First-Order Tree-Type Dependence between Variables and Classification Performance

Sarunas Raudys and Ausra Saudargiene

Abstract—Structuralization of the covariance matrix reduces the number of parameters to be estimated from the training data and does not affect an increase in the generalization error asymptotically as both the number of dimensions and training sample size grow. A method to benefit from approximately correct assumptions about the first order tree dependence between components of the feature vector is proposed. We use a structured estimate of the covariance matrix to decorrelate and scale the data and to train a single-layer perceptron in the transformed feature space. We show that training the perceptron can reduce negative effects of inexact a priori information. Experiments performed with 13 artificial and 10 real world data sets show that the first-order tree-type dependence model is the most preferable one out of two dozen of the covariance matrix structures investigated.

Index Terms—First-order tree-type dependence, a priori information, classification, generalization, sample size, dimensionality.

1 Introduction

More than 30 years ago, Chow and Liu [2] proposed approximation of a p-dimensional discrete probability distribution by the first-order tree-type (FOTT) dependence to reduce the number of parameters estimated from the data. They assumed that each component of the feature vector depends directly upon only one other component. In a vivid example, one could say that each variable has only one "boss" component that it depends upon. Denote p as a dimensionality of a p-variate vector $\boldsymbol{X} = (x_1, x_2, \ldots, x_p)^T$ to be classified. Then, the probability density function can be written as a product of $p-1$ second order distribution densities:

$$f(x_1, x_2, x_3, \ldots, x_p) = f(x_1) f(x_2 \mid x_{m_2}) f(x_3 \mid x_{m_3}) \ldots f(x_p \mid x_{m_p}) \quad 0 \le m_j \le p, \tag{1}$$

where sequence $\{m_2, \ldots, m_p\}$ constitutes a graph of connections (a permutation of integers $1, 2, \ldots, p$) and $f(x_1|x_0) = f(x_1)$ (we assume that variables $x_1, x_2, \ldots, x_p$ are ranked in a such way that $m_j < j$, $j = 2, 3, \ldots, p$). Chow and Liu [2] applied their model to hand-printed numeral recognition and obtained a significant improvement for a finite training set size. Prochorskas et al. [15] tried to predict outcomes of heart attacks and found that the classifier based on the FOTT dependence model outperformed other classifiers. In spite of the positive qualities of the FOTT dependence model, it has remained unnoticed in the pattern recognition literature. One of the reasons for this is that the dependency structures between the variables are more complicated in real data sets.

A goal of this paper is to find a way to utilize a priori information contained in the first-order tree-type dependence hypothesis when *this information is only partially correct*. We analyze an asymptotic behavior of the generalization properties of the Fisher LDF with a sample covariance matrix (CM), structured by the FOTT dependence model and show that the usefulness of this model depends both on the learning set size and deviation of the true data from the data model. In order to save the constructive information contained in the FOTT dependence postulation, we suggest using the structuralized covariance matrix for the whitening data transformation and training a nonlinear single-layer perceptron (SLP) in the transformed feature space. We compare the FOTT model with 20 other methods of CM structuralization and show that inaccuracies caused by incorrect postulation of the dependence structure can be reduced in SLP training.

• *The authors are with the Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania.*
E-mail: raudys@das.mii.lt, ausrsaud@takas.lt.

Manuscript received 1 Sept. 1999; revised 3 Apr. 2000; accepted 13 Sept. 2000.
Recommended for acceptance by P. Meer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110516.

2 Theoretical Background

To design the standard Fisher LDF

$$g(\boldsymbol{X}) = (\boldsymbol{X} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2))^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

we have to evaluate $2p$ components of the mean vectors and $p(p+1)/2$ elements of the covariance matrix, which is common for both pattern classes. From (1), it follows that the FOTT dependence model requires estimation of only $2p-1$ different nonzero elements in order to define the covariance matrix. Let Σ_{Tree} be the FOTT representation of the true covariance matrix, and let us use estimate $\hat{\Sigma}_{\mathrm{Tree}}$ instead of conventional sample CM. Then,

$$g_{\mathrm{Tree}}(\boldsymbol{X}) = \left(\boldsymbol{X} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^T \hat{\Sigma}^{-1}_{\mathrm{Tree}} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{2}$$

Sparse symmetric inverse matrix $\hat{\Sigma}^{-1}_{\mathrm{Tree}}$ has p nonzero elements on its diagonal and $p-1$ nonzero distinct elements outside the diagonal. The remaining elements are equal to zero. The matrix $\hat{\Sigma}^{-1}_{\mathrm{Tree}}$ can be represented as a product, i.e., $\hat{\Sigma}^{-1}_{\mathrm{Tree}} = \mathbf{C}^T\mathbf{C}$, where $\mathbf{C} = ((c_{ij}))$ and

$$c_{ij} = \begin{cases} \hat{\sigma}_{ii}(1 - \hat{\rho}^2_{im_i}))^{-\frac{1}{2}} & \text{if} \quad j = i, \\ \hat{\sigma}_{im_i im_i}(1 - \hat{\rho}^2_{im_i}))^{-\frac{1}{2}} & \text{if} \quad j = m_i, \\ 0 & \text{if} \quad j \neq i, m_i, \end{cases} \tag{3}$$

where $\hat{\rho}_{im_i} = \hat{\sigma}_{im_i}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{m_i m_i}}$ is the correlation and $\mathbf{S} = ((\hat{\sigma}_{ij}))$ [27].

To estimate the graph $\{m_2, \ldots, m_p\}$, Zarudskij [28] suggested using a stepwise algorithm developed by Kruskal [9] for the construction of trees with maximum total branch weight. Let $\{|\hat{\rho}_{12}|, |\hat{\rho}_{13}|, |\hat{\rho}_{14}|, , ..., |\hat{\rho}_{\mathrm{p-1,p}}|\}$ be the absolute values of $\hat{\rho}_{ij}$. Then, the first step selects the branch with the maximum weight $|\hat{\rho}_{\mathrm{ij}}|$, while the lth step $(2 \le l \le p-1)$ selects another maximum weighted branch $|\hat{\rho}_{ls}|$, which is different from all the branches selected during the previous steps and does not form a cycle with them.

Consider GCCM data with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and a common CM Σ. Since the vector $\boldsymbol{X}$ is Gaussian, the LDF is a Gaussian random variable with mean values $Eg_{\mathrm{Tree}}(\boldsymbol{X}) = (-1)^i \boldsymbol{\mu}^T \Sigma^{-1}_{\mathrm{Tree}} \boldsymbol{\mu}/2$ (i = 1, 2) and variance $Vg_{\mathrm{Tree}}(\boldsymbol{X}) = \boldsymbol{\mu}^T \Sigma^{-1}_{\mathrm{Tree}} \Sigma \Sigma^{-1}_{\mathrm{Tree}} \boldsymbol{\mu}$ ($\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is a difference between the mean vectors). Then, the asymptotic classification error

$$P^{\mathrm{Tree}}_{\infty} = \Phi\{-\frac{1}{2}\delta^{\mathrm{Tree}}\}, \tag{4}$$

where $\delta^{\mathrm{Tree}} = \boldsymbol{\mu}^T \Sigma^{-1}_{\mathrm{Tree}} \boldsymbol{\mu} / \sqrt{\boldsymbol{\mu}^T \Sigma^{-1}_{\mathrm{Tree}} \Sigma \Sigma^{-1}_{\mathrm{Tree}} \boldsymbol{\mu}}$ is a modified Mahalanobis distance and $\Phi\{a\}$ is a standard Gaussian cumulative distribution function.

If the assumptions about the FOTT dependence are correct, the asymptotic behavior of the classifier coincides with the optimum Bayes rule. Application of approximately correct a priori information about the dependence structure between the

feature components ($\Sigma_{\text{Tree}} \neq \Sigma$) leads to $\delta^{\text{Tree}} < \delta = \sqrt{\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu}}$ (a Mahalanobis distance) and increases the asymptotic error. Hence, it is not viable to structuralize the CM for very large training sets.

An Example. Let $p = 40$, variances be

$$\sigma_{jj} = (9(j-1)/(p-1)+1)^2,$$

correlations $\rho_{21} = 0.9$, $\rho_{31} = -0.6$, and remaining the correlations in the graph $\{m_2, \ldots, m_p\}$ are equal to zero; $\boldsymbol{\mu}_j = c\sqrt{\lambda_j/p}(p-j)/(p/2-1)$ (the first features are most informative [6]), where c is determined in order to obtain the Mahalanobis distance δ = 3.76—this corresponds to the Bayes error $P_B = 0.03$. Under the hypothesis that there is an FOTT dependence between x_1, x_2, x_3 ($m_2 = 1$, $m_3 = 1$) from (3), we can calculate $\rho_{21} = -0.54$. Let, in reality, $\rho_{23} = -0.22445$. Equation (4) shows that $\delta^{\text{Tree}} = 2.68$ and $P_\infty^{\text{Tree}} = 0.09$.

If the sample available for classifier design is finite, the estimated classification rule parameters are inexact. A double asymptotic analysis (both the dimensionality p and the training sample size $N = N_2 = N_1$ are increasing without bounds, $p/N = c$, $c < \infty$) has shown that the expected classification error (generalization error) of the standard Fisher LDF can be determined approximately by the following equation [3], [4], [17], [22], [23]:

$$EP_N^{\text{F}} \approx \Phi\left\{-\frac{\delta}{2}\frac{1}{\sqrt{T_\mu T_\Sigma}}\right\}, \tag{5}$$

where $T_\mu = 1 + 2p/(\delta^2 N)$, is a term due to estimation of the means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $T_\Sigma = 1 + p/(2N-p)$ arises due to estimation of the covariance matrix Σ.

In the case of known covariance matrix, instead of the sample estimate $\mathbf{S}$ in (5) we use Σ (exact CM) and the term T_Σ vanishes [16], [17], [3], [4]. Thus, to design the classifier, we need much fewer training samples. Equation (5) shows that if N is close to $p/2$, the standard Fisher rule does not work. One possible strategy to design the classifier is to ignore all correlations and use only diagonal elements of $\mathbf{S}$ when sample size is small. Then, we have following LDF:

$$\begin{aligned}\hat{g}_D(\boldsymbol{X}) &= \left(\boldsymbol{X} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^T \boldsymbol{D}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= \sum_{j=1}^{p}\left(x_j - \frac{1}{2}\left(\bar{x}_{1j} + \bar{x}_{2j}\right)\right)\left(\bar{x}_{1j} - \bar{x}_{1j}\right)/\hat{\sigma}_{jj},\end{aligned} \tag{6}$$

where $\boldsymbol{D}$ is a diagonal matrix composed from elements $\hat{\sigma}_{ii}$ (variances) of the matrix $\mathbf{S} = ((\hat{\sigma}_{ij}))$.

To obtain a simple and easy to comprehend expression for the generalization error of the classifier according to DF (6), assume $\boldsymbol{\Sigma} = \mathbf{I}$ (identity matrix). For a finite training set size, means $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ are Gaussian random vectors: $\bar{\mathbf{x}}_i \sim N(\boldsymbol{\mu}_i, \mathbf{I}/N)$ and elements $\hat{\sigma}_{jj}$ are distributed as $(2N-2)^{-1}\chi^2_{2N-2}$. Expected values $E(\hat{\sigma}_{jj})^{-1} = \frac{2N-2}{2N-4}$, $E(\hat{\sigma}_{jj})^{-2} = \frac{(2N-2)^2}{(2N-4)(2N-6)}$. In a high dimensional case, the distribution of the discriminant function (6) as a sum of independent random variables approaches Gaussian distribution with means and variance

$$\begin{aligned}&Eg_D(\boldsymbol{X}) \mid \boldsymbol{X} \in \omega_i) = (-1)^{i-1}\frac{1}{2}\frac{2N-2}{2N-4}\boldsymbol{\mu}^T\boldsymbol{\mu} = (-1)^{i-1}\frac{1}{2}\frac{2N-2}{2N-4}\delta^2, \\ &\qquad (i = 1, 2) \\ &Vg_D(\boldsymbol{X}) \mid \boldsymbol{X} \in \omega_i) = \\ &\left(\frac{2N-2}{2N-4}\right)^2 \sum_{j=1}^{p}\left(\frac{2N-4}{2N-6}\left(\mu_j^2 + \frac{2}{N}\right)\left(\frac{\mu_j^2}{4} + 1 + \frac{1}{2N}\right) + \frac{\mu_j^4}{4}\right).\end{aligned}$$

Therefore, the expected generalization error of the classifier according to DF (6) is

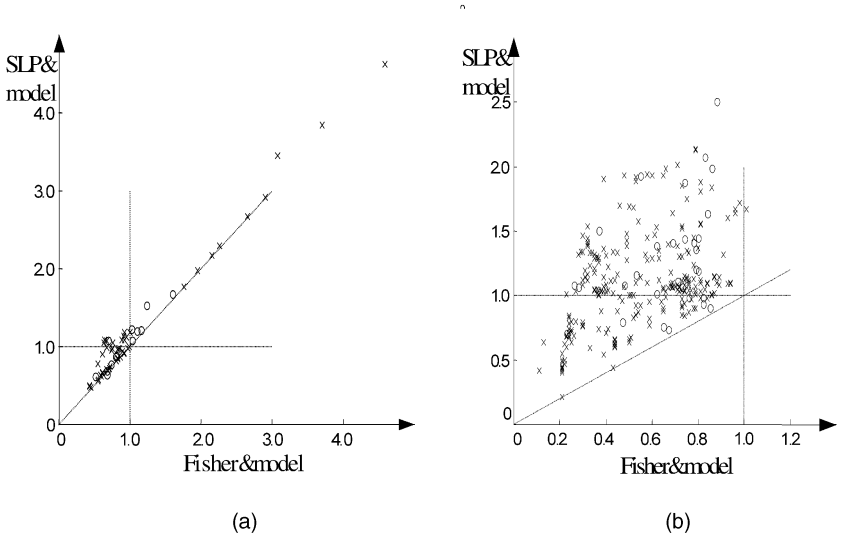
$$EP_N^D \approx \Phi\left\{-\frac{\delta}{2\sqrt{\left(1+\frac{1}{N-3}\right)\left(1+\frac{1}{N}\left(1+\frac{2p}{\delta^2}\right)+\frac{p}{\delta^2 N^2}\right)+\frac{1}{\delta^2}\sum_{j=1}^{p}\mu_j^4\frac{1}{4(N-3)}}}\right\}. \tag{7}$$

Ignoring terms containing $1/N$ and p/N^2, (7) asymptotically (as $p \to \infty$ and $N \to \infty$) becomes equivalent to (5) with $T_\Sigma = 1$. This means that *the estimation of the variances $\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp}$ does not increase the generalization error asymptotically*. In a finite dimensional case, however, the terms rejected affect the generalization error and cannot be ignored. Nevertheless, (7) advocates that the influence of the variance (common for both pattern classes) estimation is less significant than that of the means (different in both pattern classes) estimation. This important conclusion was generalized by means of double asymptotic analysis for a number of other classifier models [5], [12], [13], [26], [27]. Asymptotically, (5) with $T_\Sigma = 1$ is also valid for the FOTT linear discriminant function [27]: *the estimation of p variances and $p-1$ correlations does not increase the generalization error asymptotically*. Our analysis shows that (7) with more terms is more suitable for the FOTT dependence model than the purely asymptotic expression (5) with $T_\Sigma = 1$ in a finite dimensional case.

Calculations for the Bayes error $P_B = 0.03$ and the asymptotic error $P_\infty^{\text{Tree}} = 0.09$ using (7) with $T_\Sigma = 2N/(2N-p)$ and (7) with δ^{Tree} defined in (4) show that, in small training set size, the generalization error of the FOTT Fisher LDF is lower that standard Fisher LDF (e.g., for $N = 25$, $EP_N^F \approx 0.228$, and $EP_N^{\text{Tree}} \approx 0.137$). If $N = 36$, we have $EP_N^F \approx EP_N^{\text{Tree}} \approx 0.125$. In large learning set cases, however, the standard Fisher LFD outperforms the FOTT Fisher LDF (e.g., for $N = 100$, $EP_N^F \approx 0.052$, and $EP_N^{\text{Tree}} \approx 0.103$). We see that the structured covariance matrix is viable for application only if the training set size is rather small.

If the FOTT model is inaccurate for the real data and the training sample size is small, other methods for the CM structuralization might be useful. There exist a number of possibilities to describe the covariance matrix with a small number of parameters. One can split the features into blocks and assume the blocks to be independent. This results in a block-diagonal representation of the covariance matrix. Another possibility is to assume the variables $x_1, x_2, \ldots, x_p$ to be realizations of a stationary random process. The CM will then have the Toeplitz structure described by only p parameters. Moreover, one can assume that the process can be described by the rth order autoregression (AR) model, qth order moving average (MA), pqth order ARMA, circular models [14], [18], [24], [25], additive noise model (all diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ are equal to σ_1 and nondiagonal – to $\sigma_2, \sigma_1 > \sigma_2 > 0$), etc. In all these CM models, the number of parameters p_{model} describing the matrix Σ_{model} is substantially reduced. If the assumptions about the structure are correct, the asymptotic error corresponding to this model $P_\infty^{model} = PB$. If the assumptions are partially correct, we have $P_\infty^{model} = \Phi\{-\delta^{model}/2\} > P_B$, the Bayes error. Therefore, in certain pattern recognition problems, some models with structuralized CM can outperform the FOTT dependence model.

Consequently, instead of the dilemma of choosing between the standard Fisher classifier and its FOTT modification, the following question arises: *Which method for CM structuralization do we use in each particular situation*? The answer depends on the particular pattern recognition problem and on the training set size. In principle, one could use (5) and (7) to calculate the expected generalization error and choose an appropriate model in the GCCM case. In practice, however, three difficulties arise. The first



(a) (b)

Fig. 1. A bivariate distribution of the gain parameters $\gamma = P_N^{\mathrm{RDA}}/P_N^{\mathrm{classifier}}$ for the constrained Fisher LDF and SLP, trained optimally in the transformed feature space: (a) artificial data sets, (b) real world data sets. The circles indicate the first-order tree-type dependence model of the CM structuralization.

difficulty (the smaller one) lies in the fact that the asymptotic equations (5) and (7) are exact only in high-dimensional case. In fact, the coefficient T_μ and (5) contain terms with $1/N$ and p/N^2, which disappear as $p \to \infty$ and $N \to \infty$ [17]. Equation (5) gives only approximate values if δ^{model} is used instead of δ. More serious difficulties are: Parameters of the GCCM model to be used in (5), (7) are unknown, the GCCM pattern classes do not occur in real problems.

3 Experiments with Artificial Gaussian Data

Our goal is to obtain benchmark performance estimates of FOTT dependence model and validate simulation methodology used subsequently in the experiments with real data and the single-layer perceptron classifier. Generalization properties of the Fisher LDF with FOTT structured CM are compared with standard statistical linear methods: the Euclidean distance classifier (EDC), the standard Fisher LDF (if $2N < p$, we use a pseudoinversion of CM), and its modifications with structured sample CM estimates according to the standard Toeplitz, circular, and the first and fourth order autoregression models. As *a benchmark method* for estimation of the relative efficacy of the CM structuralization, we use an optimized standard linear regularized discriminant analysis (RDA), where, in LDF, we use a ridge estimate of the CM, $\mathbf{S} + \lambda\mathbf{I}$ [6], [11]. In our experiments, selection of the optimum λ_{opt} value is based on the lowest generalization error P_N, calculated for 50 values of $\lambda = \lambda_0/(1-\lambda_0)$, where λ_0 is in the interval $[0; 0.98]$. For the artificial data sets, we know $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$. Thus, means of any sample-based LDF g($\boldsymbol{X}$) "A" with the weights $w_0^{\mathrm{A}}, \boldsymbol{w}^{\mathrm{A}}$ will be $E\, g(\boldsymbol{X}) = \boldsymbol{\mu}_i^{\mathrm{A}} + x^T\boldsymbol{w}^{\mathrm{A}}$ $(i = 1, 2)$ and a variance $Vg(\boldsymbol{X}) = (\boldsymbol{w}^{\mathrm{A}})^T\Sigma^{-1}\boldsymbol{w}^{\mathrm{A}}$. Thus, the generalization error can be calculated analytically:

$$P_N(w_0^{\mathrm{A}}, \boldsymbol{w}^{\mathrm{A}}) = \frac{1}{2}\Phi\left\{-\frac{w_0^{\mathrm{A}} + (\boldsymbol{w}^{\mathrm{A}})^T\boldsymbol{\mu}_1}{\sqrt{(\boldsymbol{w}^{\mathrm{A}})^T\Sigma^{-1}\boldsymbol{w}^{\mathrm{A}}}}\right\} + \frac{1}{2}\Phi\left\{\frac{w_0^{\mathrm{A}} + (\boldsymbol{w}^{\mathrm{A}})^T\boldsymbol{\mu}_2}{\sqrt{(\boldsymbol{w}^{\mathrm{A}})^T\Sigma^{-1}\boldsymbol{w}^{\mathrm{A}}}}\right\} \tag{8}$$

In the experiments, we used two-category case artificial 40-variate GCCM populations with the FOTT dependence, Toeplitz, circular (for these data models we selected three sets of the model's parameters), first and fourth order AR structures (two models). In addition, we investigated two 40-variate artificial correlated GCCM data models as Friedman [6] did in analysis of RDA for the quadratic case (for details see [24], [25]). In all cases, the Mahalanobis distance δ = 3.76 (the Bayes error $P_B = 0.03$). We concentrated our analysis mainly on a case when the number of learning examples is small: N = 13, p = 40. To evaluate the effectiveness of structuralization we used a ratio $\gamma = P_N^{\mathrm{RDA}}/P_N^{\mathrm{classifier}}$, where P_N^{RDA} and $P_N^{\mathrm{classifier}}$ are the generalization errors of the optimized standard RDA and a classifier under investigation, respectively. In RDA, our benchmark method, we utilized the additional information in order to evaluate λ_{opt}. Thus, the RDA estimates were optimistically biased.

In total, we considered 13 artificial data models and 7 + 1 classifiers. All series of experiments were repeated 25 times with different randomly chosen learning sets. Results of 7 × 13 series of the experiments are presented on a horizontal axis of Fig. 1a. The experiments affirm that proper hypotheses about the data structure improve the generalization accuracy. For the three FOTT data sets, correct assumptions reduce the generalization error by 1.1, 1.6, and 1.23 times compared to the optimized RDA, while improper structuralization methods (Toeplitz, circular, autoregression) were not successful. A gain of 4.59 times on average in 25 experiments for the *AR4* data model indicates that, in principle, the proper structuralization can be very useful. For the *unstructured* Friedman data, however, all structuralization methods investigated were *ineffective*. In spite of the fact that we utilized additional information in the RDA design, correct structuralization helped to outperform the optimized RDA.

4 Whitening Data Transformation

In solving real world problems we do not know the structure of the covariance matrix in advance. Usually, our hypotheses would be inaccurate. Can we save the information contained in approximately correct assumptions and increase the generalization accuracy of the classifier? The answer is positive and, for this, we propose the use of a nonlinear single-layer perceptron. To design the SLP classifier, we do not need to consider the data structure and we still can get a good classifier for non-Gaussian data sets. Statistical hypotheses about the data may be incorporated into the SLP training process by structuring the sample CM and using it for a *whitening transformation*, which decorrelates and scales the data. The theoretical background of the integrated approach lies in a recently discovered fact that SLP evolves from a

simple statistical classifier to a more complex ones in the course of adaptive training [19].

4.1 Single-Layer Perceptron in the Transformed Feature Space

A single-layer perceptron has p inputs $x_1, \ldots, x_p$ and one output $o = f(\boldsymbol{w}^T\boldsymbol{X} + w_0)$, where $f(net)$ is a nonlinear activation function, e.g., $f(net) = \tanh(net)$. To find the weights, we have to minimize a certain loss function. The most popular one is the following sum of squares:

$$cost_l = \frac{1}{N_1 + N_2}\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left(t_j^i - f\left(\boldsymbol{w}^T\boldsymbol{x}_j^{(i)} + w_0\right)\right)^2, \tag{9}$$

where $t_j^{(i)}$ is a desired output for $\boldsymbol{x}_j^{(i)}$ (e.g., $t_j^{(1)} = 1$, $t_j^{(2)} = -1$), the jth training vector from ω_i.

The weights are updated according to standard total gradient rule $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\partial cost_l/\partial w$, where η is a learning rate. Iterative training of SLP becomes very slow when the data are almost singular, i.e., in cases of high dimensionality and a small training set [10]. The eigenvalues of the sample covariance matrix $\mathbf{S}$ are essentially different and the ratio between the largest λ_{max} and the smallest λ_{min} eigenvalues becomes large. Gradient descent training of the SLP converges when $0 < \eta < 1/\lambda_{max}$ [10]. Since the recommended optimum value of the learning rate is $\eta < 1/(2\lambda_{max})$, the convergence is extremely slow. This drawback can be improved by the data rotation and scaling. Let us perform a singular value decomposition of $\mathbf{S} : \mathbf{T}^T\mathbf{S}\mathbf{T} = \mathbf{D}$, where $\mathbf{T}$ is the $p \times p$ eigenvector matrix and $\mathbf{D}$ is the $p \times p$ diagonal eigenvalue matrix. Then, $\mathbf{S} = \mathbf{T}\mathbf{D}\mathbf{T}^T$. Consequently, the sample covariance matrix $\mathbf{S}_y$ of linearly transformed training vectors $\boldsymbol{y} = \mathbf{D}^{-1/2}\mathbf{T}^T\boldsymbol{x}$ is the identity matrix: $\mathbf{D}^{-1/2}\mathbf{T}^T\mathbf{S}\mathbf{T}\mathbf{D}^{-1/2} = \mathbf{I}$. Similarly, $\bar{\mathbf{y}}_1 = \mathbf{D}^{-1/2}\mathbf{T}^T\bar{\mathbf{x}}_1$, $\bar{\mathbf{y}}_2 = \mathbf{D}^{-1/2}\mathbf{T}^T\bar{\mathbf{x}}_2$. This is the so-called *whitening transformation* [7]. Let us train the SLP in the new space $\boldsymbol{y}$. If the center of the training data, $1/2(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$, is moved to zero, the total gradient training is applied, "symmetrical" targets $t^{(2)} = -t^{(1)}$ (for tanh activation function) are used, and weights are initialized with zero values, then, after the first iteration, we obtain a classifier equivalent to the EDC, which is the simplest statistical classifier [19]:

$$g^{(t=1)}(\boldsymbol{Y}) = (\boldsymbol{Y} - 1/2(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2))^T(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)) \times constant. \tag{10}$$

Utilization of representations $\mathbf{S}^{-1} = \mathbf{T}\mathbf{D}^{-1}\mathbf{T}^T$, $\bar{\mathbf{y}}_i = \mathbf{D}^{-1/2}\mathbf{T}^T\bar{\mathbf{x}}_i$, and $\boldsymbol{Y} = \mathbf{D}^{-1/2}\mathbf{T}^T\boldsymbol{X}$ leads to:

$$\begin{aligned} g^{t=1)}(\boldsymbol{Y}) &= (\boldsymbol{X} - 1/2(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2))^T\mathbf{T}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{T}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))k \\ &= (\boldsymbol{Y} - 1/2(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2))^T\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2))k. \end{aligned} \tag{11}$$

Equation (11) shows that SLP, trained one iteration in the transformed feature space $\boldsymbol{y}$, is equivalent to the standard Fisher LDF in the original feature space (OFS) $\boldsymbol{x}$. If we transform the data using matrix $\mathbf{G}_{\mathrm{RDA}} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^T$, $\boldsymbol{y} = \mathbf{G}_{\mathrm{RDA}}\mathbf{x}$, then, after the first iteration, we obtain a classifier which is equivalent to the standard RDA in OFS. Let us perform the data transformation using matrix $\mathbf{G}_{\mathrm{Tree}} = \mathbf{D}_{\mathrm{Tree}}^{-1/2}\mathbf{T}_{\mathrm{Tree}}^{\mathrm{T}} : \boldsymbol{y} = \mathbf{G}_{\mathrm{Tree}}\mathbf{x}$, where $\mathbf{D}_{\mathrm{Tree}}^{-1/2}$ and $\mathbf{T}_{\mathrm{Tree}}$ are the $p \times p$ diagonal eigenvalue and the $p \times p$ eigenvector matrices of the FOTT structured sample CM $\hat{\Sigma}_{\mathrm{Tree}}$. After the first iteration, the SLP performs as the EDC in the transformed feature space and as the constrained Fisher LDF in the original feature space. If the correct data transformation is applied, the generalization accuracy of the SLP increases. This is explained by the fact that the EDC, which is obtained at the very beginning of the SLP training, is the optimum classifier for spherical populations. Moreover, EDC has good properties for small learning sets, its generalization error is determined by (5) with $T_\Sigma = 1$. On the other hand, for the spherical Gaussian data, the EDC serves as a very good initialization of SLP. It was shown that a successful initialization can be very advantageous in reducing the generalization error [21]. The necessary condition to use good initial values is to stop training optimally. Moreover, equalized eigenvalues of the covariance matrix provide favorable conditions for fast training.

During further training, the SLP produces the following discriminant function:

$$\begin{aligned} g^{(t)}(\boldsymbol{Y}) &= (\boldsymbol{Y} - 1/2(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2))^T(\lambda_1\mathbf{I} + \mathbf{S}_{\mathbf{y}Tree})^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2))k_t \\ &= (\boldsymbol{X} - 1/2(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2))^T(\lambda_1\hat{\Sigma}_{\mathrm{Tree}} + \mathbf{S})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)k_t, \end{aligned} \tag{12}$$

where t is a number of training iterations, k_t is a scalar constant, $\lambda_1 = \frac{2}{(t-1)\eta}\frac{N}{N-1}$, and $\mathbf{S}_{\mathbf{y}Tree} = \mathbf{D}_{\mathrm{Tree}}^{-1/2}\mathbf{T}_{\mathrm{Tree}}^T\mathbf{S}\,\mathbf{T}_{\mathrm{Tree}}\mathbf{D}_{\mathrm{Tree}}^{-1/2}$.

Equation (12) shows that, after t iterations, we have the Fisher LDF with regularized and structured sample CM. With an increase in the number of iterations, the effect of structuralization of the covariance matrix vanishes. The resulting decision rule approaches the standard Fisher LDF. Thus, ***training the SLP can diminish negative influence of inaccurate assumptions***.

In SLP training, the weights are increasing continuously and the activation function *f*(*net*) starts acting in a nonlinear region. We obtain a robust classification rule that is insensitive to outliers. Later, we approach a minimum empirical error classifier and, at the very end, even a maximum margin classifier [19]. The last three classifiers can outperform the standard parametric statistical rules if the data are non-Gaussian. From the theoretical considerations above, it follows that data transformation, together with an optimally stopped SLP, may lead to a significant improvement of the generalization properties compared to the RDA and constrained Fisher LDF when the data are non-Gaussian, the covariance matrices of the populations are distinct, and/or the structure of CM is postulated incorrectly.

4.2 Experiments with Artificial Gaussian Data

In the experiments, we utilized the same artificial data as described in Section 3. We moved the learning data center to zero, initialized SLP with the zero weight vector, used a sigmoid activation function $f(net) = 1/(1 + \mathrm{e}^{-net})$, and trained the SLP according to the standard total gradient delta learning rule with targets 0 and 1. In order to approach the maximum margin classifier quickly, we increased the learning step η exponentially with each iteration number t [19]: $\eta = 0.2 \times 1.03^t$, where $t_{max} = 500$. The SLP was trained in the original feature space and in the transformed feature space $\boldsymbol{y} = \mathbf{D}_{model}^{-1/2}\mathbf{T}_{model}^T\boldsymbol{x}$, where $\mathbf{D}_{model}$ and $\mathbf{T}_{model}^T$ are the eigenvalue and eigenvector matrices of the structured sample CM, respectively. The optimum stopping moment t_{opt} was determined from the estimates of the generalization error $P_N(w_0, \boldsymbol{w})$ calculated analytically (8). As in the RDA, our benchmark method, additional information used to evaluate t_{opt} makes the SLP performance estimates optimistically biased. Thus, the effect of the bias becomes almost irrelevant and we obtain more or less fair comparison of both methods, the RDA and SLP.

A scatter diagram in Fig. 1 shows an efficiency of the whitening transformation and the subsequent use of the optimally stopped SLP if correct models are applied, e.g., the FOTT structuralization resulted in a gain $\gamma_{\mathrm{Tree}} = 1.23$ for the Gaussian tree-type data *Tr3*. For the integrated approach, the gain is higher: $\gamma_{Tree\&SLP} = 1.54$. The approach leads to decreased generalization, even if the assumptions about the CM structure are inexact. The Friedman data do not have any of the CM structures considered and, here, no gain was achieved. The experimental results confirm the theoretical presumption that the efficacy of the optimally stopped SLP in the original feature space is almost the same as that of the optimum RDA. Higher values of the generalization error of the SLP are typically associated with these few cases when 500 iterations are not sufficient for training (large eigenvalues! see Section 4.1).

TABLE 1
Mean Generalization Error EP_N^{RDA} of the Standard Linear RDA and the Relative Efficacy $\gamma = P_N^{\mathrm{RDA}}/P_N^{\mathrm{classifier}}$ for the EDC, the Standard and Constrained Fisher LDF and SLP, Trained Optimally in the Original and Transformed Feature Space

Classification method		EP^{RDA}	γEDC	γFisher LDF	γFisher LDF& Tree	γSLP in OFS	γSLP in TFS by **G**	γSLP in TFS by $\mathbf{G}_{Tree}$	γSLP in TFS by $\mathbf{G}_{compet}$
Data	*p/N*								
Vowels	*28/14*	0.07	0.70	0.49	0.82	1.06	0.20	0.99	**1.12**
Vowels	*28/54*	0.03	0.35	0.63	0.48	1.05	0.84	1.08	**1.12**
Sonar	*60/30*	0.19	0.68	0.65	0.76	**1.11**	0.61	0.99	1.08
Sonar	*60/80*	0.11	0.41	0.92	0.55	1.45	1.89	1.93	**1.98**
Ionosphere.	*33/16*	0.14	0.67	0.52	0.82	1.05	0.48	0.94	**1.10**
Ionosphere	*33/66*	0.10	0.57	0.90	0.79	1.31	1.23	**1.36**	1.35
Musk	*166/83*	0.14	0.47	0.53	0.62	0.97	0.52	**1.02**	1.00
Musk	*166/180*	0.08	0.25	0.98	0.37	0.74	1.44	**1.51**	0.80
Satellite image	*36/18*	0.04	0.90	0.24	0.67	1.12	0.17	0.74	**1.24**
Satellite image	*36/72*	0.04	0.81	0.85	0.73	1.87	0.94	1.07	**2.14**
Mammogram	*65/20*	0.09	0.37	0.54	0.84	1.02	0.47	**1.64**	1.23
Mammogram	*65/25*	0.07	0.26	0.51	0.69	1.19	0.43	1.41	**1.54**
Thyroid	*18/9*	0.04	0.88	0.56	0.85	1.04	0.40	0.91	**1.10**
Thyroid	*18/36*	0.03	0.81	0.42	0.74	1.57	1.14	1.44	1.84
Lung noise	*66/33*	0.23	0.68	0.75	0.80	1.07	0.64	1.19	**1.64**
Lung noise	*66/132*	0.07	0.21	1.00	0.26	0.56	1.90	1.08	1.07
Phonetic	*96/48*	0.02	0.67	0.76	0.78	1.13	0.10	1.41	1.64
Phonetic	*96/80*	0.02	0.79	0.43	0.71	1.03	0.46	1.11	1.68
Stock	*92/46*	0.22	0.44	1.12	0.80	0.62	0.68	1.45	**1.57**
Stock	*92/184*	0.10	0.21	0.95	0.83	0.44	1.41	**2.08**	1.34

CM is structuralized by the first-order tree-type dependence model and the most suitable other model.

TABLE 2
Mean Generalization Error EP_N^{RDA} of the Standard Linear RDA and the Relative Efficacy $\gamma = P_N^{\mathrm{RDA}}/P_N^{\mathrm{classifier}}$ for the Standard and Constrained Fisher LDF and SLP, Trained Optimally in the Original and Transformed Feature Space

Classification method		EP^{RDA}	γ Fisher LDF& BD	γ Fisher LDF& BD&Tree	γ SLP in TFS by $\mathbf{G}_{BD}$	γ SLP in TFS by $\mathbf{G}_{BD\&Tree}$
Data	*p/N*					
Mammogram	*65/20*	0.09	0.98	0.53	0.56	1.16
Mammogram	*65/25*	0.07	1.29	0.62	0.92	1.39
Thyroid	*18/9*	0.04	1.02	0.65	0.68	0.76
Thyroid	*18/36*	0.03	1.09	0.86	1.61	**1.99**
Lung noise	*66/33*	0.22	1.31	0.79	1.62	1.21
Lung noise	*66/132*	0.07	0.56	0.28	**2.05**	1.07
Phonetic	*96/48*	0.02	0.83	0.88	1.01	**2.51**
Phonetic	*96/80*	0.02	1.07	0.74	1.68	**1.88**
Stock	*92/46*	0.21	0.55	0.47	1.00	0.80
Stock	*92/184*	0.10	0.31	0.23	1.30	0.71

CM is structured by the block-diagonal model and block-diagonal model with blocks constrained by the first-order tree-type dependence.

5 Experiments with Real Data Sets

We have performed extensive simulation experiments with 10 real data sets in order to study the usefulness of the first-order tree-type dependence model applied in the Fisher LDF design and the SLP training. The real world pattern classification problems have not been selected purposely to fit the structuralization assumptions considered. Details concerning the data sets can be found in [20], [25]. We considered the conventional FOTT dependence model and its block-diagonal (BD) representation, where the components of the feature vector are split into separate independent blocks. All features inside one block are assumed to be FOTT dependent. The performance of the Fisher LDF with the FOTT structured sample CM is compared with the following classifiers: the EDC, the standard Fisher LDF, and 20 constrained Fisher LDF, designed by applying 20 structuralization models (general Toeplitz, circular, first and fourth order AR, first order and second order moving average, first (1, 1) order and second (2, 2) order ARMA, additive noise models in conventional and block-diagonal (BD) representations), BD and multivariate Markov models (see, e.g., [14], [18], [24], [25]).

Experiments were repeated 25 times. In the majority of the experiments, the training set constitutes a small part of a "general population." The learning set size was small compared to the dimensionality. In order to compare relative efficacy of different approaches, we adopted a point of view that *each data set represents*

a general population. This approach is frequently utilized in the statistical analysis. The optimal parameters, λ_{opt} and t_{opt}, for the RDA and SLP were found from the generalization error estimates calculated on the entire data.

Results with CM structuralization are presented in Table 1 and Table 2 (lefthand blocks of the columns) and Fig. 1b (horizontal axis). The highest average gain achieved is $\gamma_{LDA\&BD} = 1.31$ in the classification of 66-variant Lung noise data (N = 33), using the 6×11 block-diagonal representation of CM. However, the covariance matrix structuralization leads to a gain only in five series of the experiments out of 211 series performed. We emphasize that the efficiency estimates γ are pessimistic since additional information is utilized to find optimum λ for the RDA, our benchmark method. Nevertheless, the results show that the assumptions about the structure of the CM, equality of CM of different pattern classes, as well as about the normality of the distributions, have not been retained precisely.

In Section 4.1, we presented arguments that data transformations and optimally stopped SLP can be especially useful if hypotheses about the data structure are not exact. Experiments with real world problems confirm this theory. We obtain a noticeable increase in the efficacy of the CM structuralization by a joint use of the data transformation and SLP. The righthand blocks in Table 1 and Table 2 contain the gain parameter γ for the SLP, trained optimally in the original and transformed feature space. Maximum γ values are presented in bold.

In Table 1, we denoted: $\mathbf{G}, \mathbf{G}_{compet}$-transformation matrices of the conventional sample CM and the most suitable competing model, respectively, and, in Table 2, $\mathbf{G}_{BD}, \mathbf{G}_{BD\&Tree}$-transformation matrices of the block-diagonal model and the first-order tree-type dependence model in block-diagonal representation, respectively. To reveal the gain obtained from the assumptions about the FOTT dependence structure, we included into Table 1 the maximum efficacy, $\gamma_{compet\&SLP}$, achieved using the most suitable competing covariance matrix model out of 20 investigated ones.

The highest average gain, $\gamma_{BD\&Tree\&SLP} = 2.51$ (the mean in 25 experiments), is obtained in classifying the 96-variate Phonetic data (learning set size N = 32) and using the 4×24 block-diagonal representation of CM with the FOTT structuralization. In general, the first-order tree-type dependence model improves the classification accuracy of SLP in 14 cases out of 20 investigated; in eight cases, it is found to be the best model among two dozen of the CM structure models considered. The effectiveness of the structuralization is illustrated by a bivariate distribution of the gain parameters $\gamma = P_N^{\mathrm{RDA}} / P_N^{\mathrm{classifier}}$, which are evaluated in the experiments with 10 different data sets, two learning-set sizes for each type of the data, and 22 matrix structuralization models (Fig. 1b). In this figure, $\gamma_{model\&Fisher}$ corresponds to the efficacy of the structuralization in the Fisher LDF design and $\gamma_{model\&SLP}$ shows the efficacy of the methodology applied in the SLP training procedure. The major part of the points ($\gamma_{model\&Fisher}$, $\gamma_{model\&SLP}$) is situated in the lefthand upper part of the graph. It means, that the additional use of SLP helps to "ameliorate" the inexact hypotheses about the data structure and improves generalization accuracy almost in all experiments. The simulation studies indicate that the FOTT dependence model may be adequate for many types of real data. The use of the FOTT transformations and the SLP classifier introduces significant changes in the perception of this CM approximation method, whose usefulness has so far been greatly undermined.

6 Concluding Remarks

In most real problems, the assumptions about the FOTT dependence structure, CM common for both populations, and the Gaussian distributions are violated. Inexact assumptions decrease the generalization accuracy. We suggest integrating the positive qualities of the *statistical and neural* methods and rehabilitating the FOTT dependence model. Instead of using statistical methods and multivariate models of the CM (the FOTT dependence and others) to design the statistical classifier directly, we suggest application of these methods to perform the data whitening and then to train the nonlinear SLP in the transformed feature space. This modus operandi helps retain the information contained in approximately correct assumptions about the structure of the covariance matrix and improves the generalization properties. The following three main keys are the basis of success:

1. The good initial weighs of SLP result in decrease of the generalization error;
2. During adaptive training, SLP performs as seven statistical classifiers of increased complexity; the last ones are appropriate to classify non-Gaussian data;
3. The use of correct or even approximately correct statistical models for data whitening results in good initial weights of SLP and reduces the size of the required training set.

A necessary condition to save a priori information is to stop the SLP training in time. The first hints that one should decorrelate the inputs were from visual cortex investigators in biological systems [1]. In order to improve convergence properties of the gradient descent algorithm, Halkaaer and Winter [8] suggested transforming the input data by removing the mean and decreasing correlations across the input variables. In our approach, we decorrelate and scale the data in order to reduce the generalization error. Among unsolved problems, we will mention a best split of the design set into training and the validation sets, the accuracy problems, a special generation of the validation set for selection the best structuralization model, the optimal parameters, λ_{opt} and t_{opt}, for the RDA and SLP.

Acknowledgments

The authors are grateful to Dr. Algimantas Rudzionis from Kaunas University of Technology, Professor Bulent Sankur from Bogazici University, Istanbul, Professor Mineichi Kudo from Hokaydo University, Professor Jack Sklansky from the University of California, Irvine, and Professor Allen Long from South Bank University, London, for providing real world data sets.

References

[1] J.J. Atick and A.N. Redlich, "Towards a Theory of Early Visual Processing," *Neural Computation*, vol. 2, pp. 308-320, 1990.

[2] C.K. Chow and C.N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Trans. Information Theory*, vol. 14, pp. 462-467, 1968.

[3] A.D. Deev, "Representation of Statistics of Discriminant Analysis and Asymptotic Expansions in Dimensionalities Comparable with Sample Size," *Reports of Academy of Sciences of the USSR*, vol. 195, no. 4, pp. 756-762, 1970 (in Russian).

[4] A.D. Deev, "Asymptotic Expansions for Distributions of Statistics **W**, **M**, **W*** in Discriminant Analysis," *Statistical Methods of Classification*, J.N. Blagoveshenskij, ed., vol. 31, pp. 6-57, Moscow: Moscow Univ. Press, 1972 (in Russian).

[5] A.D. Deev, "Discriminant Function Designed on Independent Blocks of Variables," *Eng. Cybernetics (Proc. Academy of Sciences of the USSR)*, no. 12, pp. 153-156, 1974 (in Russian).

[6] J.M. Friedman, "Regularized Discriminant Analysis," *J. Am. Statistical Assoc.*, vol. 84, pp. 165-175, 1989.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. New York: Academic Press, 1990.

[8] S. Halkaaer and O. Winter, "The Effect of Correlated Input Data on the Dynamics of Learning," *Advances in Neural Information Processing Systems*, M.C. Mozer, M.I. Jordan, and T. Petsche, eds., vol. 9, pp. 169-175, Cambridge, Mass.: MIT Press, A Radford Book, 1996.

[9] I.B. Kruskal Jr., "On the Shortest Spanning Subtree of a Graph and the Travelling Salesman Problem," *Proc. Am. Math. Soc.*, vol. 7, pp. 48-50, 1956.

[10] Y. le Cun, I. Kanter, and S. Solla, "Eigenvalues of Covariance Matrices: Application to Neural-Network Learning," *Physical Review Letters,* vol. 66, no. 18, pp. 2396-2399, 1991.
[11] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley, 1992.
[12] L.D. Meshalkin, "Assignment of Numerical Values to Nominal Variables," *Statistical Problems Control,* S. Raudys and L. Meshalkin, eds., vol. 14, pp. 49-56, Vilnius: Inst. of Math. and Cybernetics Press, 1976 (in Russian).
[13] L.D. Meshalkin and V.I. Serdobolskij, "Errors in Classifying Multivariate Observations," *Theory of Probabilities and Its Applications,* vol. 23, no. 4, pp. 772-781, 1978 (in Russian).
[14] D. Morgera and D.B. Cooper, "Structurized Estimation: Sample Size Reduction for Adaptive Pattern Classification," *IEEE Trans. Information Theory,* vol. 23, pp. 728-741, 1977.
[15] R. Prochorskas, V. Ziuznis, and N. Misiuniene, "Use of Different Classifiers to Predict Outcomes of Heart Attacks," *Problems of Ischemic Heart Diseases,* pp. 216-267, Vilnius, Lithuania: Mokslas Publishing House, 1976 (in Russian).
[16] S. Raudys, "On Determining Training Sample Size of Linear Classifier," *Computing Systems,* N.G. Zagoruiko ed., vol. 28, pp. 79-87, Inst. of Math. Press, Novosibirsk: Nauka, 1967 (in Russian).
[17] S. Raudys, "On the Amount of a priori Information in Designing the Classification Algorithm," *Eng. Cybernetics (Proc. Academy of Sciences of the USSR),* no. 4, pp. 168-174, 1972 (in Russian).
[18] S. Raudys, "Methods to Overcome Dimensionality Problems in Statistical Pattern Recognition: A Review," *Zavodskaya Laboratorya (Factory Lab., Interdisciplinary USSR J.),* no. 3, pp. 45 & 49-55, Moscow: Nauka, 1991 (in Russian).
[19] S. Raudys, "Evolution and Generalization of a Single Neuron. I. SLP as Seven Statistical Classifiers," *Neural Networks,* vol. 11, pp. 283-296, 1998.
[20] S. Raudys, "Scaled Rotation Regularization," *Pattern Recognition,* vol. 33, pp. 1989-1998, 2000.
[21] S. Raudys and S. Amari, "Effect of Initial Values in Simple Perception," *Proc. 1998 IEEE World Congress Computational Intelligence, IJCNN '98,* pp. 1530-1535, 1998.
[22] S. Raudys and A.K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, pp. 252-264, 1991.
[23] S. Raudys and V. Pikelis, "On Dimensionality, Sample Size, Classification Error and Complexity of the Classification Algorithm in Pattern Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 2, pp. 242-252, 1980.
[24] S. Raudys and A. Saudargiene, "Structures of the Covariance Matrices in the Classifier Design," *Proc. Joint IAPR Int'l Workshops/SSPR '98 and SPR '98,* pp. 583-592, 1998.
[25] A. Saudargiene, "Structurization of the Covariance Matrix by Process Type and Block-Diagonal Models in the Classifier Design," *Informatica,* vol. 10, no. 2, pp. 245-269, Vilnius: Inst. of Math. and Informatics Press, 1999.
[26] V.I. Serdobolskij, "The Moments of Discriminant Function and Classification for a Large Number of Variables," S. Raudys, ed., vol. 38, pp. 27-51, Vilnius: Statistical Problems of Control. Inst. of Math. and Cyb. Press, 1979 (in Russian).
[27] V.I. Zarudskij, "The Use of Models of Simple Dependence Problems in Classification," *Statistical Problems of Control,* S. Raudys, ed., vol. 38, pp. 53-75, Vilnius: Inst. of Math. and Cyb. Press, 1979 (in Russian).
[28] V.I. Zarudskij, "Determination of Some Graph Connections for Normal Vectors in Large Dimensional Case," *Algorithmic and Programic Supply of Applied Multivariate Statistical Analysis,* S.A. Aivazian, ed., pp. 189-208, Moscow: Nauka, 1980 (in Russian).