}\right)\frac{\delta^2/4+(1+\delta^2/4)(p-1)}{2N\delta},$$

where $\phi(c)$ is the standard $N(0,1)$ Gaussian density function.

Wyman et al. [159] have shown that the double asymptotic expression (3.7) outperforms Efron's expression in accuracy. Recently in the West, Koolaard and Lawoko [71] have compared the expected error rates of the EDC and SLDF via asymptotic expansions, and Fujikoshi [37] and Fujikoshi and Seo [38] have employed a double asymptotic approach to derive expected error-rate approximations for the SLDF. Also, Viollaz et al. [157] have derived an asymptotic approximation for EP_N^{F}.

The nonoptimality of the plug-in approach to statistical classifier construction is even more clearly demonstrated in the case of the *sample quadratic discriminant function* (SQDF) defined as

$$g_{\mathrm{Q}}(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_2)'\boldsymbol{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\} - \{(\mathbf{x}-\bar{\mathbf{x}}_1)'\boldsymbol{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) + \ln\left(\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}\right) + 2\ln(\alpha_1/\alpha_2). \tag{3.9}$$

If one applies the SQDF, one tacitly assumes that the two-population covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unequal. When $N_1 = N_2 = N$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Raudys [111] derived the following approximation for the asymptotic expected PMC of the SQDF

defined in (3.9), which is

$$EP_N^{\mathrm{Q}} \approx \Phi\left(-\frac{\delta}{2}\left[\left(1+\frac{p+2}{N-p-4}\right)\left(1+\frac{1}{N}\left(1+\frac{2p}{\delta^2}\right)+\frac{p}{N^2\delta^2}\right.\right.\right.$$
$$\left.\left.\left.+\frac{(\delta^4/2)+p+p(\delta^2+p)}{(N-p-4)\delta^2}\right)\right]^{-1/2}\right). \tag{3.10}$$

In (3.10) one can see that for a small dimension p and a large Mahalanobis distance δ, the functional relationship between the training-set size N and the dimension p is nearly linear. However, for a large dimension p, the functional relationship between N and p is quadratic, i.e., p^2/N. In spite of the comparatively high accuracy, in comparison with the exact expected errors given in [128] expression (3.10) provides slightly optimistic estimates. Takeshita and Toriwaki [145] compared a modification of expression (3.10) with a similar expression derived by Fukunaga [39]. He determined that for pattern-class configurations having small error rates, the double asymptotic approach yields more accurate expected error-rate approximations.

For linear classifiers the double asymptotic approach yields surprisingly accurate expected error-rate approximations, even for the case when N/p is small. That is, the expected error-rate approximations derived by former Soviet Union researchers apply not only to the large training-sample size case, but also to the small training-sample size situation. Pikelis [100] used his exact formula in the form of an infinite sum to calculate the expected error of the SLDF. He then found that the simple expected error-rate approximations (3.7) and (3.8) outperform the Okamoto (1963) expansion in accuracy. Also, Wyman et al. [159] used a simulation study to compare the small sample efficacy of several asymptotic expansions for the SLDF and found that the expansions based on the double asymptotic approach are superior to the asymptotic expansions where p is held constant and $N \to \infty$. Viollaz et al. compared the expected error-rate approximations of Raudys, Deev, and Okamoto to their expected error-rate approximation and found that their approximation along with those of Raudys and Deev gave excellent results even for the very small training-sample size case when $N \approx p$.

Another important characteristic of the conditional PMC for a classifier is its variance. Researchers in the Western Hemisphere have made important contributions to this topic. Efron [29] represented the increase in the conditional error rate as a quadratic form of normal variables. He presented asymptotic expressions for the mean and the variance of the conditional classification error. He also compared the SLDF with the logistic regression classification procedure for the case where the actual pattern classes are multivariate Gaussian. The approximate standard deviation of the conditional error is

$$\sqrt{VP_N^{\mathrm{F}}} \approx \frac{\phi(\delta/2)}{\sqrt{2}\delta N}\sqrt{\frac{\delta^4}{16}+\left(1+\frac{\delta^2}{4}\right)^2(p-1)}. \tag{3.11}$$

Formula (3.12) shows an interesting behavior for the conditional error rate. From the double asymptotic approach, it follows that when both N and p are simultaneously and proportionally increasing, the expected PMC tends to a constant value. The standard deviation, however, tends to zero. This fact suggests that, regardless of which randomly selected training set one will use in the high-dimensional case, one will obtain approximately the same conditional error. Thus in the high-dimensional cases, one needs to be concerned only to the expected PMC, and not the variance of the conditional classification error. This is an important conclusion for practitioners.

O'Neil [97] generalized Efron's representation for the case when the discriminant function is nonlinear and the actual pattern classes are not Gaussian. Sayre [135] used Efron's representation to derive an expression for the variance of conditional PMC of the SLDF. The conditional PMC variance was also studied by McLachlan [87] and Schervish [136] (also see [89]).

The plug-in estimate of multivariate Gaussian density is biased. Lumelskij [80] proposed an unbiased estimate. The asymptotic expected PMC expression for the SQDF becomes much more complicated than (3.10) for the case when $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and $N_1 \neq N_2$ [111]. The asymptotic expected PMC approximations for the SQDF indicate that increasing the training-sample size N_i in one pattern class and keeping the other sample size N_j fixed for $i,j,=1,2,\ i \neq j$, can actually increase EP_N^{Q} [45,116]. This phenomenon is caused by the fact that when $N_i \ll N_j$, $i,j=1,2,\ i \neq j$, $g_{\mathrm{Q}}(\mathbf{x})$ is a very biased estimator of the Bayes DF (2.1) from which the plug-in SQDF is derived. The major component of this bias is induced by inverse sample covariance matrices $\boldsymbol{S}_1^{-1}$ and $\boldsymbol{S}_2^{-1}$. Therefore, the expectation of the SQDF with respect to random matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ [113] is

$$Eg_{\mathrm{Q}}(\mathbf{x}) = b_1(\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\} - b_2\{(\mathbf{x}-\bar{\mathbf{x}}_1)\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) + \ln\left(\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}\right) + b_3 + 2\ln(\alpha_1/\alpha_2), \tag{3.12}$$

where

$$b_1 = \frac{N_2-1}{N_2-p-2},$$

$$b_2 = \frac{N_1-1}{N_1-p-2},$$

$$b_3 = \left(\sum_{j=1}^{p}\psi\left(\frac{N_2-j}{2}\right) - \sum_{j=1}^{p}\psi\left(\frac{N_1-j}{2}\right) + p\log\left(\frac{N_1-1}{N_2-1}\right)\right)$$

and $\psi(r)$ is the psi Euler (diagamma) function [144].

For instance, assuming multivariate Gaussian class densities of dimension $p=40$ with $\mathbf{\Sigma}_2 = 2\mathbf{\Sigma}_1$ and $N_1 = N_2 = 200$, we obtain $EP_N^{\mathrm{Q}} = 0.151$. For this configuration with $N_1 = 200$ and $N_2 = 20,000$, we have that $EP_N^{\mathrm{Q}} = 0.170$. However, for the same configuration, if $N_1 = 20,000$ and $N_2 = 200$, the result is $EP_N^{\mathrm{Q}} = 0.074$ [116]. A

less-biased quadratic discriminant function follows directly from (3.12) in the form of

$$g_{\mathrm{Q}}(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_2)'\boldsymbol{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\}/b_2 - \{(\mathbf{x}-\bar{\mathbf{x}}_1)'\boldsymbol{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)/b_1$$

$$+ \ln\left(\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}\right) - b_3 + 2\ln(\alpha_1/\alpha_2).$$

Simulation experiments have shown that this bias modification of the SQDF works well if the overlap between the populations is small and the difference between the expected and asymptotic error is not too large ($EP_N^{\mathrm{Q}} < 2P_\infty^{\mathrm{Q}}$).

The dimensionality of the parameter vectors θ_i in $f_i(\mathbf{x}|\theta_i)$, $i = 1, 2$, is an important factor that determines the complexity of parametric-based statistical discriminant functions. The statistical classifier's sensitivity to the training-sample sizes becomes obvious when one compares the approximate expected PMC expressions (3.3) and (3.7). The assumption of $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}$ used to design the SEDC in (3.4) is much too restrictive for many practical DA and PR applications. The less-restrictive assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{D}$, where $\mathbf{D}$ is a diagonal covariance matrix, is more palatable. Under this diagonal common covariance assumption, one will likely use the classifier

$$g_{\mathrm{LD}}(\mathbf{x}) = \mathbf{x}'\hat{\mathbf{D}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\hat{\mathbf{D}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

where $\hat{\mathbf{D}}$ is a diagonal matrix composed of the diagonal elements of the sample covariance matrix. Raudys [111] showed that an expression for the asymptotic expected PMC for $g_{\mathrm{LD}}(\mathbf{x})$ is

$$EP_N^{\mathrm{LD}} \approx \Phi\left(-\frac{\delta}{2}\left[1 + \frac{1}{N}\left(1 + \frac{2p}{\delta^2}\right) + \frac{p}{N^2\delta^2} + \frac{\delta_{(4)}}{4(N-3)}\right]^{-1/2}\right), \tag{3.13}$$

where $\delta_{(4)}^2 = \frac{1}{\delta^2}\sum_{k=1}^{p}\frac{(\mu_{1k}-\mu_{2k})^4}{\sigma_k^2}$ and σ_k^2 is the variance for the kth feature, $k = 1, 2, \ldots, p$. For small training-sample sizes, the term $\delta_{(4)}^2/4/(N-3)$ can become quite large. However, for very large values of the training-sample size N, the terms $p/N^2/\delta^2$ and $\delta_{(4)}^2/4/(N-3)$ are essentially zero and, therefore, (3.13) reduces to

$$EP_N^{\mathrm{LD}} \approx \Phi\left(-\frac{\delta}{2}\left[1 + \frac{2p}{N\delta^2}\right]^{-1/2}\right). \tag{3.14}$$

One can now readily see that the approximation for EP_N^{LD} given in (3.14) is identical to expression (3.3), which is an approximate expression for the Euclidean distance classifier $EP_N^{\mathrm{F}^*}$. Thus, estimating the variances in $g_{\mathrm{LD}}(\mathbf{x})$ increases the expected error rate only slightly.

When one does not assume equal diagonal covariance structures for both pattern-class models but assumes only that each pattern-class model has a unique diagonal covariance matrix ($\mathbf{D}_2 \neq \mathbf{D}_1$), instead using (3.9) one might choose

the quadratic classifier

$$g_{\mathrm{QD}}(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_2)'\hat{\mathbf{D}}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_2)'\hat{\mathbf{D}}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) + \ln\left(\frac{|\hat{\mathbf{D}}_2|}{|\hat{\mathbf{D}}_1|}\right) + 2\ln(\alpha_1/\alpha_2). \tag{3.15}$$

Assuming multivariate normality of the pattern classes and that the true covariance matrices are diagonal and common for both populations ($\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1 = \mathbf{D}$), $\alpha_1 = \alpha_2$, $N_2 = N_1 = N$, Raudys [111] derived an asymptotic expected PMC approximation for the quadratic discriminant function (3.15), which is

$$EP_N^{\mathrm{QD}} \approx \Phi\left(-\frac{\delta}{2}\left[1+\frac{4p}{N\delta^2}\right]^{-1/2}\right). \tag{3.16}$$

From expressions (3.14) and (3.16), one can derive the important conclusion that under the equal diagonal covariance matrix assumption, the estimation of the p feature variances does not significantly affect the expected PMC of the classifier g_{LD}. This result was generalized by Deev [21]. Deev analyzed two Gaussian pattern-class models with common covariance structures composed of h independent blocks of the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 & \dots & 0 \\ 0 & \mathbf{\Sigma}_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{\Sigma}_h \end{bmatrix}. \tag{3.17}$$

Assuming equal block-diagonal covariance structures, one can utilize the statistical classifier

$$g_{\mathrm{B}}(\mathbf{x}) = \sum_{i=1}^{h} \mathbf{x}_i'\mathbf{S}_i^{-1}(\bar{\mathbf{x}}_{(1i)} - \bar{\mathbf{x}}_{(2i)}) - \frac{1}{2}(\bar{\mathbf{x}}_{(1i)} + \bar{\mathbf{x}}_{(2i)})'\mathbf{S}_i^{-1}(\bar{\mathbf{x}}_{(1i)} - \bar{\mathbf{x}}_{(2i)}) + c, \tag{3.18}$$

where $\bar{\mathbf{x}}_{(1i)}$ and $\bar{\mathbf{x}}_{(2i)}$, $i = 1, 2, \dots, h$, are p_j-dimensional sample-mean vectors for each block of the two assumed pattern classes, $\mathbf{S}_i$ is the sample maximum likelihood estimate of $\mathbf{\Sigma}_i$ for each block of the covariance structure, and c is a threshold constant. Asymptotically, as the dimensions p_j and the training-sample sizes N_1 and N_2 become arbitrarily large in a manner such that ratio $p_j/N_i \to \lambda_{ij}$, an expected PMC asymptotic approximation for block model (3.18) is

$$EP_N^{\mathrm{B}} \approx \alpha_1 \Phi\left(\frac{2c - \sum_{j=1}^{h}(\delta_j^2 - \lambda_{1j} + \lambda_{2j})}{2\sqrt{\sum_{j=1}^{h}(\delta_j^2 + \lambda_{1j} + \lambda_{2j})\frac{N_1+N_2}{N_1+N_2-p_j}}}\right) + \alpha_2 \Phi\left(\frac{-2c - \sum_{j=1}^{h}(\delta_j^2 + \lambda_{1j} - \lambda_{2j})}{2\sqrt{\sum_{j=1}^{h}(\delta_j^2 + \lambda_{1j} + \lambda_{2j})\frac{N_1+N_2}{N_1+N_2-p_j}}}\right). \tag{3.19}$$

Note that in expression (3.19), instead of the single term $T_S = \frac{N_1+N_2}{N_1+N_2-p}$, as in (3.8), we have the multiple terms $T_{S_j} = \frac{N_1+N_2}{N_1+N_2-p_j}$, $j = 1, 2, \ldots, h$.

A very interesting model allowing one to reduce the number of parameters to be estimated is the tree-dependence model introduced for discrete variables by Chow and Liu [15]. This model assumes a covariance structure in which each component of the feature vector $\mathbf{x}$ depends on only one other component. Zarudskij [160,161] represented the inverse of the first-order tree-dependence covariance matrix of a multivariate Gaussian distribution in a special manner that allows simple estimation. To denote his representation of $\mathbf{\Sigma}^{-1}$, let $\mathbf{\Sigma} = [(\sigma_{ij})]$ and $\mathbf{\Sigma}^{-1} = \mathbf{C}'\mathbf{C}$, where

$$c_{ij} = \begin{cases} \frac{1}{\sqrt{\sigma_{ij}(1-r^2_{jm_i})}} & \text{if } j = i, \\ \frac{-r_{im_i}}{\sqrt{\sigma_{ij}(1-r^2_{jm_i})}} & \text{if } j = m_i, \\ 0 & \text{if } j \neq i, m_i, \end{cases}$$

and $r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. Then, as $p \to \infty$, $N_1 \to \infty$, and $N_2 \to \infty$ such that $p/N_i \to \lambda_i$, $i = 1, 2$, Zarudskij [160,161] showed that for the tree-dependence model classifier, we have

$$EP_N^{\mathrm{TD}} \approx \alpha_1 \Phi\left(-\frac{\delta^2 - \lambda_1 + \lambda_2 - 2c}{2\sqrt{\delta^2 + \lambda_1 + \lambda_2}}\right) + \alpha_2 \Phi\left(-\frac{\delta^2 + \lambda_1 - \lambda_2 + 2c}{2\sqrt{\delta^2 + \lambda_1 + \lambda_2}}\right). \tag{3.20}$$

If $N_2 = N_1$ for very large values of the training-sample size N, expression (3.20) reduces to (3.3). Meshalkin [91] analyzed a model of independent discrete variables, where each variable can have m_j states, $j = 1, 2, \ldots, p$. When the maximum likelihood method is used to estimate the class conditional probability of each state and when one uses a plug-in classification rule, then asymptotically as $N_i \to \infty$, $i = 1, 2$, and $p \to \infty$ such that $\sum_{j=1}^{p} (m_j - 1)/N_i \to \lambda_i^*$, $i = 1, 2$, the expected PMC is approximated as

$$EP_N^{\mathrm{DS}} \approx \Phi\left(-\frac{F}{2\sqrt{F + \lambda_1^* + \lambda_2^*}}\right), \tag{3.21}$$

where $F = \sum_{j=1}^{p} \sum_{s=1}^{k_j} \frac{(P_{1js} - P_{2js})^2}{(P_{1js} + P_{2js})}$ is a discrete analog of the squared Mahalanobis distance and P_{ijs} is the probability that the jth variable in the class Π_i assumes the sth state.

In some applications, values of the discrete variables are obtained after quantization of continuous components of the feature vector $\mathbf{x}$. If one can employ a priori information concerning the pattern-class probability density function, then, instead of expression (3.21), as $N_i \to \infty$, $i = 1, 2$, and $p \to \infty$ such that $p/N_i \to \lambda_i$,

$i = 1, 2$, the expected PMC is approximated as

$$EP_N^{\mathrm{DIS}} \approx \Phi\left(-\frac{\delta^2}{2\sqrt{\delta^2+\lambda_1+\lambda_2}}\right).$$

A generalization of this result was derived by Meshalkin and Serdobolskij [93] and Serdobolskij [92]. For multivariate feature vectors possessing the block independence covariance structure (3.17) with some two pages of regularity conditions (that we omit), they analyzed the performance of the discriminant function

$$g_{\mathrm{BL}}(\mathbf{x}) = \sum_{j=1}^{h} \ln\frac{f_j(\mathbf{x}, \hat{\theta}_{1l_j}, \hat{\theta}^*_{s_j}|\Pi_1)}{f_j(\mathbf{x}, \hat{\theta}_{2l_j}, \hat{\theta}^*_{s_j}|\Pi_2)} + c,$$

where h is the number of independent blocks of the vector $\mathbf{x}$, $\hat{\theta}_{il_j}$ is the l_j-variate parameter vectors (assumed to be different in the two competing pattern classes), $\hat{\theta}^*_{il_j}$ is the s_j-variate parameter vectors (assumed to be common for both pattern classes), and c is a classification threshold. Let $l = \sum_{j=1}^{h} l_j$, $s = \sum_{j=1}^{h} s_j$, and $N_1 > N_2$. Thus, if $h \to \infty$, $N_1 \to \infty$, and $N_2 \to \infty$ such that $l/N_1 \to \lambda_1$, $l/N_2 \to \lambda_2$, $s/N_1 \to t_1$, and $s/N_2 \to t_2$, then the conditional PMC tends to (3.20) where δ now denotes a generalization of the Mahalanobis distance between the two populations.

This fundamental result of Meshalkin and Serdobolskij is exact if both the sample size and the dimensionality are large. For moderate values of p and N, the more complex asymptotic formulae such as (3.10) give slightly higher accuracy. Expression (3.20) is important if one analyzes discriminant functions with structured covariance matrices (SCM) such as the tree-type dependency structure. In this case, one can describe the resulting discriminant functions assuming the SCM model by relatively few parameters. According to Meshalkin and Serdobolskij's result, the SLDF under the SCM model is relatively insensitive to the training-sample size. Han [49], Morgera and Cooper [94], and Ge and Simpson [41] have carried out investigations of classification rules derived under the SCM model. Their results agree that classifiers designed by applying the SCM model are less sensitive to relatively small training-sample sizes.

Other asymptotic expansions for general families of class densities $f_i(\mathbf{x}|\theta_i)$ were obtained by Golcov and Troitsky [44] and Kharin and Duchinskas [68]. The latter two Soviet Union researchers showed that for regular density families with a fixed observation space, the expected PMC tends to the Bayes error as $\frac{\varsigma_1}{N_1}, \frac{\varsigma_2}{N_2}$, where ς_1 and ς_2 depend on the prior probabilities α_1 and α_2 and the parameters of the pattern-class densities. They applied a Chibisov [14] expansion of the maximum likelihood estimates to obtain the coefficients ς_1 and ς_2. This approach was later used by Kharin [66] to obtain an explicit expected PMC expression for the multivariate Gaussian case and by Kharin [61,67] to obtain an expected PMC expression when the actual pattern-class densities differ from the assumed pattern-class densities.

Apart from the derivation of asymptotic expected PMC approximations, former Soviet Union researchers have developed roughly "exact" expected PMC expressions written in the form of infinite series or integrals of certain functions. Such

expressions have been formulated by Raudys [109] for the SEDC, Pikelis [99] for the SLDF with independent features, Raudys and Pikelis [127] for the SLDF, and Raudys [113] for the SQDF. These expected PMC expressions are summarized and tabulated in Raudys and Pikelis [128].

4. Regularized statistical discriminant analysis for the SLDF

We now briefly consider the two-population classification rules designed to deal with situations when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, and the training-sample sizes are very small relative to the feature dimension p. When the total training-sample size $n = N_1 + N_2$ is less than p, the pooled-sample covariance matrix $\mathbf{S}$ is singular and, thus, a problem for the expression of the SLDF arises. Harley [51] and DiPillo [25,26] formulated the problem of regularization of the covariance matrix in the small training-sample size situation. They suggested that one replace the estimator $\mathbf{S}^{-1}$ with the estimator $\mathbf{S}_{\mathrm{R}}^{-1}$, where $\mathbf{S}_{\mathrm{R}} = \mathbf{S} + t \times \mathbf{I}$, $\mathbf{I}$ is the $p \times p$ identity matrix, and $t > 0$ is the regularization parameter. This estimator yields the *sample regularized linear discriminant function* (SRLDF)

$$g_{\mathrm{RL}}(\mathbf{x}) = \mathbf{x}'\mathbf{S}_{\mathrm{R}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}_{\mathrm{R}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{4.1}$$

For more details concerning linear and nonlinear regularized discriminant analysis, see [36,78,79,89,134]. Barsov [7,8] analyzed the performance of the SRLDF when the pattern classes are Gaussian with a common unknown covariance matrix and different known mean vectors. He showed that if $\alpha_1 = \alpha_2$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known but $\boldsymbol{\Sigma}$ is unknown, then asymptotically for $p \to \infty$, $n = N_1 + N_2 \to \infty$, and $p/n \to \lambda < 1$, an expected PMC approximation for the SRLDF is

$$EP_N^{\mathrm{RL}} \approx \Phi\left(-\frac{\boldsymbol{\Lambda}'\mathbf{A}(t)\boldsymbol{\Lambda}}{2\sqrt{\boldsymbol{\Lambda}'\mathbf{A}^2(t)\boldsymbol{\Lambda}}}\left[1 - \frac{t^2 m^2(t)}{n}(\operatorname{tr}\mathbf{A}^2(t))\right]^{1/2}\right), \tag{4.2}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathbf{A}(t) = (\boldsymbol{\Sigma}^{-1} + tm(t)\mathbf{I})^{-1}$, $m(t)$ satisfies the equation $t(\frac{1}{m(t)} - 1) \cong (1/n)[\operatorname{tr}\mathbf{A}(t)]$, and $\alpha \approx \beta$ is defined as $\alpha = \beta(1 + o(1))$ asymptotically. Since $\boldsymbol{\Sigma}$ is unknown, Barsov [8] proposed estimating the optimal regularization parameter t by minimizing the estimate of the asymptotic expected error rate

$$\hat{E}P_N^{\mathrm{RL}} \approx \Phi\left(-\frac{\boldsymbol{\Lambda}'\mathbf{S}_{\mathrm{R}}^{-1}\boldsymbol{\Lambda}}{2\sqrt{\boldsymbol{\Lambda}'\mathbf{S}_{\mathrm{R}}^{-1}\mathbf{S}\mathbf{S}_{\mathrm{R}}^{-1}\boldsymbol{\Lambda}}}\left(1 - \frac{p}{n} + \frac{\operatorname{tr}(\mathbf{S}_{\mathrm{R}}^{-1})}{n}\right)^{1/2}\right).$$

Another approach to solving the regularization problem for the SLDF has been formulated by Serdobolskij [139], who proposed using a generalized regularized estimator of $\boldsymbol{\Sigma}$ of the form $\mathbf{S}_{GR} = \int_{\Omega}(\mathbf{S} + t\mathbf{S})\,\mathrm{d}\eta(t)$, where $\eta(t)$ is a weighting function. Serdobolskij [139] also proposed a method of determining the optimal weighting function $\eta(t)$.

Raudys and Skurikhina [131] and Raudys et al. [132] obtained an explicit asymptotic approximation for EP_N^{RL} as a function of the parameters of two multivariate Gaussian populations with a common covariance structure, common training-sample sizes, and the regularization parameter t. They derived an expression for the conditional PMC in terms of the conditional mean and variance of the SRLDF. However, when both the dimension p and the training sample sizes N are large, the mean and variance of the SRLDF tend to constants, and both the conditional and expected PMCs are approximated as

$$EP_N^{\mathrm{RL}} \approx \Phi\left(-\frac{\delta}{2}\frac{T_t}{\sqrt{T_\mu T_S}}\right), \tag{4.3}$$

where

$$T_\mu = 1 + \frac{1}{N}\left(1 + \frac{2p}{\delta^2}\right) + \frac{p}{N^2\,\delta^2}, \quad T_S = \frac{2N}{2N-p},$$
$$T_t = \frac{(1 + 2tb_1T_S + t\,\mathrm{tr}(\mathbf{D}^{-1})/N)^{1/2}}{1 + tb_2T_S},$$
$$b_i = \frac{\boldsymbol{\mu}\mathbf{D}^{-1}\boldsymbol{\mu}}{\delta^2}\gamma_i + \frac{\mathrm{tr}\,\mathbf{D}^{-1}}{2N-p},$$
$$\gamma_1 = 1 \quad \text{and} \quad \gamma_2 = \left(1 + \frac{2\,\mathrm{tr}\,\mathbf{D}^{-1}}{N\boldsymbol{\mu}\mathbf{D}^{-1}\boldsymbol{\mu}}\right)\left(1 + \frac{2p}{N\delta^2}\right)^{-1}.$$

We note that a portion of expression (4.3) coincides with the analogical expression for the standard Fisher discriminant function without regularization as described in expression (3.7). The difference between (4.3) and (3.7) occurs in the additional term T_t, that is responsible for the regularized estimate of the covariance matrix. This asymptotic expected PMC approximation is an explicit function of the regularization constant t. Therefore, (4.3) can be used as a criterion function to obtain an approximately optimal regularization value of t, that is,

$$t_{\mathrm{opt}} = \frac{1}{\beta_1 T_S} - \frac{1}{\beta_2 T_S + \frac{\mathrm{tr}\,\mathbf{D}^{-1}}{4N}}. \tag{4.4}$$

Note in expression (4.4) that t_{opt} is a monotonic decreasing function of the common training-sample size N. Using the asymptotic expected PMC expressions (4.3) and (4.4), one can analytically examine the differences in the expected PMCs between the SRLDF defined in (4.1) and the SLDF defined in (3.6) for various covariance structures and ratios of the training-sample sizes to the dimensionality. We note also that the derivation of expression (4.3) is based on a Taylor series expansion of $(\mathbf{S} + t\mathbf{I})^{-1}$ that is accurate only for very small values of t.

An alternative to a regularized SLDF in the case when n is less than p is the standard Fisher classifier using the pseudoinverse of the pooled-sample covariance matrix $\mathbf{S}$. Raudys and Duin [125] have shown that the expected error rate of such a linear classifier has a peaking behavior. That is, the expected error rate decreases as N ranges from 1 to $p/2$ and then increases as N ranges from $p/2$ to p [125].

5. Density estimation-based statistical classifiers

In many applications we have data contaminated by "noise," i.e., the measurement vectors have many atypical observations, or outliers (see, e.g., [54]). In some cases the class indexes of the training vectors are determined with errors. For this type of data, special decision-making rules should be derived. Randles et al. [105] suggested generalized linear and quadratic discriminant functions using robust estimates. Two additional important DA topics studied by Soviet Union researchers are the topics of classifier design and error-rate analysis of robust and nonparametric statistical classifiers. Kharin [63,67] suggested the SLDF when each training sample is contaminated by observations from the other class and derived an asymptotic approximation for the expected PMC.

To consider his results, let $1-t_i$ be the probability that an observation vector labeled as belonging to class i actually belongs to class j, and let t_i be the probability that an observation is mislabeled as belonging to population Π_i, $i=1,2$. Kharin [65,67] derived an expression for the expected PMC of the standard linear Fisher discriminant function with parameters estimated from the contaminated data

$$EP_N^{\mathrm{LC}} = P_{\mathrm{B}} + \gamma(t_1,t_2) + \sum_{i=1}^{2}\left(\frac{\alpha_i}{N_i} + \frac{t_i\beta_i}{N_i}\right) + o(\tau_0^2),$$

where $\tau_0 = \max\{\tau_1,\tau_2\}$, $\tau_i = \max\{t_i,(1+t_i)/N_i\}$, δ,α_i, and β_i are functions of the pattern-class densities $f_i(\mathbf{x}|\theta_i)$, and γ depends on the contamination levels t_1 and t_2.

Kharin suggested modifying the traditional plug-in SLDF rule to reduce the negative influence of contaminated training samples. For example, for the SLDF with known covariance but unknown means, Kharin has proposed using the estimators $\tilde{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i + (-1)^i t_i(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, $i=1,2$. For the case when $\boldsymbol{\Sigma}_2 \neq \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_i$ is unknown, $i=1,2$, the resulting quadratic discriminant function is considerably more complex than the plug-in SQDF defined in (3.9) [65]. For details see [67].

Popular nonparametric techniques to classify multimodal populations are the nonparametric classifiers such as the k-nearest neighbor ($k-NN$) classifier and the kernel-density-estimation classifier using the Parzen window [28,39]. Raudys et al. [130] compared 13 different classification algorithms and found that the SLDF and PW classifiers perform the best for the classification of several real-world pattern recognition problems. Kharin [62–64] analyzed the expected PMC's of the $k-NN$ and the Parzen window (PW) classifiers. The statistical discriminant function using the PW is

$$g_{\mathrm{PW}}(\mathbf{x}) = \sum_{i=1}^{2}\frac{(-1)^{3-i}}{N_i}\sum_{j=1}^{N_i} K\left(\frac{(\mathbf{x}-\mathbf{x}_{ij})'(\mathbf{x}-\mathbf{x}_{ij})}{t}\right),$$

where $K(\cdot)$ is the kernel function, t is the smoothing parameter, and $\mathbf{x}_{ij}$ is the jth training observation from the ith pattern-class population. Kharin's asymptotic expected PMC expression for $g_{\mathrm{PW}}(\mathbf{x})$ is

$$EP_N^{\mathrm{PW}} = P_{\mathrm{B}} + t^4\Delta P_\infty + \frac{H}{\lambda^p N} + O(1/N), \tag{5.1}$$

where P_B is the Bayes error and ΔP_∞ and H are functions of the pattern-class densities $f_i(\mathbf{x}|\theta_i)$, $i=1,2$. Kharin utilized expression (5.1) as a criterion function and assumed that $t=bN^{-\gamma}$ in order to determine the asymptotically optimal values of $\gamma_{\text{opt}}=\frac{1}{p+4}$ and $b_{\text{opt}}=\frac{p\tau}{4}$, where τ is a function of the pattern-class densities $f_i(\mathbf{x}|\theta_i)$, $i=1,2$.

Raudys [116] determined that when the ratio of the training-sample size N to the dimension p is small, the distribution of the PW density function estimators $\hat{f}_i^{\text{PW}}(\mathbf{x}|\theta_i)$, $i=1,2$, is highly skewed. This phenomenon explains the poor accuracy of the asymptotic approximation EP_N^{PW} given in (5.1). Also, Raudys [117] performed a simulation study with multivariate spherical Gaussian data which shows that dependent on the dimensionality p, the distance between the pattern classes δ, and the value of the smoothing parameter t, the quantity EP_N^{PW} tends to the asymptotic error $P_\infty^{\text{PW}}(t)$ at the rate of $N^{-1/3}$, $N^{-1/2}$, or even N^{-1}. Theoretically, and via a Monte Carlo simulation study, Raudys [116,119] showed that the training-sample size required to achieve a specific value of $EP_N^{\text{PW}}/P_\infty^{\text{PW}}$ grows exponentially as a function of the dimensionality p at the rate

$$[\alpha(EP_N^{\text{PW}}/P_\infty^{\text{PW}},t)]^p, \tag{5.2}$$

where ratio $\kappa=EP_N^{\text{PW}}/P_\infty^{\text{PW}}$ is the learning quality parameter given in Section 3. However, the function $\alpha(\kappa,t)$ must be determined by trial and error from experiments when the necessary parameters are unknown.

A practical method of determining t_{opt} for the classifier PW was proposed by Raudys [114] and Raudys and Jain [126]. This method is to approximate t_{opt} by determining the smoothing parameter t that corresponds to the minimum leave-one-out conditional PMC estimate evaluated over a finite number of smoothing parameter values t_i, $i=1,2,\ldots,r$. However, this method was quite computationally intense. Therefore, for estimating the classification error in a high-dimensional feature space, Raudys [114] proposed calculating the distances $D_{\alpha\beta}^{ij}=(\mathbf{x}_{ij}-\mathbf{x}_{\alpha\beta})'(\mathbf{x}_{ij}-\mathbf{x}_{\alpha\beta})$ between a pair of training vectors, $\mathbf{x}_{ij}$ and $\mathbf{x}_{\alpha\beta}$, to determine the contribution for all r nonparametric density estimates $K\left(\frac{D_{\alpha\beta}^{ij}}{t_1}\right), K\left(\frac{D_{\alpha\beta}^{ij}}{t_2}\right), \ldots, K\left(\frac{D_{\alpha\beta}^{ij}}{t_r}\right)$. Thus, utilizing this approach, one computes the distance $D_{\alpha\beta}^{ij}$ only once and, therefore, lessens the computational demand.

When the measurement feature vectors are dependent and categorical, each taking k_j distinct states or values, the conventional statistical classification rule is a multinomial-based rule (see [16,77]). To design this type of classifier where there are $m=\prod_{j=1}^p k_j$ possible categories, one must estimate m conditional probabilities p_{ij}, $j=1,2,\ldots,m$, for each pattern class. Matrosov [82] obtained several error bounds for the conditional PMC of a multinomial-based classifier.

In [46,104,123], exact expressions for the expected PMC and the expectation of the resubstitution error-rate estimator were derived, and these expectations were tabulated for a model of an unfavorable distribution of the values $p_{11},\ldots,p_{2m}$ (see

also [123, Section 3.8]). The training-sample sizes necessary to achieve a given learning quality grow exponentially with an increase in the dimensionality of the discrete feature vector p, as in the expression (5.2) for the PW classifier, provided the number of states k_j is the same for all features. The analysis indicates that for the realistic distribution of the values $p_{11}, \ldots, p_{2m}$, the multinomial classifier can be designed using comparatively small training-sample sizes.

6. Minimum empirical-error linear classifiers

An alternative approach to the parametric-based plug-in SDA is to determine a linear classification rule of the form

$$g_{\mathrm{AL}}(x) = w_0 + \sum_{i=1}^{p} w_i x_i, \tag{6.1}$$

where one estimates the classifier weights, or coefficients, w_0 and $\mathbf{w} = (w_1, w_2, \ldots, w_p)'$ by minimizing an empirical error rate, such as the resubstitution error-rate estimate $\hat{P}_{\mathrm{R}}$ or some other empirical loss function. Minimization of the standard sum-of-squares loss function (adaline algorithm, [158])

$$\mathrm{sse} = \sum_{i=1}^{2} \sum_{j=1}^{N_i} (y_{ij} - (\mathbf{w}'\mathbf{x}_{ij} + w_0))^2, \tag{6.2}$$

where y_{ij} is equal to either -1 or 1 (depending on the class membership of $\mathbf{x}_{ij}$) can lead to the SLDF [70]. However, the loss function (6.2) does not minimize the number of misclassifications in the training set.

Smith [141] changed the quadratic loss function and suggested two modifications of the standard sum-of-squares cost functions: a relaxation and a fixed increment algorithm where distant-atypical observations have a smaller contribution to the loss function. In artificial neural network training [53,123], one uses the following modification of the quadratic loss function:

$$\mathrm{sse} = \sum_{i=1}^{2} \sum_{j=1}^{N_i} (y_{ij} - f(\mathbf{w}'\mathbf{x}_{ij} + w_0))^2,$$

where $f(c)$ is a nonlinear activation function, such as the sigmoid function $f(c) = 1/(1 + \exp(-c))$, that varies between 0 and 1. While minimizing the cost function of a nonlinear single-layer perceptron classifier by means of an iterative-gradient-descent optimization algorithm, one can obtain seven known statistical classification rules. These seven statistical classifiers are (1) the EDC, (2) the SRLDF, (3) the SLDF, and (4) its modification using the pseudoinverse of the pooled-sample covariance matrix, (5) a robust linear classifier, as well as (6) the minimum empirical error, and (7) maximum margin classifiers [122,123]. References to a dozen other statistical procedures that obtain the minimum empirical error classifier can be found in [121].

As in the parametric-based classifier design approach, the conditional PMC of the classifier, denoted by P_N, depends on both the training-sample sizes N_1, N_2 and the dimensionality p. The first result concerning this topic was published by Widrow and Hoff [158], who concluded that the sample size required to achieve a given signal to noise ratio for the adaline-type algorithm should increase proportionally to the number of inputs. Smith [142] analyzed the relaxation and fixed increment algorithms and derived asymptotic expressions for the expected PMC that are similar to those derived by Efron [29]. Smith concluded that these two algorithms are more sensitive to the training-sample size than the standard adaline algorithm.

Due to the nonlinearity of the loss function, one encounters great difficulty in obtaining analytical expressions for the error rate similar to the parametric classifiers based on multivariate Gaussian pattern classes. A number of bounds for the actual and estimated error rates of the minimum empirical-error classifier were obtained by Vapnik and Chervonenkis [154,155] and Vapnik [152]. One such upper bound on the conditional PMC of (6.1) is

$$P_N < \hat{P}_R + \left(\frac{p(\ln(n/p)+1) - \ln\eta}{2n}\right) \times \left(1 + \left[1 + \frac{4n\hat{P}_R}{p(\ln(n/p)+1) - \ln\eta}\right]^{1/2}\right) \tag{6.3}$$

with probability $1-\eta$, $(n = N_1 + N_2)$.

The upper bound (6.3) has been obtained for the least-favorable distribution of training-pattern vectors and results in a very pessimistic estimate of the number of samples required to adequately train the classifier. Therefore, a number of modifications of this bound have been suggested. E.g., Cherkassky and Mulier [13, Section 4.3.1] propose that one use

$$P_N < \hat{P}_R + \left(\frac{a_1 p(\ln(a_2 n/p)+1) - \ln\eta}{2n}\right) \times \left(1 + \left[1 + \frac{4n\hat{P}_R}{a_1 p(\ln(a_2 n/p)+1) - \ln\eta}\right]^{1/2}\right),$$

where the constants a_1, a_2 must be in the range $0 < a_1 \leqslant 4$, $0 < a_2 \leqslant 2$. Unfortunately, for actual real-data classification problems, good empirical values for a_1 and a_2 are unknown. Cherkassky and Mulier also noted that the above bounds are tighter when the training set size N is large. Both bounds mentioned agree with conclusions that follow from asymptotic analysis of the parametric rules. Namely, the increase in the expected error rate depends on the ratio N/p. The reader can find more details on this topic in the original papers by Vapnik and Chervonenkis [154,155], and in booksa by Vapnik [152,153], Vidyasagar [156], and Cherkassky and Mulier [13].

Raudys [120,121] derived an asymptotic approximation for the expected PMC (the generalization error) of a linear *zero empirical error* (ZEE) classifier for a specific

input data model. He assumed the data model configuration of two spherical multivariate Gaussian pattern classes $N(\boldsymbol{\mu}_1, \mathbf{I})$ and $N(\boldsymbol{\mu}_2, \mathbf{I})$ and training-sample sizes $N_1 = N_2 = N$, and analyzed the hypothetical randomized training scenario. For this training method one repeatedly generates many random discriminant hyperplanes $w_0 + \mathbf{w}'\mathbf{x} = 0$, where $\mathbf{w} = (w_1, w_2, \ldots, w_p)'$, according to a certain prior density function $\psi_{\text{prior}}(w_0, \mathbf{w})$. One then selects those discriminant hyperplanes that classify the training-data vectors without error (i.e., $\hat{P}_R = 0$). Researchers have considered two types of prior densities $\psi_{\text{prior}}(w_0, \mathbf{w})$ in order to obtain numerical values of the expected PMC: (1) the multivariate Gaussian density and (2) the density of the weight vector of the SEDC classifier trained on additional data sets. The analysis of the second model is motivated by the fact that after the first learning iteration, a single-layer perceptron can realize the decision boundary of the SEDC [122,123]. Use of the Gaussian prior implies that no additional information is used to design this type of classifier.

For the first training model, where the prior distribution of the weights is spherically Gaussian, the expected PMC was calculated from the exact theoretical formulae [121,124]. To compare results with formulae for the parametric classifiers presented in a very simple way (Eqs. (3.3), (3.7), and (3.10)), Raudys [121] used the tabulated data to approximate the expected PMC for the ZEE classifier by means of the similar formula

$$EP_N^{\text{ZEE}} \approx \Phi\left(-\frac{\delta}{2}\left[1 + (1.6 + 0.18\delta)\left(\frac{p}{N}\right)^{1.8-\delta/5}\right]^{-1/2}\right). \tag{6.4}$$

This approximation is comparable to the expected PMC approximations (3.2), (3.7), and (3.10). Simple but accurate, asymptotic expressions for the ZEE classifier was proposed by Diciunas and Raudys [24] and Diciunas [23]

$$EP_N^{\text{ZEE}} \approx \Phi\left(-\frac{\delta}{2}\right) + \frac{1}{4}\Phi\left(\frac{\delta}{2}\right)\frac{p}{N}. \tag{6.5}$$

In addition, exact analytical formulae for expected PMC of linear nonzero empirical error and ZEE classifiers with arbitrary margin width were derived [23]. Upon examination of numerical values calculated from these expressions, one can see that in the very small training-sample size cases, the ZEE linear-classifier approach outperforms the SLDF.

During the last dozen years, considerable attention was given to the complexity versus sample size problems in the artificial neural network literature. Researchers have derived a number of asymptotic formulae for the expected classification error by utilizing different mathematical and theoretical physics techniques. These techniques include the double asymptotic approach (thermodynamical limit), the random search optimization procedure, the replica symmetry technique, and an annealed approximation method [3–5,47,52,76,90].

7. Estimation of the classification error rate

Estimation of the conditional PMC of a classifier constructed from sample data is one of the most important aspects of designing any discrimination algorithm. The simplest error-rate estimator is to classify test observations to determine the empirical frequency of misclassifications. This method is called the hold-out error counting estimator, which we denote by $\hat{P}_{\mathrm{H}}$. If the training samples are used to evaluate the error rate instead of a set of test data vectors, this error-rate estimator is called the resubstitution error-counting estimator, denoted by $\hat{P}_{\mathrm{R}}$. Unfortunately, for small training-sample sizes, one adapts to training data, and the resubstitution error-rate estimator becomes optimistically biased. A number of alternative conditional PMC estimation methods have appeared in the discriminant analysis and statistical pattern recognition literature [30,39,43,50,73,89,147]. The most popular unbiased conditional PMC estimator is the hold-out, or cross-validation, method. However, when the training-sample size is small relative to the feature dimensionality p, an experimenter takes a large risk by splitting the available data into training and hold-out samples.

An alternative to the hold-out method is the leave-one-out error-rate estimator $\hat{P}_{\mathrm{L}}$ proposed by Lachenbruch and Mickey [73]. This estimator consists of creating all possible classifiers of interest using $n-1$ observations, applying each classifier to the corresponding hold-out observation, and then counting the proportion of the hold-out observations which are misclassified. In the Soviet statistical classification literature, this procedure was known as the *sliding egzam* conditional PMC estimator and was proposed by Brailovskij [11]. In a generalization of this method, one skips subsets of training vectors from the training data sequentially and uses these subsets to estimate the classification error. In the pattern recognition literature, this approach is known as the *rotation* method. In the ANN literature it is known as the *k-fold cross-validation* method.

Another well-known classifier-performance estimator is a bootstrap error-rate estimator, proposed by Efron [30]. The bootstrap conditional PMC estimate is obtained if one subsamples r times from a random training sample of size N_i, $i=1,2$, and then uses these subsamples to estimate the mean difference Δ between the resubstitution estimated error rate $\hat{P}_{\mathrm{R}\alpha}$ and the conditional estimated error rate $\hat{P}_{N\alpha}$. The resulting bootstrap bias estimator is

$$\hat{\Delta}_{N\mathrm{R}}^{\mathrm{B}} = \frac{1}{r}\sum_{\alpha=1}^{r}(\hat{P}_{N\alpha} - \hat{P}_{\mathrm{R}\alpha}). \tag{7.1}$$

The bias estimator $\hat{\Delta}_{N\mathrm{R}}^{\mathrm{B}}$ is then added to the resubstitution error-rate estimator $\hat{P}_{\mathrm{R}}$ to reduce the optimistic bias of the resubstitution estimator. An early analog of this error-rate estimation technique was proposed by Pinsker [101].

Researchers have thoroughly studied parametric error-rate estimators for the two-group Gaussian population model with common covariance matrices. The

estimators are of the form

$$\hat{P}^{\mathrm{F}}_{\mathrm{Method}} = \Phi(-\hat{D}^{(\mathrm{Method})}/2), \tag{7.2}$$

where Φ are the standard normal cumulative distribution function and various error-rate estimators utilize distinctly different Mahalanobis-distance estimates $\hat{D}^{(\mathrm{Method})}$. For the *D-method* given in (7.2), one utilizes an estimator of the Mahalanobis distance of the form

$$\hat{D}^2 = (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Because the estimator $\hat{D}^2$ is biased, a number of modifications to several parametric error estimates have been suggested. The Western research on this topic is summarized in a number of review papers and books mentioned above. Enukov [31,32] analyzed the performance of a sample Mahalanobis-distance-based parametric error-rate estimator $\hat{P}^{\mathrm{F}}_D = \Phi(-\hat{D}/2)$ for the SLDF asymptotically when $N_i \to \infty$, $i = 1, 2$, and $p \to \infty$ and found it to be biased:

$$E\hat{P}^{\mathrm{F}}_D \approx \Phi\left(-\frac{\delta}{2}\left[\left(1 + \frac{2p}{N^2\delta^2}\right)\frac{2N}{2N-p}\right]^{1/2}\right). \tag{7.3}$$

He proposed an unbiased estimate of the expected classification error based on expression (3.8) of the form

$$\hat{P}^{\mathrm{F}}_{\mathrm{unbiased}} = \Phi\left(-\frac{\hat{D}}{2}\left(1 + \frac{2p}{N^2\hat{D}^2}\right)\frac{2N}{2N-p}\right).$$

Bulygin [12] extended Enukov's results to the many-categories case (k pattern classes). Asymptotic expansions of the first and second moments of $\hat{P}_{\mathrm{R}}, \hat{P}_{\mathrm{H}}$, and $\hat{P}_{\mathrm{L}}$ in terms of $1/N_i$, $i = 1, 2$, for regular pattern-class densities were presented in [27].

Raudys [112] considered the resubstitution error-counting estimator and showed that for the SEDC, as both $N \to \infty$ and $p \to \infty$,

$$E\hat{P}^{\mathrm{F}*}_{\mathrm{R}} \approx E\hat{P}^{\mathrm{F}*}_D \approx \Phi\left(-\frac{\delta}{2}\left(1 + \frac{2p}{N^2\delta^2}\right)^{1/2}\right) \tag{7.4}$$

if the common pattern-class covariance matrix is known. Also, for the SLDF Pivoriunas and Raudys [102] considered the resubstitution method and found that $E\hat{P}^{\mathrm{F}}_{\mathrm{R}} \approx E\hat{P}^{\mathrm{F}}_D$ if the common pattern-class covariance matrix is estimated. For both the statistical classifiers SEDC and SLDF, the expectation of the LOO conditional PMC estimator converges to the asymptotic expected PMC expressions (3.3) and (3.7), respectively. Comparison of expressions (7.3) and (7.4) with (3.2) and (3.7), respectively, shows that for the EDC and SLDF constructed with large training-sample sizes, one can see that $E\hat{P}^{\mathrm{F}}_D$ and $E\hat{P}^{\mathrm{F}}_{\mathrm{R}}$ are symmetric with respect to the asymptotic PMC, P_∞. Similar observations were made for other classification rules [5]. Thus, the expressions $\hat{P}^{*}_{\infty} = \frac{\hat{P}_{\mathrm{R}} + \hat{P}_{\mathrm{L}}}{2}$ and $\hat{P}^{**}_{\infty} = \kappa\hat{P}_{\mathrm{R}}$ have been suggested as estimators of the asymptotic PMC, P_∞. An entity $E\hat{P}_N = (\kappa^2)\hat{P}_{\mathrm{R}}$ has been suggested

as an estimator of the expected PMC, where $\kappa = EP_N/P_\infty$ is the learning ratio that can be determined from asymptotic expansions, exact formulae, or tables [128].

The dispersion of different error-rate estimators is also very important. For parametric estimators such as the D-method, $\hat{P}_D^{\mathrm{F}} = \Phi(-\hat{D}/2)$, one should consider that the sample Mahalanobis distance D^2 is a noncentral F random variable. Using the first terms of the Taylor expansion, one can show that

$$V[\hat{P}_D^{\mathrm{F}}] = \frac{\phi(\delta/2\sqrt{T_\mu T_S})^2}{16N} T_S(\delta^2 T_\mu T_S + 8). \tag{7.5}$$

Similar expressions can be found for other more sophisticated parametric estimates [30–32,73,89]. From theoretical analysis and simulation studies, Raudys and Vaitukaitis [133] also concluded that the variances of four error-counting estimators (the resubstitution, hold-out, LOO, and bootstrap bias-corrected) are all well approximated by

$$V(\hat{P}_i) \approx \frac{E(\hat{P}_i)(1 - E(\hat{P}_i))}{n}, \tag{7.6}$$

where $\hat{P}_i$, $i = 1, 2, 3, 4$, is one of four error-rate estimators given immediately above, E denotes the expectation operator, and n is the number of observation vectors utilized to calculate the number of misclassifications.

Expression (7.6) explains why the resubstitution error-rate estimator has the smallest variance when compared to the three other nonparametric methods. In general, the larger the mean value of the error-rate estimator, the larger its variance. Expressions (7.6) and (7.5) also yield conditions where the variances of the parametric-based conditional PMC estimators are smaller than the nonparametric (error-counting) estimators. These facts, combined with the variance of the conditional classification error given in (3.11), constitute theoretical evidence that the variance of the bootstrap error-rate estimator is as large as the variance of the LOO estimator. Raudys and Vaitukaitis [133] and Raudys [118] also determined that the variance of the bias-correcting term (7.1) of the bootstrap estimator is

$$V(\hat{\Delta}_{N\mathrm{R}}^{\mathrm{B}}) \approx \frac{P_\infty(1 - P_\infty/\kappa^2)}{nr\kappa^2}, \tag{7.7}$$

where $\kappa = EP_N/P_\infty$, the learning quantity discussed above.

Expression (7.7) can be used to determine r, the number of bootstrap subsamples. A simple calculation shows that $10 < r < 20$ is usually a sufficiently large number of bootstrap subsamples to estimate the bias correction term.

8. Model complexity and feature subset selection

Rao [106] first emphasized the problems that can arise in cases where the number of training samples is close to the number of dimensions. Allais [2] and Hughes [55] were the first to formulate and analyze the problem of classifier-performance dependence on the dimensionality and training-sample sizes. The dependency was

also explored by Van Ness and Simpson [151] and Van Ness [149,150]. With an increase in the number of features p and fixed training-sample sizes N_1 and N_2, the expected classification error EP_N^{F} diminishes at first, then levels off and afterward begins to increase. This relationship between the training-sample size and feature dimension is known as the "peaking phenomenon" of a statistical classifier. In the former USSR this classifier property was first discovered by Lbov [75].

A similar conclusion is valid for the model (the classifier) complexity. From expressions (3.3) and (3.7), one can see that the SLDF is more sensitive to the training-sample size N than the SEDC and that $P_{\infty}^{\mathrm{E}} \geqslant P_{\infty}^{\mathrm{F}}$. Therefore, if the training ratios N_i/p, $i = 1, 2$, are small, one may prefer to use the simpler SEDC, which does not require the estimation of $\mathbf{\Sigma}$ and, therefore, has fewer parameters to estimate than the SLDF. On the other hand, if N_i/p, $i = 1, 2$, are large, one may prefer the more complex SLDF, which includes directional information in the estimated covariance matrix $\mathbf{S}$, as the statistical classifier of choice. The intersection of the expected error rates EP_N^{F} and EP_N^{E}, when plotted versus the training-sample size N, resembles scissors and, therefore, this phenomenon was later called the scissors effect.

Apparently, Raudys [110] and Kanal and Chandrasekaran [59] were the first to formulate the concept of matching the classifier complexity with the training-sample size and data dimensionality. Vapnik and Chervonenkis [155] formulated the necessity of choosing the classifier's model complexity in accordance with the training-sample size as a principle of structural risk minimization. Here, a sequence of the classifiers of increasing complexity is tested. The model is selected according to the smallest sum of the empirical error and an analytically determined penalty term. The problem of matching the classifier complexity with the training-sample size is also known as a "bias variance dilemma," or "Occam's razor," among other names (see, e.g., [9]).

A considerable amount of research has been done by former Soviet Union researchers concerning feature-space dimension reduction, or classification-rule complexity reduction. Dimension reduction has long been proposed as necessary when the training-sample sizes are small in comparison with the feature dimensionality. One method of combating this "curse of dimensionality" is to perform some type of variable selection to decrease the data dimensionality and, therefore, supposedly decrease the conditional PMC [75]. We shall refer to this variable selection process as feature-subset selection.

An alternative to feature selection is to design a small number of simple-structured classifiers and fuse their decision by means of a trainable fusion rule. This approach is popular in the West under the name multiple classification systems [69], combining classifiers, etc. In the former Soviet Union, this research direction began in the early 1980s. Results of the first decade of research in this area are summarized in the Rastrigin and Erenstein monograph [107].

In the West many algorithms have been proposed for feature subset selection in statistical discriminant analysis. The most widely researched algorithms have been for parametric classification algorithms and are based on statistical hypothesis tests. These methods are usually referred to as stepwise variable selection methods. Papers

concerning stepwise variable selection algorithms include Brailovskij [11], Lbov [74], Urbakh [148], Enukov [31], McKay and Campbell [83,84], Costanza and Afifi [17], Costanza and Ashikaga [18], Farver and Dunn [34], Habbema and Hermans [48], McLachlan [88], Fatti and Hawkins [35], Ganeshanandam and Krzanowski [40], Tcheponis et al. [146], and Aivazyan et al. [1].

Both the tasks of model selection and feature subset selection are performed using an inexact sample-based performance estimator. In the model selection problem, we choose the important features using a validation set to evaluate the performance of the competing features or variates [33]. Therefore, $P_{\text{true}}^{\text{selection}}$, the actual performance of the model selected (measured on an independent test set) is worse than $P_{\text{apparent}}^{\text{selection}}$, the "apparent" performance of this model measured on the validation set. The problem of the adaptation bias in feature subset selection was first considered in Meshalkin [92]. The concepts of apparent and true selection errors $P_{\text{true}}^{\text{selection}}$, and $P_{\text{apparent}}^{\text{selection}}$, respectively, were introduced in [115].

Raudys [115] explored the feature-subset-selection problem assuming the pattern-class models of two multivariate spherically Gaussian classes $\mathrm{N}(\boldsymbol{\mu}_1, \mathbf{I})$ and $\mathrm{N}(\boldsymbol{\mu}_2, \mathbf{I})$, where $\mathbf{I}$ is the p-dimensional identity matrix and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})'$, $i = 1, 2$. He further assumed that the individual feature-quality values, $\Delta\mu_j = \mu_{1j} - \mu_{2j}$, were $\mathrm{N}(0, \sigma^2)$ random variables. Feature subsets of size q ($q \ll p$) were selected with estimated feature-quality values $\Delta\bar{\mathbf{x}}_j = \bar{\mathbf{x}}_{1j} - \bar{\mathbf{x}}_{2j}$. Assuming this model, he showed that as $p \to \infty$ and $N = N_1 = N_2 \to \infty$, the expected PMC tends to

$$EP_N^{\text{Selection}} \approx \Phi\left(\frac{-\delta}{2}\left[1 + \frac{2q}{\delta^2 N} \times \frac{\delta_1^2}{\sigma^2}\right]^{-1/2}\right), \tag{8.1}$$

where δ_1^2 is the mean feature-quality value (a variance of $\Delta\mu_j$) in the case where the exact values $\Delta\boldsymbol{\mu}_j$ are used to select the q best features and $\delta^2 = q\delta_1^2$ is the squared Mahalanobis distance if the ideal feature-subset selection is performed.

In (8.1) the term $2q/(\delta^2 N)$ represents the increase in the expected PMC due to suboptimal training of the classifier (cf. (8.1)–(3.3)). The term δ_1^2/σ^2 is greater than one and measures the increase in the expected PMC due to imperfect feature selection. Raudys [115] compared numerical values of the terms $1 + 2q/(\delta^2 N)$ and δ_1^2/σ^2 and performed simulation studies to demonstrate that the contribution of the term δ_1^2/σ^2 to expression (8.1) is usually larger than the term $1 + 2q/(\delta^2 N)$. Thus, one can now understand the interesting and little-known result that for simple-structured classification rules, an inaccurate feature-subset selection process increases the expected PMC more than imperfect training of a simple classifier. During the last few years, this aspect of the model selection problem has attained considerable attention in a portion of the scientific community and is known by the phrase "there is no free lunch" (e.g., a heated debate has occurred in *Neural Computation*). One cannot select the best model using only one data set. Additional information (validation set, hypotheses about the data structure, etc.) must be employed to make the correct model selection decision.

Serdobolskij [138,139] obtained important fundamental results on the efficacy of feature-subset selection. He analyzed a pattern-class model in which the observation vector $\mathbf{x}$ is composed of independent blocks of variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_h$. That is, the pattern-class density models are of the form $f(\mathbf{x}|\theta_i) = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_h|\Pi_i) = \prod_{j=1}^{h} f_i(\mathbf{x}_j|\theta_i)$. Serdobolskij used this pattern-class density model and the discriminant function

$$g(\mathbf{x}) = \sum_{i=1}^{h} \eta(u_j) \ln \frac{f_j(\mathbf{x}_j|\hat{\theta}_{1j})}{f_j(\mathbf{x}_j|\hat{\theta}_{2j})},$$

where $f_i(\mathbf{x}_j|\hat{\theta}_{ij})$ is the ith pattern-class density of the observation vector $\mathbf{x}_j$ and $\hat{\theta}_{ij}$ is the estimator of the p_h-variate pattern-class density parameter vector corresponding to the ith class. Also, the jth block is $u_{ij} = \frac{1}{2} n\hat{B}_j$, where $\hat{B}_j = \int \ln \frac{f_j(\mathbf{x}_j|\hat{\theta}_{1j})}{f_j(\mathbf{x}_j|\hat{\theta}_{2j})} (f_j(\mathbf{x}_j|\hat{\theta}_{1j}) - f_j(\mathbf{x}_j|\hat{\theta}_{2j}))\mu(d\mathbf{x}) \geqslant 0$ is a separability measure between the two pattern classes, $\eta(u_j)$ is a weighting function of contributions of separate blocks of variables, and $\mu(d\mathbf{x})$ is an absolutely continuous probability measure. If $\eta(u_j) = 0$ or 1, we have feature selection, and for continuous $\eta(u_j)$ we have feature weighting.

Serdobolskij [138, Theorem 6 and Example 2] assumed that the individual contributions u_j are random variables with cumulative distribution function $R(u)$ and found an optimal weighting function $\eta_{\text{opt}}(u) = \frac{\sigma(u)}{\pi(u)}$ yielding a minimal expected PMC

$$EP_N = \Phi(-\sqrt{M(\eta_0)}/2),$$

where $M(\eta_0) = \frac{2h}{n} \int_0^\infty \frac{\sigma^2(u)}{\pi(u)} du$, $\sigma(u) = \int \beta^2 f_{p_h+2}^{\beta}(u)\, dR(\beta^2)$, $\pi(u) = u \int f_{p_h}^{\beta}(u)\, dR(\beta^2)$, and $f_m^{\beta}(u)$ is the noncentral chi-square density with m degrees of freedom and noncentrality parameter β. Serdobolskij proved that the contribution's density function,

$$dR(\beta^2) = \left(\frac{\gamma}{2\pi}\right)^{p_h/2} \exp\left(-\frac{\gamma\beta^2}{2}\right) \alpha\vec{\beta},$$

where $\vec{\beta} \in R^{p_h}$, $|\vec{\beta}| = \beta$, and $\gamma > 0$, is the only density function for which weighting does not reduce the classification error. Thus, to obtain the minimal error rate, we must choose $\eta_{\text{opt}}(u_j) = \text{constant}$. Accordingly, the Gaussian distribution of the individual contributions u_j is only one model where no gain in the model (feature) selection can be obtained. In a number of theorems and examples, Serdobolskij [138,139] extended these results to feature selection and error-rate estimation.

Bulygin [12] also investigated the distribution of the separability estimator $\hat{\mathbf{V}}$, composed of pairwise sample Mahalanobis distances, and proposed several optimality criteria for the determination of the discriminatory power of single features, determination of the optimal dimensionality for finite training-sample sizes, and selection of informative feature subsets. One such optimality criterion is

$$G(\mathbf{f}_k, k) = \frac{n-k-2}{n-2k-1} \hat{\mathbf{V}}(\hat{\mathbf{V}} - k\mathbf{D}_n^{-1})^{-1}(\hat{\mathbf{V}} + (n-3)\mathbf{D}_n^{-1}),$$

where $\mathbf{D}_n = [(d_{nij})]$ with $d_{nij} = n\delta_{ij} - \frac{N_i N_j}{n+k-2}$, where $n = \sum_{i=1}^{k} N_i$, δ_{ij} is Kronecker's delta and $\mathbf{f}_k$ $(k<p)$ is the particular subset of features used in the calculation of the pairwise estimated Mahalanobis distances.

Raudys [115] and Raudys and Pikelis [129] considered the general problem of model selection using sample-based conditional error-rate estimators. Expected values of the actual, $P_{\text{true}}^{\text{selection}}$, and the apparent, $P_{\text{apparent}}^{\text{selection}}$, classification errors were obtained and tabulated. The actual error, $P_{\text{true}}^{\text{selection}}$, decreases as the number of the models increases. The difference between $P_{\text{true}}^{\text{selection}}$ and $P_{\text{apparent}}^{\text{selection}}$ increases with m, where m is the number of models compared empirically. In the finite validation-set-size case, little is gained when m is chosen to be large. Thus, the studies referred to immediately above demonstrate the dubious nature of an estimated decrease in the expected PMC supposedly attained when one performs feature-subset or model selection. This idea is somewhat contradictory to the idea familiar in the western DA literature of using subset selection to obtain better performance of a classifier designed with a small or limited training-sample size. Estes [33] and Murry [95] first considered the feature-selection problem experimentally and cautioned against choosing m to be large.

9. Comments

We have attempted to briefly present some of the more important past and current results derived by former Soviet Union researchers in DA and SPR. We have emphasized the findings of these scientists in error-rate analysis because their approach to deriving error-rate approximations differs from that taken in the West. Of course, we have excluded many interesting DA and SPR topics and important results studied and derived by these investigators.

Interestingly, the work of Western DA and SPR researchers appears to be better known by their former Soviet Union counterparts than vice versa. This fact is easily understood in light of the difficulty of Western DA and SPR researchers in obtaining the pertinent research literature in the Soviet Union-based journals and books. One can only hope that with the advent of sophisticated electronic mail systems and advanced computer technology, researchers in both the East and the West will become more aware of each others' research, not only in DA and SPR, but also in all other statistical science topics.

Acknowledgments

The authors wish to thank Professor Ingram Olkin for his encouragement to write this paper. We also wish to thank an anonymous referee for his insightful criticisms and his suggestion to include a partial review of the Western discriminant analysis literature. Finally, we wish to thank John W. Seaman, Jr., and Joy L. Young for their helpful suggestions in the preparation of this manuscript.

References

[1] S.A. Aivazyan, V.M. Buchstaber, I.S. Yenukov, L.D. Meshalkin, Applied Statistics: Classification and Reduction of Dimensionality, Finansy and Statistika, Moscow, 1989.

[2] D.C. Allais, The problem of too many measurements in pattern recognition, IEEE Internat. Cen. Rec. Part 7 (1966) 124–130.

[3] S. Amari, A universal theorem on learning curves, Neural Networks 6 (1993) 161–166.

[4] S. Amari, N. Fujita, S. Shinomoto, Four types of learning curves, Neural Comput. 4 (1992) 605–618.

[5] S. Amari, N. Murata, Statistical theory of learning curves under entropy loss criterion, Neural Comput. 5 (1993) 140–153.

[6] T.W. Anderson, An asymptotic expansion of the distribution of the studentized classification statistic W, Ann. Statist. 1 (1973) 964–972.

[7] D.A. Barsov, Asymptotic distributions of statistics of discriminant analysis for ellipsoidal distributions, Proceedings of the Third International Vilnius Conference on Probability Theory and Mathematical Statistics, Vol. 1, 1981, pp. 35–36 (in Russian).

[8] D.A. Barsov, Estimates of covariance matrices and classification errors for spherically symmetrical distributions, The Second All-Union School-Seminar Algorithms and Software of Applied Multivariate Statistical Analysis, 1983, pp. 214–216 (in Russian).

[9] M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.

[10] A. Bowker, Representation of Hotelling's T2 and Anderson's classification statistics W in terms of simple statistics, in: H. Solomon (Ed.), Studies in Item Analysis and Prediction, Stanford University Press, Stanford, 1961, pp. 285–292.

[11] V.L. Brailovskij, An object recognition algorithm with many parameters and its applications, Eng. Cybernet. (USSR) 2 (1964) 22–30.

[12] V.P. Bulygin, An asymptotic analysis of a large number of classes in problems of statistical classification, Sci. Papers Statist. 45 (1983) 248–252 (in Russian).

[13] V. Cherkassky, F. Mulier, Learning from Data: Concepts, Theory, and Methods, Wiley, New York, 1998.

[14] D.M. Chibisov, An asymptotic expansion for a class of estimators containing maximum likelihood estimators, Theory Probab. Appl. 28 (1973) 303–311 (in Russian).

[15] C.K. Chow, C.N. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inform. Theory IT-14 (1968) 462–467.

[16] W.G. Cochran, C. Hopkins, Some classification problems with multivariate qualitative data, Biometrics 17 (1961) 11–31.

[17] M.C. Costanza, A.A. Afifi, Comparison of stopping rules in forward stepwise discriminant analysis, J. Amer. Statist. Assoc. 74 (1979) 777–785.

[18] M.C. Costanza, T. Ashikaga, Monte carlo study of forward stepwise discrimination based on small samples, Comput. Math. Appl. 12A (1986) 245–252.

[19] A.D. Deev, Representation of statistics of discriminant analysis and asymptotic expansions in dimensionalities comparable with sample size, Rep. Acad. Sci. USSR 195 (4) (1970) 756–762 (in Russian).

[20] A.D. Deev, Asymptotic expansions for distributions of statistics W, M, and W* in discriminant analysis, in: Yu.N. Blagoveshenskij (Ed.), Statistical Methods of Classification, Vol. 31, Moscow University Press, Moscow, 1972, pp. 6–57 (in Russian).

[21] A.D. Deev, Discriminant function designed on independent blocks of variables, Eng. Cybernet. (USSR) 12 (1974) 153–156 (in Russian).

[22] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.

[23] V. Diciunas, Generalization performance of statistical neural classifiers, Ph.D. Thesis, Vilnius University, Vilnius, Lithuania, 2002.

[24] V. Diciunas, S. Raudys, Generalization error or randomized linear empirical error classifier, Informatica 11 (2000) 381–396.

[25] P.J. DiPillo, The application of bias to discriminant analysis, Comm. Statist.—Theory Methods A 5 (1976) 843–854.
[26] P.J. DiPillo, Biased discriminant analysis: evaluation of the optimum probability of misclassification, Commun Statist.—Theory Methods A 8 (1979) 1447–1457.
[27] K. Duchinskas, Asymptotic expansions for the second moments of estimates of error rate in classification, Statist. Problems Control 93 (1990) 92–100 (in Russian).
[28] R.O. Duda, P.E. Hart, D. Stork, Pattern Classification, Wiley, New York, 2000.
[29] B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, J. Amer. Statist. Assoc. 70 (1975) 892–898.
[30] B. Efron, The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
[31] I.S. Enukov, A choice of a set of measurements with maximal discriminating power in the case of limited learning sample size, in: S.A. Aivazian (Ed.), Multivariate Statistics, Nauka, Moscow, 1973, pp. 394–397 (in Russian).
[32] I.S. Enukov, A choice of the decision rule in the case of limited sample size, Statist. Problems Control 14 (1976) 127–136 (in Russian).
[33] S.E. Estes, Measurement selection for linear discriminant used in pattern classification, Ph.D. Dissertation, Stanford University, 1965.
[34] T.B. Farver, O.J. Dunn, Stepwise variable selection in classification, Biomed. J. 21 (1979) 145–153.
[35] L.P. Fatti, D.M. Hawkins, Variable selection in heteroscedastic discriminant analysis, J. Amer. Statist. Assoc. 81 (1986) 494–500.
[36] J.M. Friedman, Regularized discriminant analysis, J. Amer. Statist. Assoc. 84 (1989) 165–175.
[37] Y. Fujikoshi, Error bounds for asymptotic approximations of the linear discriminant function when the sample sizes and dimensionality are large, J. Multivariate Anal. 73 (2000) 1–17.
[38] Y. Fujikoshi, T. Seo, Asymptotic approximations for EPMC's of the linear and the quadratic functions when the sample sizes and the dimension are large, Random Oper. Stochastic Equations 6 (1998) 269–280.
[39] K. Fukunaga, Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990.
[40] S. Ganeshanandam, W.J. Krzanowski, On selecting variables and assessing their performance in linear discriminant analysis, Austral. J. Statist. 31 (1989) 433–447.
[41] N. Ge, D.G. Simpson, Correlation and high-dimensional consistency in pattern recognition, J. Amer. Statist. Assoc. 93 (1998) 995–1006.
[42] S. Geisser, Posterior odds for multivariate normal classification, J. Roy. Statist. Soc. Ser. B 21 (1964) 69–76.
[43] N. Glick, Additive estimators for probabilities of correct classification, Pattern Recognition 10 (1978) 211–222.
[44] V. Golcov, E. Troitsky, Asymptotic expansion for a probability density function for adaptive classification statistics, Statist. Problems Control 14 (1976) 11–32 (in Russian).
[45] V. Grabauskas, Personal communication, Department of Data Analysis, Institute of Mathematics and Cybernetics, Academy of Sciences of Lithuania, 1982.
[46] D. Griskevicius, S. Raudys, On the expected probability of the classification error of the classifier for discrete variables, Statist. Problems Control 38 (1979) 95–112 (in Russian).
[47] H. Gu, H. Takahashi, How bad may learning curves be?, IEEE Trans. Pattern Anal. Mach. Intell. 20 (2000) 52–64.
[48] J.D.F. Habbema, J. Hermans, Selection of variables in discriminant analysis by F-statistic and error rate, Technometrics 19 (1977) 487–493.
[49] C.P. Han, Distribution of discriminant function in circular models, Ann. Inst. Stat. Math. 22 (1970) 117–175.
[50] D.J. Hand, Recent advances in error rate estimation, Pattern Recognition Lett. 22 (1986) 335–346.
[51] T.J. Harley, Pseudo-estimates versus pseudo-inverses for singular sample covariance matrices, Sect. 2, in Report No 5, Contract DA-36-039-SC-90742, AD427172 (Sept, 1963): also MS Thesis, Moore School of Electrical Engineering, University of Pennsylvania, 1965.

[52] D. Haussler, M. Kearns, H.S. Seung, N. Tishby, Rigorous learning curves from statistical mechanics, in: Proceedings of the Seventh Annual ACM Conference on Computer Learning Theory, 1994, pp. 76–87.
[53] J. Hertz, A. Krogh, R.G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, Reading, MA, 1991.
[54] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[55] G.F. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Inform. Theory IT-14 (1965) 55–63.
[56] A.K. Jain, B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, Handbook of Statistics, Vol. 2, North-Holland, Amsterdam, 1982, pp. 835–855.
[57] S. John, Errors in discrimination, Ann. Math. Statist. 32 (1961) 1125–1144.
[58] K. Juskevicius, Investigation of the sensitivity of a minimum distance piecewise-linear classifier to the limitations of learning sample size, Statist. Problems Control 61 (1983) 89–129 (in Russian).
[59] L. Kanal, B. Chandrasekaran, On dimensionality and sample size in statistical pattern classification, Pattern Recognition 3 (1971) 238–255.
[60] D.G. Keehn, A note on learning for Gaussian properties, IEEE Trans. Inform. Theory 11 (1965) 126–131.
[61] Yu.S. Kharin, On the robustness of decision rules in discriminant analysis, Proceedings of the Second All-Union Scientific-Technical Conference on Applications of Multivariate Statistical Analysis in Economics and Estimation of Quality Production, 1981, pp. 270–272 (in Russian).
[62] Yu.S. Kharin, The risk expansion for the nonparametric classifier, Adaption in the Systems of Control and Decision Making, Nauka, Novosibirsk, 1982 (in Russian).
[63] Yu.S. Kharin, Asymptotic expansions for the risk of parametric and nonparametric decision functions, Transactions of the Ninth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Publishing House of the Czechoslovak Academy of Sciences, Prague, Czechoslovakia, 1983a, pp. 11–16.
[64] Yu.S. Kharin, Investigation and optimization of Rosenblatt-Parzen classifier with the aid of asymptotic expansion, Automat. Remote Control 11 (1983b) 91–99 (in Russian).
[65] Yu.S. Kharin, On the stability of classification decision rules in the presence of training sample mislabeling, Automat. Remote Control 11 (1983c) 100–110 (in Russian).
[66] Yu.S. Kharin, The investigation of risk for statistical classifiers using minimum estimators, Theory Problems Appl. 28 (1984) 623–630 (in Russian).
[67] Yu.S. Kharin, Robustness in Statistical Pattern Recognition, Kluwer Academic Publishers, Dordrecht, 1996.
[68] Yu.S. Kharin, K. Duchinskas, The asymptotic expansion of the risk for the classifier designed by the use maximum likelihood estimates, Statist. Problems Control 38 (1979) 77–94 (in Russian).
[69] J. Kittler, F. Roli, Multiple Classifier Systems (2001, eds), Springer Lecture Notes in Computer Science, Vols. 1857, 2096 (2000), Springer, London.
[70] J.S. Koford, G.F. Groner, The use of an adaptive threshold element to design a linear optimal pattern classifier, IEEE Trans. Inform. Theory IT-12 (1966) 42–50.
[71] J.P. Koolaard, C.R.O. Lawoko, The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study, Commun. Statist.—Theory Methods A 25 (1996) 2989–3011.
[72] V.A. Kovalevskij, A statistical approach to the problem of learning in pattern recognition, Pattern Recognition Constr. Reading Automata 1 (1966) 13–14 (in Russian).
[73] P.A. Lachenbruch, R.M. Mickey, Estimation of error rates in discriminant analysis, Technometrics 10 (1968) 1–11.
[74] G.S. Lbov, A selection of an efficient system of statistically dependent variables, in: N.G. Zagoruiko (Ed.), Computing Systems, Vol. 19, Institute of Mathematics, Novosibirsk, 1965, pp. 21–34 (in Russian).
[75] G.S. Lbov, On representativeness of the sample size while choosing the effective measurement system, Comput. Systems 22 (1966) 39–58 (in Russian).

[76] E. Levin, N. Tishby, S.A. Solla, A statistical approach to generalization in layered neural networks, Proc. IEEE 78 (1990) 2133–2150.
[77] H. Linhart, Techniques for discriminant analysis with discrete variables, Metrika 2 (1959) 138–140.
[78] W.L. Loh, On linear discriminant analysis with adaptive ridge classification rules, J. Multivariate Anal. 53 (1995) 264–278.
[79] W.L. Loh, Linear discrimination with adaptive ridge classification rules, J. Multivariate Anal. 62 (1997) 169–180.
[80] Ya.P. Lumelskij, Unbiased consistent estimates of probabilities in the case of the multivariate normal distribution, Vestnik of Moscow State Univ. 6 (1968) 14–17 (in Russian).
[81] V.R. Marco, D.M. Young, D.W. Turner, The Euclidean distance classifier: an alternative to the linear discriminant function, Commun Statist.—Comput. Simulations 6 (1987) 485–505.
[82] V.L. Matrosov, Optimal algebraic algorithms based on calculation of estimates, Rep. Acad. Sci. USSR 262 (4) (1982) 818–822 (in Russian).
[83] R.J. McKay, N.A. Campbell, Variable selection techniques in discriminant analysis, I. Description, British J. Math. Statist. Psychol. 35 (1982) 1–29.
[84] R.J. McKay, N.A. Campbell, Variable selection techniques in discriminant analysis, II. Allocation, British J. Math. Statist. Psychol. 35 (1982) 30–41.
[85] G.J. McLachlan, An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function, Austral. J. Statist. 14 (1972) 68–72.
[86] G.J. McLachlan, An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis, Austral. J. Statist. 15 (1973) 210–214.
[87] G.J. McLachlan, The asymptotic distributions of the conditional error rate and risk in discriminant analysis, Biometrika 61 (1974) 131–135.
[88] G.J. McLachlan, On the relationship between the F test and the overall error rate for variable selection in two-group discriminant analysis, Biometrics 36 (1980) 501–510.
[89] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.
[90] R. Meir, Empirical risk minimization versus maximum-likelihood estimation: a case study, Neural Comput. 1 (1995) 144–157.
[91] L.D. Meshalkin, Assignment of numerical values to nominal variables, Statist. Problems Control 14 (1976) 49–56 (in Russian).
[92] L.D. Meshalkin, Theory of statistical analysis of a chronic progressive disease, USSR Doctoral Dissertation, Moscow State University, Moscow, 1997 (in Russian).
[93] L.D. Meshalkin, V.I. Serdobolskij, Errors in classifying multivariate observations, Theory Probab. Appl. 23 (1978) 772–781 (in Russian).
[94] D. Morgera, D.B. Cooper, Structurized estimation: sample size reduction for adaptive pattern classification, IEEE Trans. Inform. Theory 23 (1977) 728–741.
[95] G.D. Murray, A caution note on selection of variables in discriminant analysis, Appl. Statist. 26 (1977) 246–250.
[96] M. Okamoto, An asymptotic expansion for the distribution of linear discriminant function, Ann. Math. Statist. 34 (1963) 1286–1301 (Correction Ann. Math. Statist. 39 (1968) 1358–1359).
[97] T.J. O'Neil, A general distribution of the error rate of a classification procedure with application to logistic regression discrimination, J. Amer. Statist. Assoc. 75 (1980) 154–160.
[98] M. Opper, D. Haussler, Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise, Proceedings of the Fourth Annual ACM Conference on Computer Learning Theory, 1991, pp. 75–87.
[99] V. Pikelis, The errors of a linear classifier with independent measurements when the learning sample size is small, Statist. Problems Control 5 (1973) 69–101 (in Russian).
[100] V. Pikelis, Comparison of methods of computing the expected classification errors, Automat. Remote Control 5 (1976) 59–63 (in Russian).
[101] I.Sh. Pinsker, Estimation of learning method and learning sample, Simulation Automat. Anal. Electrocardiograms (1973) 13–23 (in Russian).

[102] V. Pivoriunas, S. Raudys, On the accuracy of the leave-one-out estimator, Statist. Problems Control 27 (1978) 53–70 (in Russian).
[103] V.S. Pugachev, Statistical problems of pattern recognition theory, Proceedings of the Third All-Union Conference on Automatic Control, Nauka, Moscow, 1967 (in Russian).
[104] S. Putinaite, On the expected and apparent probabilities of misclassification of a classifier for discrete variables, Statist. Problems Control 93 (1990) 101–111 (in Russian).
[105] R.H. Randles, J.D. Brofitt, I.S. Ramberg, R.V. Hogg, Generalized linear and quadratic discriminant functions using robust estimates, J. Amer. Statist. Assoc. 73 (1978) 564–568.
[106] C.R. Rao, On some problems arising of discrimination with multiple characters, Sankya 9 (1949) 343–365.
[107] L.A. Rastrigin, R.Ch. Erenstein, Method of Collective Recognition, Energoizdat, Moscow, 1981 (in Russian).
[108] S. Raudys, On determining training sample size of linear classifier, Comput. Systems 28 (1967) 79–87 (in Russian).
[109] S. Raudys, Analysis of the dependence of classification probability upon training sample size, VINITI (Moscow) 313 (1968) (in Russian).
[110] S. Raudys, On the problems of sample size in pattern recognition, Proceedings of the Second All-Union Conference on Statistical Methods in Control Theory, Nauka, Moscow, 1970, pp. 64–67 (in Russian).
[111] S. Raudys, On the amount of a priori information in designing the classification algorithm, Proc. Acad. Sci. USSR (1972) 168–174 (in Russian).
[112] S. Raudys, A comparison of two methods to estimate the probabilities of classification errors, Automat. Remote Control 10 (1973) 49–53 (in Russian).
[113] S. Raudys, The classification errors of a quadratic discriminant function, Statist. Problems Control 14 (1976) 33–47 (in Russian).
[114] S. Raudys, Investigation of a nonparametric classifier when the sample size is limited, Statist. Problems Control 14 (1976) 117–126 (in Russian).
[115] S. Raudys, Classification errors when features are selected, Statist. Problems Control 38 (1979) 9–26 (in Russian);
S. Raudys, Influence of sample size on the accuracy of model selection in pattern recognition, Statist. Problems Control 50 (1981) 9–30 (in Russian).
[116] S. Raudys, Statistical pattern recognition: small sample design problems, Unpublished Manuscript, Institute of Mathematics and Cybernetics, Vilnius, 1984, pp. 480.
[117] S. Raudys, The influence of sample size on classification performance: a review, Statist. Problems Control 66 (1984) 9–42 (in Russian).
[118] S. Raudys, On the accuracy of a bootstrap estimate of the classification error, Proceedings of Ninth International Joint Conference on Pattern Recognition, IEEE Press, Los Alamitos, 1988, pp. 1230–1232.
[119] S. Raudys, On the effectiveness of Parzen window classifier, Internat. J. Informatica (Institute of Mathematics and Informatics, Vilnius) 2 (1991) 434–454.
[120] S. Raudys, On shape of pattern error function, initializations, and intrinsic dimensionality in ANN classifier design, Informatica 4 (1993) 360–383.
[121] S. Raudys, On dimensionality, sample size and classification error of nonparametric linear classification algorithms, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-19 (1997) 667–671.
[122] S. Raudys, Evolution and generalization of a single neurone, Part I. SLP as seven statistical classifiers, Neural Networks 11 (1998) 283–296.
[123] S. Raudys, Statistical and Neural Classifiers: An Integrated Approach to Design, Springer, London, 2001.
[124] S. Raudys and V. Diciunas, Expected error of minimal empirical error and maximal margin classifiers, ICPR13, Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, August 25–29, Vol. 2, Track B: Pattern Recognition and Signal Analysis, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 875–879.

[125] S. Raudys, R.P.W. Duin, On expected classification error of the Fisher classifier with pseudo-inverse covariance matrix, Pattern Recognition Lett. 19 (1998) 385–392.
[126] S. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252–264.
[127] S. Raudys, V. Pikelis, Tabulation of the probability of misclassification for the linear discriminant function, Statist. Problems Control 11 (1975) 81–120 (in Russian).
[128] S. Raudys, V. Pikelis, On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1980) 242–252.
[129] S. Raudys, V. Pikelis, Collective selection of the test version of a pattern recognition system, Pattern Recognition Lett. 1 (1982) 7–13.
[130] S. Raudys, V. Pikelis, K. Juskevicius, Experimental comparison of thirteen classification algorithms, Statist. Problems Control 11 (1975) 35–80 (in Russian).
[131] S. Raudys, M. Skurikhina, Small sample properties of ridge-estimate of the covariance matrix in statistical and neural net classification. New Trends in Probability and Statistics, Vol. 3, Proceedings of the Fifth Tartu Conference on Multivariate Statistics and Matrices in Statistics, 1994, pp. 237–245.
[132] S. Raudys, M. Skurikhina, T. Cibas, P. Gallinari, Ridge estimates of the covariance matrix and regularization of artificial neural network classifier, Pattern Recognition Image Process. Moscow No. 4 (1995) 633–650.
[133] S. Raudys, V. Vaitukaitis, Methods of estimating the probability of misclassification, Statist. Problems Control 66 (1984) 43–65 (in Russian).
[134] W. Raynes, T. Greene, Covariance pooling and stabilization for classification, Comput. Statist. Data Anal. 11 (1991) 17–42.
[135] J.W. Sayre, The distribution of the actual error rates in linear discriminant analysis, J. Amer. Statist. Assoc. 75 (1980) 201–205.
[136] M.J. Schervish, Asymptotic expansions for the means and variances of error rates, Biometrika 69 (1981) 295–299.
[137] V.I. Serdobolskij, The moments of discriminant functions and classification for a large number of variables, In Statist. Problems Control 38 (1979) 27–51 (in Russian).
[138] V.I. Serdobolskij, Discriminant analysis with a large number of variables, Rep. Acad. Sci. USSR 254 (1980) 39–44 (in Russian).
[139] V.I. Serdobolskij, On minimal error probability in discriminant analysis, Rep. Acad. Sci. USSR 270 (1983) 1066–1070 (in Russian).
[140] R. Sitgreaves, Some results on the distribution of the W-classification statistics, Studies in Item Selection and Prediction, Stanford University Press, Stanford, 1961, pp. 241–261.
[141] F.W. Smith, Design of minimum-error optimal classifiers for patterns from distributions with Gaussian tails, IEEE Trans. Inform. Theory IT-17 (1971) 701–707.
[142] F.W. Smith, Small-sample optimality of design techniques for linear classifiers of Gaussian patterns, IEEE Trans. Inform. Theory IT-18 (1972) 118–126.
[143] H. Solomon, Probability and statistics in psychometric research, in: I.J. Neyman (Ed.), Proceedings of the Third Berkley Symposium on Mathematical statistics and Probability, Univ. California Press, Berkley, 1956, pp. 169–184.
[144] J. Spanier, K.B. Oldham, An Atlas of Functions, Hemisphere Publishing Corporation, New York, 1987.
[145] T. Takeshita, J. Toriwaki, Experimental study of performance of pattern classifiers and the size of design samples, Pattern Recognition Lett. 16 (1995) 307–312.
[146] K. Tcheponis, D. Zhvirenaite, B.S. Busygin, L. Miroshnichenko, in: S. Raudys (Ed.), Methods, Criteria and Algorithms Used for Feature Transformation, Extraction and Selection in Data Analysis, Institute of Mathematics and Informatics Press, Vilnius, 1988.
[147] G.T. Toussaint, Bibliography on estimation of misclassification, IEEE Trans. Inform. Theory IT-20 (1974) 472–479.
[148] V.Yu. Urbakh, Linear discriminant analysis: loss of discriminating power when a variate is limited, Biometrics 27 (1971) 531–534.

[149] J.W. Van Ness, Dimensionality and classification performance with independent coordinates, IEEE Trans. System Cybernet. SC-7 (1977) 560–564.
[150] J.W. Van Ness, On the effects of dimension in discriminant analysis for unequal covariance populations, Technometrics 21 (1979) 119–127.
[151] J.W. Van Ness, C. Simpson, On the effects of dimension in discriminant analysis, Technometrics 18 (1976) 175–187.
[152] V.N. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer, Berlin, 1982.
[153] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, Berlin, 1995.
[154] V.N. Vapnik, D.Ya. Chervonenkis, Algorithms with full memory and recurrence algorithms in the problem to train pattern recognition, Automat. Remote Control 4 (1968) 95–106 (in Russian).
[155] V.N. Vapnik, D.Ya. Chervonenkis, Theory of Pattern Recognition—Statistical Learning Problems, Nauka, Moscow, 1974 (in Russian).
[156] M. Vidyasagar, A Theory of Learning and Generalization, Springer, London, 1997.
[157] A.J. Viollaz, A.M. Sfer, S.M. Salvatierra, An approximation of the unconditional error rates of the sample linear discriminant function, Comm. Statist.—Theory Methods A 24 (1995) 1941–1969.
[158] B. Widrow, M.E. Hoff, Adaptive switching circuits, WESCON Convent. Rec. 4 (1960) 96–104.
[159] F. Wyman, D.M. Young, D.W. Turner, A comparison of asymptotic error rate expansions for the sample linear discriminant function, Pattern Recognition 23 (1990) 775–783.
[160] V.I. Zarudskij, Classification of normal vectors with a simple structure in multidimensional space, Applied Multivariate Statistical Analysis, Nauka, Moscow, 1978, pp. 37–51 (in Russian).
[161] V.I. Zarudskij, The use of models of simple dependence problems of classification, Statist. Problems Control 38 (1979) 33–75 (in Russian).