INFORMATICA, 1999, Vol. 10, No. 2, 245–269

Structurization of the Covariance Matrix by Process Type and Block-Diagonal Models in the Classifier Design

Aušra SAUDARGIENĖ
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: idauda@vdu.lt

Abstract. Structurization of the sample covariance matrix reduces the number of the parameters to be estimated and, in a case the structurization assumptions are correct, improves small sample properties of a statistical linear classifier. Structured estimates of the sample covariance matrix are used to decorellate and scale the data, and to train a single layer perceptron classifier afterwards. In most from ten real world pattern classification problems tested, the structurization methodology applied together with the data transformations and subsequent use of the optimally stopped single layer perceptron resulted in a significant gain in comparison with the best statistical linear classifier – the regularized discriminant analysis.

Key words: regularized discriminant analysis, single layer perceptron, generalization, covariance matrix, dimensionality, learning-set size.

1. Introduction

For the case of two Gaussian populations $N(\boldsymbol{\mu}_1, \Sigma_1)$, $N(\boldsymbol{\mu}_2, \Sigma_2)$ the asymptotically optimal classification rule is a quadratic discriminant function. When the covariance matrices are identical, the resulting decision algorithm is linear:

$$g^F(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ is the vector to be classified, and Σ stands for the covariance matrix common for both classes.

If we assume that p components $x_1, x_2, \ldots, x_p$ of the feature vector $\boldsymbol{x}$ are mutually independent and have equal variances, the simplest – Euclidean distance classifier is obtained:

$$g^E(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{2}$$

In the procedure of the statistical classifiers design, one needs to estimate the parameters associated with the class probability densities. When the data dimensionality p is

high and the number of samples N available for the classifier design is limited, these estimates are inaccurate and result in low generalization accuracy. Generalization properties of the classifier depend on the relationship between N and p : N/p for the linear classification rules and N^2/p for the quadratic classification rules (Raudys 1972; Raudys and Pikelis, 1980). Therefore a ratio learning-set/dimensionality plays an essential role in selecting and constructing the classification algorithm.

In practice, the parameters of the classifiers (1)–(2) – the covariance matrix Σ and mean vectors $\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2$ are substituted by their "plug-in" estimates $\boldsymbol{S}, \overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2$. This approach results in optimal properties of the classifier if the training-set size is very large. In small learning-set cases, the variability of the covariance matrix estimate becomes high resulting in strong dependence of the generalization accuracy on the particular learning-set. If the number of learning samples used to estimate the covariance matrix is smaller than the dimensionality, the covariance matrix becomes singular and one can not invert it.

A number of methods exists to cope with the problems arising in the case of high dimensionality in comparison with the learning-set size:

- *Dimensionality reduction.* Various feature extraction and selection techniques are used. Feature extraction algorithms perform linear and non-linear transformations of the original p-variate vectors, and feature selection methods find the most informative subsets of features (e.g., Sammon, 1969; Bryant and Guseman, 1979).
- *Pseudoinversion of the covariance matrix*. The inverse of the covariance matrix estimate is replaced by the pseudoinversion $\boldsymbol{S}^{PSEUDO} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T} = \boldsymbol{T}'\begin{pmatrix} \boldsymbol{d}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{T}$, where $\boldsymbol{T}$ is a $p \times p$ matrix of eigenvectors, and $\boldsymbol{D} = \begin{pmatrix} \boldsymbol{d} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$ is a $p \times p$ matrix of eigenvalues, obtained by a singular value decomposition of the estimated covariance matrix $\boldsymbol{S} : \boldsymbol{TST}' = \boldsymbol{D}$, and $\boldsymbol{d}$ is a diagonal matrix with $2N - 2$ non-zero eigenvalues in its diagonal (Fukunaga, 1972).
- *Regularization of the sample covariance matrix*. Covariance matrix estimate $\boldsymbol{S}$ is substituted by a ridge estimate $\boldsymbol{S}^{RDA} = \boldsymbol{S} + \lambda\boldsymbol{I}$ in order to gain stability, where $\boldsymbol{I}$ is $p \times p$ identity matrix and λ is a positive regularization constant. The ratio of the largest and smallest eigenvalues of $\boldsymbol{S}$ is reduced by this technique. The resulting decision rule is called a regularized discriminant analysis (RDA). It is a very good classification algorithm (Friedman, 1989).
- *Structurization of the sample covariance matrix*. This method considerably reduces the number of parameters estimated from the learning-set, and can improve the small sample properties of the classifier if correct hypothesis on the structure of the covariance matrix are made. The covariance matrix is transformed into a certain form that is described by a smaller number of distinct elements. We will call them the parameters of the structured covariance matrix. Various models of the structurization exist: Diagonal, Toeplitz, Circular, autoregression, moving average, autoregression-moving average, Block-diagonal (Morgera and Cooper, 1977; Raudys, 1991; Raudys and Saudargienė, 1998).

Theoretical studies (Raudys, 1972; Deev, 1974; Meshalkin and Serdobolskij, 1978) showed that parameters of the probability density function asymptotically do not effect the generalization error, if these parameters are common for both classes and their number increases linearly with the increase in the dimensionality. Asymptotic analysis was performed with an approach where the dimensionality and the training-set size tend to infinity, i.e., $p \to \infty,\ N \to \infty$, and $N/p = \beta$ (β is a positive constant). It means if the number of the parameters describing the structured covariance matrix is proportional to the dimensionality, the structurization of the covariance matrix may result in the significant improvement of the generalization properties of the classifier in the high dimensional and large learning-set case. Therefore if correct hypothesis on the data structure are used to structurize the sample covariance matrix, the generalization accuracy of the classifier increases. Moreover, the structured sample covariance matrix can be applied to transform the data into spherical populations for a single layer perceptron training procedure and can reduce the generalization error even in a case when assumptions on the data nature differ from reality.

There exists a great variety of the models of the covariance matrix. A statistical verification of the model correspondence to the data is a procedure requiring a lot of computer time. Moreover, the goal when solving the classification problems is to minimize the probability of misclassification, but not the criteria of the model compliance. Therefore the selection of the model, resulting the lowest generalization error, can be performed experimentally.

The influence of the structurization of the covariance matrix on the generalization error of the linear discriminant analysis (LDA) and the optimally stopped single layer perceptron (SLP) was investigated in our previous work (Raudys and Saudargienė, 1998). Simulation experiments were performed in two-category case with artificial multivariate Gaussian populations sharing the common covariance matrix. In addition, several real-world data sets were used. Sample covariance matrix was structured by Toeplitz, Circular, first order tree dependence, autoregression, Block-diagonal models and substituted into LDA as well as used to decorellate and scale the data for the subsequent SLP training.

In this paper we continue to analyze the efficacy of the structurization technique applied in LDA design and SLP training process. The following models of the structurization of the covariance matrix are analyzed: Toeplitz, Circular, autoregression, Block-diagonal and the new ones are introduced: moving average, autoregression-moving average, Block-diagonal with structured blocks, Block-diagonal Markov. A usefulness of analytical expressions of the asymptotic probability of misclassification of Fisher LDA is considered, and difficulties arising in estimating the expected probability of misclassification are discussed. The proposed classifier design methodology is applied to ten multidimensional real world data sets.

Notations and abbreviations

p – dimensionality	SLP – single layer perceptron
N – number of learning vectors in one class	OFS – original feature space
CM – covariance matrix	TFS – transformed feature space
LDA – linear discriminant analysis	AR – autoregression model
RDA – regularized discriminant analysis	MA – moving average model
PMC – probability of misclassification	BD – Block-diagonal model
ARMA – autoregression moving average model	EDC – Euclidean distance classifier

2. Models of the Covariance Matrix

An objective of structurization of the covariance matrix is to reduce the number of the parameters to be estimated from the learning-set. In this paper we concentrate our investigations on the models of the covariance matrix that describe a stationary random process or posses the Block-diagonal structure.

2.1. *Covariance Matrices for Stationary Random Processes*

Let $\{\boldsymbol{X}_t,\ t = 0, \pm 1, \pm 2, \ldots\}$ be a discrete-time random signal process. If the signal has a Gaussian amplitude distribution, it is completely described by its mean value:

$$\overline{x} = E\{\boldsymbol{X}_t\} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} x_t, \tag{3}$$

and by its autocovariance function:

$$\gamma_k = E\big\{[x_t - \overline{x}][x_{t+k} - \overline{x}]\big\} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} x_t x_{t+k} - \overline{x}^2. \tag{4}$$

Autocovariance with the delay k γ_k is a covariance between the components x_t and x_{t+k} separated by k intervals of time. The variance of the process is:

$$\sigma_x^2 = \gamma_0 = E\big\{[(x_t - \overline{x})]^2\big\} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} x_t^2 - \overline{x}^2. \tag{5}$$

A random signal is stationary in the narrow sense if the probability densities and the joint probability densities are independent of a time shift. A random signal is called stationary in the wide sense if $\overline{x}$ and γ_k are independent of time (Isermann, 1981).

It is assumed an observation $x = (x_1, x_2, \ldots, x_p)'$ is taken from a stationary random sequence, i.e., the components of the p-variate feature vector x are the measurements of a stationary process at a time $t = 1, 2, \ldots, p$. The relationship between the components is described by the covariance matrix.

2.1.1. *Toeplitz Covariance Matrix of a General Form*

Toeplitz covariance matrix describes a stationary random process and has the form (Box, Jenkins, 1974):

$$\Sigma = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{pmatrix}, \tag{6}$$

where $\gamma_0 = \sigma_x^2$ is the variance of the process, and $\gamma_1, \gamma_2, \ldots, \gamma_{p-1}$ determine the dependencies between the components. Only p parameters have to be estimated from the learning-set.

2.1.2. *Circular Covariance Matrix*

The covariance matrix of a Circular structure represents a stationary periodical process and has the following form (Han, 1970):

$$\Sigma = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_2 & \gamma_1 \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_3 & \gamma_2 \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_4 & \gamma_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma 2 & \gamma_3 & \cdots & \gamma_1 & \gamma_0 \end{pmatrix}. \tag{7}$$

For p-dimensional case only $p/2$ parameters describe the covariance matrix.

The matrix can be transformed into a canonical form using $p \times p$ orthonormal transformation matrix $\boldsymbol{T} = \{t_{ij}\}$ that does not depend on Σ (Han, 1970):

$$\boldsymbol{T}\Sigma\boldsymbol{T}' = \begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \alpha_p \end{pmatrix}. \tag{8}$$

The ij^{th} element of the orthonormal matrix $\boldsymbol{T}$ is:

$$t_{ij} = \sqrt{p}\left(\cos\frac{2\pi}{p}(i-1)(j-1) + \sin\frac{2\pi}{p}(i-1)(j-1)\right). \tag{9}$$

2.1.3. *Autoregression (AR) Model*
Random process is described by difference r^{th} order autoregression equation AR(r):

$$\sum_{k=0}^{r}\left[a_k(x_{t-k}-\mu)\right] = e_t, \quad a_0 = 1, \tag{10}$$

where random variables $e_t, e_{t-1}\ldots$ are mutually independent and identically distributed $N(0,1)$.

The parameters $a_1, a_2, \ldots, a_r$ fulfil the conditions of a stationary process: the roots $z_1, z_2, \ldots, z_r$ of the characteristic polynomial $\sum\limits_{k=0}^{r} a_k z^{r-k}$ of Formula (10) lie inside the unit circle of the complex plane: $|z_k| < 1,\ k = 1, 2, \ldots, r$. An inverse of the covariance matrix is of a special form: all elements, except the ones in the main diagonal and $2r$ subdiagonals, are zeros. It is important to note, this matrix can be inverted analytically (Kligienė, 1977) employing the autoregression parameters $a_1, a_2, \ldots, a_r$:

$$\Sigma^{-1} = \begin{pmatrix}
k_{11} & \ldots & k_{1r} & k_r & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
k_{r1} & \ldots & k_{rr} & k_1 & k_2 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
k_r & \ldots & k_1 & k_0 & k_1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & \ldots & k_2 & k_1 & k_0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & \ldots & 0 & 0 & 0 & \ldots & k_0 & k_1 & k_2 & \ldots & 0 \\
0 & \ldots & 0 & 0 & 0 & \ldots & k_1 & k_0 & k_1 & \ldots & k_r \\
0 & \ldots & 0 & 0 & 0 & \ldots & k_2 & k_1 & k_{rr} & \ldots & k_{r1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & \ldots & 0 & 0 & 0 & \ldots & 0 & k_r & k_{1r} & \ldots & k_{11}
\end{pmatrix}, \tag{11}$$

where

$$k_l = \sum_{k=0}^{r-l} a_k a_{k+1}, \quad k_{st} = \sum_{k=0}^{\min(s,t)-1} a_k a_{k+|s-t|}, \quad s,t = 1,2,\ldots,r,$$

l – element or delay, r – order of the process. While calculating the sample covariance matrix the AR parameters $a_1, a_2, \ldots, a_r$ are replaced by their estimates $\widehat{a}_1, \widehat{a}_2, \ldots, \widehat{a}_r$ obtained from the autocovariance function using Yule-Walker equations (Box, Jenkins, 1974). The Markov process is a partial case of the autoregression model with $r = 1$.

2.1.4. *Moving Average (MA) Model*

The relationship between the components of the random process is determined by the sum of the weighted variables $\nu_t, \nu_{t-1}, \ldots, \nu_{t-q}$ that are mutually independent and Gaussian distributed $N(0, \sigma_\nu^2)$:

$$\sum_{k=0}^{q} \left[b_k(\nu_{t-k} - \mu) \right] = x_t, \quad b_0 = 1, \tag{12}$$

where q is the order of the moving average process MA(q). MA process is stationary without any restrictions for the parameters $b_1, b_2, \ldots, b_q$.

The covariance matrix has Toeplitz form and is composed from the variance of the process σ_x^2 in the main diagonal, and q covariances $\gamma_1, \gamma_2, \gamma_3 \ldots, \gamma_q$ in q subdiagonals. All elements in the subdiagonals, which order is higher than q, are zeros.

The variance of the process σ_x^2 is (Box, Jenkins, 1974):

$$\sigma_x^2 = (1 + b_1^2 + b_2^2 + \ldots + b_q^2)\sigma_\nu^2, \tag{13}$$

and the covariances $\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_q$ are:

$$\gamma_k = \begin{cases} (-b_k + b_1 b_{k+1} + b_2 b_{k+2} + \ldots + b_{q-k} b_q)\sigma_\nu^2, & k = 1, 2, \ldots, q, \\ 0, & k > q. \end{cases} \tag{14}$$

To find the sample covariance matrix the parameters $b_1, b_2, \ldots, b_q$ and the variance of noise σ_ν^2 are replaced by their estimates $\widehat{b}_1, \widehat{b}_2, \ldots, \widehat{b}_q, \widehat{\sigma}_\nu^2$ that are obtained from the autocovariance function in an iterative procedure (Box, Jenkins, 1974).

2.1.5. *Autoregression – Moving Average (ARMA) Model*

The dependencies between the variables of the random process are described by the r^{th} order autoregression – q^{th} order moving average equation ARMA(r, q):

$$\sum_{k=0}^{r} \left[a_k(x_{t-k} - \mu) \right] - \sum_{k=0}^{q} \left[b_k(\nu_{t-k} - \mu) \right] = 0, \quad a_0 = 1,\ b_0 = 1, \tag{15}$$

where variables $\nu_t, \nu_{t-1}, \ldots$ are mutually independent and Gaussian distributed $N(0, \sigma_v^2)$. The process is stationary under condition the roots $z_1, z_2, \ldots, z_r$ of the characteristic polynomial $\sum\limits_{k=0}^{r} a_k z^{r-k}$ of formula (15) lie inside the unit circle of the complex plane: $|z_k| < 1,\ k = 1, 2, \ldots, r$.

The covariance matrix has Toeplitz form. The covariances are calculated analytically engaging r autoregression and q moving average parameters $a_1, a_2, \ldots, a_r, b_1, b_2, \ldots,$

b_q. The variance of the process σ_x^2 and the first m covariances $\gamma_1, \gamma_2, \ldots, \gamma_m$ are found:

$$\begin{pmatrix} \sigma_x^2 \\ \gamma_1 \\ \gamma_2 \\ \ldots \\ \gamma_m \end{pmatrix} = \boldsymbol{K}^{-1}\boldsymbol{C}, \tag{16}$$

where m is equal to r if $r \geqslant q$ and to q if $r < q$, $\boldsymbol{K}$ is the matrix of the coefficients $\boldsymbol{K} = \{k_{ij}\}$ of size $(m+1) \times (m+1)$, and $\boldsymbol{C} = (c_0, c_1, c_2, \ldots, c_m)$ is the vector of free parameters.

The matrix of the coefficients $\boldsymbol{K}$ is constructed of AR coefficients:

$$\boldsymbol{K} = \begin{pmatrix} -1 & a_1 & a_2 & a_3 & \ldots & a_{r-1} & a_r \\ a_1 & a_2 - 1 & a_3 & a_4 & \ldots & a_r & 0 \\ a_2 & a_1 + a_3 & a_4 - 1 & a_5 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{r-2} & a_{r-1} + a_{r-3} & a_r + a_{r-4} & a_{r-5} & \ldots & 0 & 0 \\ a_{r-1} & a_{r-1} + a_{r-3} & a_{r-3} & a_{r-4} & \ldots & -1 & 0 \\ a_r & a_{r-1} & a_{r-2} & a_{r-3} & \ldots & a_1 & -1 \end{pmatrix} \tag{17}$$

The $m+1$ free parameters $c_0, c_1, c_2, \ldots, c_m$ are calculated:

$$c_k = \begin{cases} -\delta_\nu^2 + \sum\limits_{i=k+1}^{q} b_i \gamma_{x\nu}(k-i), & k = 0, \\ b_k \sigma_\nu^2 + \sum\limits_{i=k+1}^{q} b_i \gamma_{x\nu}(k-i), & k = 1, \ldots, q, \\ 0, & k = q+1, \ldots, m, \end{cases} \tag{18}$$

where $\gamma_{x\nu}(-n)$ is the covariance function between the sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ and the noise vector $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_p)'$ found in an iterative procedure for each value of n:

$$\gamma_{x\nu}(-n) = \sigma_\nu^2 \Big(\sum_{i=1}^{n-1} a_i \gamma_{x\nu}(-n+i) + (a_n + b_n) \Big), \quad n = 0, 1, \ldots, q;\ a_n = 0 \quad \text{if} \quad n > r. \tag{19}$$

The remaining $p - m$ covariances $\gamma_{m+1}, \gamma_{m+2}, \ldots, \gamma_p$ are expressed in a simple form:

$$\gamma_k = \sigma_\nu^2 (a_1 \gamma_{k-1} + a_2 \gamma_{k-2} + \ldots + a_r \gamma_{k-r}), \quad k = m+1, \ldots, p. \tag{20}$$

For calculation of the sample covariance matrix the estimates of the parameters and variance of a noise $\widehat{a}_1, \widehat{a}_2, \ldots, \widehat{a}_r, \widehat{b}_1, \widehat{b}_2, \ldots, \widehat{b}_q, \widehat{\sigma}_\nu^2$ are found in an iterative procedure using the autocovariance function (Box, Jenkins, 1974).

2.2. *Block-diagonal Covariance Matrices*

2.2.1. *Block-diagonal Covariance Matrix of a General Form*
The features are grouped into blocks that are assumed to be statistically independent:

$$\Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_1 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_1 \end{pmatrix}. \tag{21}$$

The number of parameters describing the Block-diagonal covariance matrix is $k = \frac{1}{2}\sum\limits_{j=1}^{h} k_j(k_j+1)$, where h is the number of independent blocks, k_j – dimensionality of the j^{th} block.

2.2.2. *Block-diagonal Covariance Matrix with Structured Blocks*
The components of the feature vectors are divided into blocks. It is assumed that blocks are statistically independent and each block posses the process type structure, i.e., is structured by any of the models presented above.

2.2.3. *Block-diagonal Markov Model*
The dependencies between the blocks are described by the Markov model. Feature vector $\boldsymbol{x} = (x_1, x_2, \dots, x_p)'$ is considered as the sequence of h mutually dependent random vectors of size k: $\boldsymbol{x} = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \dots, \boldsymbol{x}_h^*)'$, where $\boldsymbol{x}_i^* = (x_{(i-1)\times k+1}, x_{(i-1)\times k+2}, \dots, x_{(i-1)xk+k})$. It is assumed that the covariance matrix is of the form (Kligys, 1981; 1984):

$$\Sigma = \begin{pmatrix} \Sigma_0 & -\boldsymbol{A}\Sigma_0 & \boldsymbol{A}^2\Sigma_0 & \dots & (-\boldsymbol{A})^{h-1}\Sigma_0 \\ -\boldsymbol{A}\Sigma_0 & \Sigma_0 & -\boldsymbol{A}\Sigma_0 & \dots & (-\boldsymbol{A})^{h-2}\Sigma_0 \\ \dots & \dots & \dots & \dots & \dots \\ (-\boldsymbol{A})^{h-1}\Sigma_0 & (-\boldsymbol{A})^{h-2}\Sigma_0 & -\boldsymbol{A}^{h-3}\Sigma_0 & \dots & \Sigma_0 \end{pmatrix}, \tag{22}$$

where Σ_0 is the covariance matrix of the vector $\boldsymbol{x}$ and $\boldsymbol{A}$ is the matrix of the coefficients:

$$\boldsymbol{A} = \Sigma_1\Sigma_0^{-1}, \tag{23}$$

$$\Sigma_0 = E\big\{(\boldsymbol{x}_i^* - \boldsymbol{\mu}_x)'(\boldsymbol{x}_i^* - \boldsymbol{\mu}_x)\big\}, \tag{24}$$

$$\Sigma_1 = E\big\{(\boldsymbol{x}_i^* - \boldsymbol{\mu}_x)'(\boldsymbol{x}_{i+1}^* - \boldsymbol{\mu}_x)\big\}. \tag{25}$$

Matrix $\boldsymbol{A}$ fulfils the conditions of a stationary process: the roots of the characteristic polynomial $\det(\boldsymbol{I} + \boldsymbol{A}z) = 0$ lie inside the unit circle of the complex plane.

While finding the sample covariance matrix, $\boldsymbol{A}$, Σ_0 and Σ_1 are replaced by their estimates $\widehat{\boldsymbol{A}}$, $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$:

$$\widehat{\boldsymbol{A}} = \boldsymbol{S}_1\boldsymbol{S}_0^{-1}, \tag{26}$$

$$\boldsymbol{S}_0 = \frac{1}{h}\sum_{i=1}^{h}(\boldsymbol{x}_i^* - \overline{\boldsymbol{x}}_x)'(\boldsymbol{x}_i^* - \overline{\boldsymbol{x}}_x), \tag{27}$$

$$\boldsymbol{S}_1 = \frac{1}{h-1}\sum_{i=1}^{h-1}(\boldsymbol{x}_i^* - \overline{\boldsymbol{x}}_x)'(\boldsymbol{x}_{i+1}^* - \overline{\boldsymbol{x}}_x), \tag{28}$$

$$\overline{\boldsymbol{x}}_x = \frac{1}{h}\sum_{i=1}^{h}\boldsymbol{x}_i^*. \tag{29}$$

3. Asymptotic and Expected Probabilities of Misclassification

A probability of misclassification is a main criteria to select the model of the structurization while constructing the classifier. The best choice of the model is indicated by the lowest PMC in comparison with the PMC obtained using other models or standard classification algorithms. There are four distinct PMC: Bayes, conditional, expected, asymptotic (Raudys, Pikelis, 1980). Bayes PMC P_B is the probability of misclassification of the optimal Bayes classifier, constructed having complete knowledge of the underlying probability density functions. A conditional PMC P_N depends on the characteristics of the particular learning-set used to design the classifier. The expected PMC EP_N is the expectation of the conditional PMC P_N over all learning-sets of a given size N. The asymptotic PMC is a limit $P_\infty = \lim_{N\to\infty} EP_N$.

In this section we estimate a gain that can be obtained when the structurization methodology is applied to construct the linear discriminant function. For this we must find analytically the asymptotic and expected probabilities of misclassification, the error rates that characterize the classifier. Asymptotic PMC depends on the complexity of the classifier (assumptions used to construct the classifier), the dimensionality of the learning data, the true covariance matrix. Let us have the discriminant function with already found true exact parameters – mean vectors of the populations $\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2$ and the structured sample covariance matrix Σ_{ST} instead of the true covariance matrix Σ:

$$g_{ST}^F(\boldsymbol{x}) = \Big(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\Big)'\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{30}$$

Then the $g_{ST}^F(\boldsymbol{x})$ is the linear function of random normal variable $\boldsymbol{x}$ and it has the Gaussian distribution with the mean and the variance

$$E\big\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_i\big\} = \Big(\boldsymbol{\mu}_i - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\Big)\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{31}$$

$$V\big\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_i\big\} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma_{ST}^{-1}\Sigma\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{32}$$

The asymptotic probability of misclassification is:

$$P_\infty^{F\ ST} = q_1\mathrm{Prob}\big\{g_{ST}^F(\boldsymbol{x}) < 0|\boldsymbol{x}\in\pi_1\big\} + q_2\mathrm{Prob}\big\{g_{ST}^F(\boldsymbol{x}) \geqslant 0|\boldsymbol{x}\in\pi_2\big\}$$

$$= q_1\Phi\left\{\frac{E\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_1\}}{\sqrt{V\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_1\}}}\right\} + q_2\Phi\left\{\frac{E\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_2\}}{\sqrt{V\{g_{ST}^F(\boldsymbol{x})|\boldsymbol{x}\in\pi_2\}}}\right\}, \quad (33)$$

where $\Phi\{a\} = \int_{-\infty}^{a}(2\pi)^{-1/2}\sigma^{-1}\exp\{-t^2/(2\sigma^2)\}\,dt$ is a standard Gaussian cumulative distribution function, and $q_1,\ q_2$ are a priori probabilities of the classes $\pi_1,\ \pi_2$ respectively.

For $q_1 = q_2 = 0.5$ we obtain:

$$P_\infty^{F\ ST} = \Phi\left\{-\frac{1}{2}\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma_{ST}^{-1}\Sigma\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}}\right\} = \Phi\left\{-\frac{\delta^{ST}}{2}\right\}, \quad (34)$$

where $\delta^{ST} = \dfrac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma_{ST}^{-1}\Sigma\Sigma_{ST}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}}$ is a "structured" Machalanobis distance.

Formula (34) is a tool to evaluate asymptotic PMC of Fisher LDA with the structured sample CM when the assumptions on its structure are either correct or not. This is illustrated by the simulation experiments performed with two-category case artificial Gaussian 60-variate populations with common covariance matrix and real data sets. Analytical asymptotic errors of Fisher LDA, expressed by formula (34), were compared with the experimental results. All samples of the real data sets were used to estimate the means of the classes and the covariance matrix, and then to evaluate the generalization performance of LDA. When all design set is used as a learning set and then reused as a test set, the obtained error rate estimate is called resubstitution error (Raudys, 1973). Therefore, in fact we calculated resubstitution error, although in this chapter we consider it as the asymptotic error. Sample covariance matrix was structured by the process type and Block-diagonal models. Block-diagonal model, applied for artificial data, consisted of 2 blocks of sizes 10×10 and 50×50 in the diagonal; for the real Lung noise data it was represented by 6 blocks of size 11×11 in the diagonal.The results are presented in Table 1. In upper rows the analytical values of asymptotic PMC are given, and in lower rows the experimental results are listed.

Asymptotic PMC values, obtained analytically, correspond to the experimental results for artificial Gaussian data. It shows that formula (34) is appropriate while applied for Gaussian populations with common covariance matrix. For the real world data, however, it may lead to biased-sometimes too optimistically – estimates. As an example is the Lung noise data set: the asymptotic PMC, calculated analytically is $P_\infty = 0.00$ for the Block-diagonal model and shows the good separability of the classes. However, in the experiments we obtain high asymptotic PMC: $P_\infty = 0.13$.

Asymptotic behaviour of PMC is not a reliable criteria for selecting the model of the structurization while constructing the classifier in small learning-set case. It would be advantageous to estimate the expected PMC. The dependence of the generalization error on the learning-set size, dimensionality and Machalanobis distance has been obtained for

Table 1

Comparison of theoretical (above) and experimental (below) values of the resubstitution error. Dimensionality $p = 60$, Machalanobis distance δ =3.76 for artificial Gaussian populations with common covariance matrix of Circular, Toeplitz structure. Covariance matrix structured by Toeplitz, Circular, 1st order AR, 4th order AR, 1st order MA, 2nd order MA, 1st-1st order ARMA, 2nd-2nd order ARMA models, Block-diagonal models.

Model of structur./Data	Conventional S	Circular	Toeplitz	AR1	AR4
Circular	0.03	0.03	0.03	0.23	0.20
(artificial)	0.03	0.03	0.03	0.23	0.20
Toeplitz	0.03	0.05	0.03	0.03	0.03
(artificial)	0.03	0.05	0.03	0.03	0.03
Lung noise	0.00	0.01	0.01	0.01	0.01
(real world)	0.05	0.23	0.23	0.29	0.29
Sonar	0.10	0.19	0.18	0.19	0.19
(real world)	0.09	0.16	0.15	0.19	0.16

Model of structur./Data	MA1	MA2	ARMA1.1	ARMA2.2	BD
Circular	0.26	0.26	0.25	0.25	0.19
(artifical)	0.26	0.26	0.25	0.25	0.19
Toeplitz	0.03	0.03	0.03	0.03	0.03
(artifical)	0.03	0.03	0.03	0.03	0.03
Lung noise	0.01	0.01	0.01	0.01	0.00
(real world)	0.31	0.30	0.30	0.30	0.13
Sonar	0.24	0.22	0.20	0.21	–
(real world)	0.23	0.22	0.20	0.20	–

the Fisher LDA (Raudys, 1972; Deev, 1970; 1972) with equal number of learning patterns from both classes $N_1 = N_2 = N$ and assuming the populations are Gaussian with the common covariance matrix:

$$EP_N^F \approx E\left(-\frac{\delta}{2}\frac{1}{\sqrt{\left(1+\frac{2p}{N\delta^2}\right)\frac{2N}{2N-p}}}\right), \tag{35}$$

where $\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ is a M achalanobis distance. The term $1 + \dfrac{2p}{N\delta^2}$ reflects the inexact estimation of the mean vectors, and the term $\dfrac{2N}{2N-p}$ arises due to inexact estimation of the covariance matrix.

When we know that $\Sigma = \boldsymbol{I}\sigma^2$ we can use the Euclidean distance classifier (2) instead

of Fisher LDA (1) and have (Raudys, 1967):

$$EP_N^E \approx \Phi\left(-\frac{\delta}{2}\frac{1}{\sqrt{\left(1+\frac{2p}{N\delta^2}\right)}}\right). \tag{36}$$

For more general case where we apply the Euclidean distance classifier for Gaussian populations with $\Sigma \neq \boldsymbol{I}\sigma^2$,

$$EP_N^E \approx \Phi\left(-\frac{\delta^*}{2}\frac{1}{\sqrt{\left(1+\frac{2p^*}{N(\delta^*)^2}\right)}}\right), \tag{37}$$

where δ^* is a modified Euclidean distance $\delta^* = \dfrac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}}$ and p^* is a modified (intrinsic) dimensionality $p^* = \dfrac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^2(tr\Sigma^2)}{((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))^2}$. The modified dimensionality p^* may be larger or smaller than the actual dimensionality p. In extreme cases p^* can be equal to 1 or be very large. Therefore, the sensitivity of EDC to the learning-set size when $\Sigma \neq \boldsymbol{I}\sigma^2$ can be greater or smaller than the sensitivity of the EDC when populations are spherical: $\Sigma = \boldsymbol{I}\sigma^2$.

In an analogy with expression (37) for the structurization of CM technique, we approximate the expected error rate by the following expression:

$$EP_N^{F\ ST} \approx \Phi\left(-\frac{\delta^{ST}}{2}\frac{1}{\sqrt{\left(1+\frac{2p^{ST1}}{N(\delta^{ST})^2}\right)\frac{2N}{2N-p^{ST2}}}}\right), \tag{38}$$

where p^{ST1}, p^{ST2} are modified dimensionalities. We remind, for correct model of the structurization δ^{ST}, $p^{ST1}=p$, $p^{ST2}=0$ resulting in reduction of $EP_N^{F\ ST}$.

We suppose, if the model of the structurization does not correspond to the true one, the modified dimensionalities may be $1 \leqslant p^{ST1} < \infty$ and $0 \leqslant p^{ST2} < \infty$. There are no explicit expressions for p^{ST1} and p^{ST2}. These parameters, however, can be estimated by simulation. In one of experiments, the Fisher LDA was trained using N patterns from each of two 60-variate artificial Gaussian populations with the common covariance matrix of Toeplitz type. Sample CM was structured by the 4th order AR model. The mean generalization error was obtained by averaging the results over 25 experiments. The experimental values of the mean generalization error for a set of N values are presented by a curve 1 in Fig. 1. A good approximation of the dependence obtained is a curve 2, calculated analytically using formula (38) with the experimentally estimated parameters $p^{ST1}=60$ and $p^{ST2}=5$.

Simulation studies showed that for each type of the data CM structure and the model of the structurization there exist specific values of p^{ST1}, p^{ST2}. Table 2 contains the

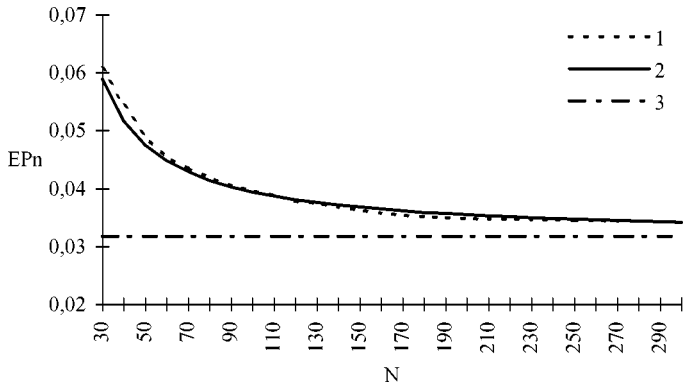

Fig. 1. The generalization error versus the learning-set size N. Simulation (1) and analytical (2) results, asymptotic error (3). Dimensionality $p = 60$, modified dimensionalities $p^{ST1} = 60$ and $p^{ST2} = 5$, Machalanobis distance $\delta = 3.76$, covariance matrix of the populations is of Toeplitz type, structurization model is the 4th order AR.

modified dimensionalities estimated by simulation for Circular, Toeplitz, Block-diagonal two-category case artificial Gaussian 60-variate data models. CM is structurized by the 4th order AR, 1st order MA and Block-diagonal models. In upper rows the values of p^{ST1} are given, and in lower rows the values of p^{ST2} are presented. When the 4th order AR structurization model is applied, a relationship between the expected PMC and the learning-set size N is expressed by (38) with $p^{ST1} = 60$ and $p^{ST2} = 15$ for the data with CM of Circular structure; $p^{ST1} = 60$ and $p^{ST2} = 5$ for the data with CM of Toeplitz structure; $p^{ST1} = 60$ and $p^{ST2} = 50$ for the data with CM of Block-diagonal structure.

We see, in the empirical (38), the modified dimensionalities p^{ST1}, p^{ST2} differ for each data type, therefore just now we can not use the formula (38) to estimate the expected PMC if the true model of CM is not known. Moreover, the parameters p^{ST1} and p^{ST2} are influenced by the learning-set size N. For example, the 4th order AR model applied for Sonar data set results $p^{ST1} = 26$ and $p^{ST2} = 0$ for $N = 31$; $p^{ST1} = 15$ and $p^{ST2} = 0$ for $N = 60$; the Block-diagonal model gives $p^{ST1} = 44$ and $p^{ST2} = 0$ for $N = 31$; $p^{ST1} = 14$ and $p^{ST2} = 0$ for $N = 60$. It means, the dependence of the expected generalization error on the true dimensionality p, learning-set size N, covariance matrix, structurization model is much more complicated.

Our analysis shows that we can not find the modified dimensionalities p^{ST1}, p^{ST2} unique for all data types and the structurization models. Therefore, for practical use we suggest to utilize the standard cross validation procedure in order to decide which model performs the best. These are our arguments:

1. Formula (38) is an approximate one. It shows main tendencies in evaluation the *mean* generalization error, however, our analysis presented above had shown that this formula is not sufficiently accurate for a wide range of values of *N, p*, and the data models. It is very difficult to obtain a simple, easy to use, analytical formula for the case when the data model differs from assumptions used to design the classification rule.

Table 2

Modified dimensionalities p^{ST1} (above) and p^{ST2} (below) for dimensionality $p = 60$, Machalanobis distance $\delta = 3.76$ for Circular, Toeplitz, Block-diagonal data models. Covariance matrix structured by the 4th order AR, 1st order MA, Block-diagonal models

Model of structurization/Data	AR4	MA1	BD
Circular	60	1	60
	15	0	50
Toeplitz	60	5	60
	5	0	50
Block-diagonal	60	1	60
(4 blocks $\times$ 15 features)	50	0	40

2. The real world data is not Gaussian with a common covariance matrix.
3. The generalization error depends on the data parameters (p^{ST1}, p^{ST2}). These parameters, however, are unknown.
4. Analytical formulae such as (35) and (38) allow to calculate *the mean values*. In real situations in the model selection, however, we need to know *the conditional classification error*. This error is a function of a random learning-set.
5. A main innovation of our approach discussed in the previous paper (Raudys and Saudargienė, 1998) and this one, is to use an optimally stopped single layer perceptron after the data transformation performed on a basis of the structurized covariance matrix estimate. In this technique, we do not need to choose a proper covariance matrix model.

Therefore, a key point of our approach is: for each concrete real world pattern recognition problem to test a number of different structures of the covariance matrix, and then to use the cross validation technique to choose the best model.

4. Simulation Experiments

4.1. *A Goal and Methodology*

The goal is to study the effect of various types of the structurization of the sample covariance matrix on the generalization performance of the linear statistical classifier and the single layer perceptron. In our experiments we compared generalization error of statistical LDA and SLP, denoted by $P_N^{classiifer}$, with the generalization error of optimized RDA P_N^{RDA}, our bench-mark method, and calculated mean relative efficacies of the structurization of the covariance matrix $\nu = \dfrac{P_N^{RDA}}{P_N^{classiifer}}$. Simulation experiments were performed with ten real world data sets. We did not selected the problems purposefully to fit the structurization assumptions considered in this paper.

The methodology of the simulation experiments is similar to that described in our previous work (Raudys and Saudargienė, 1998). We will shortly remind the main theoretical aspects of the simulation study. In a first part of our experiments, the Fisher LDA, described by formula (1), was constructed using the various structured sample estimates of the covariance matrix. For the purpose of the comparison with other methods we have included the Euclidean distance classifier, the standard Fisher LDA (if $n = N_1 + N_2 < p$, the pseudoinversion is used), and the best known linear classification method – the standard RDA with the optimal regularization parameter λ evaluated from the 50 estimates of the generalization error obtained for 50 values of the regularization parameter.

In another part of the experiments, we applied the matrix structurization methodology to improve the SLP classifier training process. Prior to training we used the structurized matrix estimates to transform the populations into spherical ones. It was shown (Raudys, 1998), that SLP may become the asymptotically optimal classifier after the first training iteration to discriminate two spherically Gaussian pattern classes. The following conditions must be fulfilled: the centre of the data is moved to the zero point; symmetrical targets $t_1 = 1 - t_2$ (for a sigmoid activation function) are used if the number of the training samples from both classes is equal $N_1 = N_2$ ($t_1 = -t_2 N_1/N_2$, if $N_1 \neq N_2$); initial weights are zeros; and a total gradient training rule is used to update the weights. Then after the first iteration SLP performs as the Euclidean distance classifier, and in following iterations it moves towards the classifiers of the increased complexity – RDA, and later to the standard Fisher LDA or Fisher LDA with a pseudoinversion of the covariance matrix if $n = N_1 + N_2 < p$.

The conditions to obtain a good, almost optimal classifier after the first training iteration include certain requirements for the data: the populations must be transformed into the spherical Gaussian ones. In our analysis a linear transformation $\boldsymbol{y} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{T}'\boldsymbol{x}$ is used to decorellate and scale the components of the feature vector $\boldsymbol{x}$, where $\boldsymbol{D}$ is a diagonal $p \times p$ matrix composed from eigenvalues of the sample covariance matrix $\boldsymbol{S}$, and $\boldsymbol{T}$ is $p \times p$ eigenvectors matrix of $\boldsymbol{S}$. Then the covariance matrix of vectors $\boldsymbol{y}$ is an identity matrix. The SLP, trained only one iteration, is equivalent to EDC in the transformed feature space and Fisher LDA in the original feature space. Use of the structured sample covariance matrix S_{ST} allows to incorporate the statistical hypothesis on the data structure into the perceptron training process. The feature vectors in the new space are found by $\boldsymbol{y} = \boldsymbol{D}_{ST}^{-\frac{1}{2}}\boldsymbol{T}'_{ST}\boldsymbol{x}$, where $\boldsymbol{D}_{ST}$ and $\boldsymbol{T}_{ST}$ are diagonal eigenvalue and eigenvectors matrix of the structured sample covariance matrix $\boldsymbol{S}_{ST}$. Then SLP after the first iteration performs as Fisher LDA with the structured sample CM in the original feature space. For example, if it is assumed that CM has Toeplitz form, SLP trained on the data $\boldsymbol{y} = \boldsymbol{D}_{Toeplitz}^{-\frac{1}{2}}\boldsymbol{T}'_{Toeplitz}\boldsymbol{x}$ after the first iteration gives the generalization performance similar to the Fisher LDA with the Toeplitz structured CM. Although at the very beginning of the training in the new feature space the SLP classifier corresponds to the Fisher LDA in the original feature space, further SLP training may significantly reduce the generalization error if assumptions on the data normality, common covariance matrix, its structure are not absolutely correct. The best performance could be achieved if SLP would be stopped optimally.

In our experiments, the SLP classifier was trained in the original and transformed feature space. The data rotation and scaling $\boldsymbol{y} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{T}'\boldsymbol{x}$ was performed using the conventional and the structured estimates of the covariance matrix. The centre of the data was moved to $\frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)$. SLP was initialized with zero weights and trained in a batch mode with the increasing learning rate $\eta = 0.2 * (1.03)^t$, where t-number of iterations ($t_{\max} = 500$). Generalization accuracy was estimated after each training iteration, and the training was stopped at the optimal moment determined by the lowest generalization error.

Learning-set was constructed from N samples selected randomly in each class (except the Phonetic data set; in this case N samples belonged to 10 from 20 randomly chosen speakers), and all samples of the data were used as the test set. The mean generalization error was computed by averaging the results over 25 experiments with each size of the learning-set.

The sample CM was structured by the models described in Section 2. The Block-diagonal model of a general form and Block-diagonal model with structured blocks were applied in the classification of 5 data sets for which a priori information on the size of the blocks was available, and the Block-diagonal Markov model was investigated for 3 data sets with the feature blocks of the equal size.

In this work, the determination of the order of the process type model and the problem of constructing the blocks in the Block-diagonal models have not been considered. The AR, MA, ARMA models with the fixed order parameters were applied in classifying the data. While applying the Block-diagonal models of structurization it was assumed that sizes of the blocks and features belonging to them are known a priori. Although the order of the model influences the structurization of the covariance matrix, the goal of our investigations was not to determine the type and the model's order, but to show that the new approach allows to obtain a minimal generalization error. The influence of the model order on the generalization error, as well as the methodology to build the blocks may be the subject for further research.

4.2. *Data Description*

We used following ten real world data sets:

Vowels data set. 800 vowels, 400 in one class, pronounced by 20 speakers, are characterized by 28 spectral and cepstral features.

Mammogram data set. The data set consists of 57 benign and 29 malignant mammograms, represented by 65 features. The first 18 features characterize classification (number, shape, size, etc) and the remaining 47 ones – a texture (histogram statistics, Gabor wavelet response, etc).

Sonar data set. Two classes describe sonar signals bounced off a metal cylinder and those bounced off a cylindrical rock (111 and 97 patterns, respectively). Each pattern is characterised by 60 features obtained by integrating energy within a particular frequency over a certain period of time.

Ionosphere data set. Two class 33-variate radar data is represented by 127 patterns in the first class and 226 patterns in the second one. The targets of the radars were free elec-

trons in the ionosphere. A "good" radar returns are those showing evidence of some type of structure in the ionosphere, and a "bad" one returns are those that do not; their signals pass through the ionosphere. Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number.

Musk data set. The data set describes a set of 92 molecules of which 47 are judged by human experts to be musks and the remaining 45 molecules are judged to be non-musks. The 166 features that describe these molecules depend upon the exact shape of the molecule. Because bonds can rotate, a single molecule can adopt many different shapes. The low-energy conformations were generated and then filtered to remove highly similar conformation. This left 476 conformations: 207 in the first class and 269 in the second one.

Satellite image data set. The data set consists of the multispectral values of pixels in 3×3 neighbourhood in the satellite images. The first class consists of 479 patterns, and the second one consists of 415 vectors. The feature vector contains 36 components that corresponds to 9 pixel values in four spectral bands.

Thyroid data set. The data set is represented by 18-variate vectors that describe 93 healthy subjects and 191 hypothyroid patients. The first 6 features are continuos, and the remaining 12 ones are binary.

Lung noise data set. Each of two classes – healthy and pathological – are described by 180 vectors, measured on 18 subjects. 66 spectral and cepstral features characterize 6 stages of the respiration cycle: early, middle, late inhalation and early, middle, late expiration.

Phonetic data set. The data set consists of 400 samples, 200 in each class. 96 cepstral features characterize consonants "m", "n" in a context of vowels "a-a", pronounced by 20 different speakers. Feature vector is divided into 4 blocks having 24 components each. Learning-set is constructed from the samples belonging to 11 randomly selected speakers.

Stock data set. 92-variate feature vector describes a history of 4 days-old 23 factors that influence the fluctuation of the share prices in the stock. 610 samples represent the decrease in the price, and 770 samples represent the increase.

These real world data sets were not selected purposefully to fit the CM models considered in this paper. Features of some of the data sets in no way are realizations of the stationary random process, and thus there the random process CM models can not be applied. Nevertheless, in order to trace an influence of an accuracy of incorrect assumptions about the CM structure on the classification results obtained we applied the random process models for all data sets.

Asymptotic error rates are presented in Table 3. The available samples were split into two sets of equal size: the first set was used as the training set to estimate the unknown parameters of the classifier, and the second one was used as the test set. The generalization performance was tested twice: on the training set and on the test set. Then the sets were interchanged, and the procedure was repeated. The mean value of these four errors was used as the asymptotic error. This method was suggested by Raudys (1976). For more exact calculation asymptotic error rates were obtained on 5 random splits of the data and then averaging the estimates. In SLP training process the feature vectors were

Table 3

Estimates of the asymptotic errors

No	Classif. method Data	Blocks Features	SLP in TFS by T_{conv}	LDA	LDA& Toeplitz	LDA& BD	LDA& BD& Toeplitz	LDA& BD& Markov
1.	Vowels	1×28	0.01	0.02	0.05	–	–	–
2.	Sonar	1×60	0.11	0.17	0.21	–	–	–
3.	Ionosphere	1×33	0.09	0.10	0.12	–	–	–
4.	Musk	1×166	0.06	0.12	0.29	–	–	–
5.	Satellite	1×36	0.02	0.15	0.05	–	–	–
6.	Mammogram	1×18+ 1× 47	0.01	0.10	0.22	0.11	0.32	–
7.	Thyroid	1× 6+ 1× 12	0.02	0.03	0.03	0.03	0.04	–
8.	Lung noise	6× 11	0.06	0.09	0.24	0.14	0.31	0.19
9.	Phonetic	4× 24	0.00	0.06	0.06	0.01	0.03	0.10
10.	Stock	4× 23	0.03	0.07	0.40	0.30	0.46	0.49

transformed by linear transformation $\boldsymbol{y} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{T}'\boldsymbol{x}$, where $\boldsymbol{D}$ and $\boldsymbol{T}$ are diagonal eigenvalue and eigenvectors matrix of the regularized data covariance matrix $\boldsymbol{S} = \boldsymbol{S} + \lambda\boldsymbol{I}$, λ – small positive regularization constant.

4.3. *Results*

The standard statistical classifiers – EDC and Fisher LDA always performed with the lower accuracy than RDA with optimal λ. In most cases we did not obtain the improvement in the generalization accuracy of the statistical LDA with the structured sample CM. In some cases, however, we achieved positive results. The gain was obtained in classification the Lung noise data when the model of the Block-diagonal covariance matrix, consisting of six blocks of size 11×11 in diagonal, was applied: the efficacy $\nu_{LDA\&BD} = 1.2$ for $N = 22$, $EP_N^{RDA} = 0.24$; $\nu_{LDA\&BD} = 1.31$ for $N = 33$, $EP_N^{RDA} = 0.23$; but $\nu_{LDA\&BD} = 0.52$ for $N = 132$, $EP_N^{RDA} = 0.07$. For some models and data sets the structurization of sample CM resulted in the generalization error that was higher, however comparable with the accuracy of RDA. For example, the gain reached 0.93 times using the ARMA1,1 model for the Satellite image data set, $N = 12$, $EP_N^{RDA} = 0.04$, and the efficacy was 0.89 loosing behind RDA while applying the MA1 model in classification the Ionosphere data for $N = 11$, $EP_N^{RDA} = 0.16$. A main reason of inefficacy of the structurization while applying the standard statistical methods lies in incorrect assumptions on the covariance matrix structures. The standard RDA with the optimal λ is a very good method, and it is difficult to improve it.

In the experiments with the SLP classifier, we obtained the significant increase in the efficacy of the structurization using the methodology in companion with the optimally stopped SLP. There we trained the perceptron on the data transformed according the structured estimate of the sample CM. In spite of the fact that assumptions on the covariance matrix structure, the Gaussian distribution of the data as well as on the equality of

the covariance matrices were not correct, the application of the structurization methodology and the optimally stopped SLP led to the improvement of the classification accuracy of SLP in comparison with RDA.

The results obtained applying the process type models of the structurization in SLP training process are presented in Table 4, and the results obtained while applying the Block-diagonal models are given in Table 5. The first two columns contain the data name and the training set size N, the third one contains the mean values of the generalization error of the standard RDA with the optimal regularization parameter λ. In the following columns the relative efficacy of the different classifiers (structurizations of the covariance matrix) are listed. For example, when the Block-diagonal model was applied for the Lung data, it yielded the efficacy $\nu_{SLP\ in\ TFS\ by\ T_{BD}} = 1.36, 1.62, 2.05$ for $N = 22$, 33, 132 respectively. ARMA1,1 model resulted no improvement in the generalization performance of LDA for the Satellite image data and learning-set sizes $N = 12$, 18, 36: $\nu_{LDA\&ARMA1,1} = 0.93, 0.87, 0.79$. After the data transformation and SLP use we obtained $\nu_{SLP\ in\ TFS\ by\ T_{ARMA1,1}} = 1.04, 1.15, 2.13$. Roughly correct a priori information on the data covariance matrix structure used to transform the data prior to training SLP results the higher generalization accuracy of the classifier. It was known a priori that 96 process type features of the Phonetic data could be divided into four independent groups. SLP, trained on the data transformed by the AR1 model, outperformed RDA up to 1.38 times: $\nu_{SLP\ in\ TFS\ by\ T_{AR1}} = 1.08, 1.38, 1.35$ for the learning-set sizes $N = 32, 48, 80$. Use of the Block-diagonal model with four structured blocks by the AR1 model increased the gain up to 1.68 times: $\nu_{SLP\ in\ TFS\ by\ T_{BD\&AR1}} = 1.64, 1.21, 1.68$.

Usually when the learning-set is small, the model parameters are estimated inexactly, and then the gain is not significant. The efficacy of the structurization increases with the increase in N. For the Sonar data and learning-set $N = 20$ the MA2 model applied in SLP training yielded the gain $\nu_{SLP\ in\ TFS\ by\ T_{MA2}} = 1.01$, while for $N = 80$ the gain was essentially higher: $\nu_{SLP\ in\ TFS\ by\ T_{MA2}} = 1.93$.

There were real world data sets, for which no one of the models investigated was suitable. An example is the Musk data, where we have not obtained the gain applying structurization methodology neither in LDA design nor in SLP training process. Only the ordinary transformation with the conventional sample CM and subsequent use of SLP improved the classification accuracy in the large training set case: $\nu_{SLP\ in\ TFS\ by\ Tconv} = 1.44$ for $N = 184$. It shows that populations do not posses a covariance matrix neither of the process type, nor of the Block-diagonal type.

LDA with the structured sample CM results the lower generalization accuracy than SLP trained in the transformed feature space and stopped optimally since the assumptions on the structurization model and Gaussian distribution of the data often do not correspond to reality. In the training process, the single layer perceptron corrects these crude hypothesis. Some models (e.g., ARMA2,2 for Stock data) resulted in decrease in the accuracy even of the optimally stopped SLP. It may indicate that the data needs more complicated model than the ones used.

Table 4

Mean generalization error EP_N^{RDA} of standard RDA and the relative efficacies $\nu = \frac{P_N^{RDA}}{P_N^{classiifer}}$ of SLP, trained in the original and transformed feature space using conventional CM and CM structured by Toeplitz, Circular, 1st order AR, 4th order AR, 1st order MA, 2nd order MA, 1st-1st order ARMA, 2nd-2nd order ARMA models (I – EP^{RDA}, II – $\nu_{SLPinOFS}$, III – $\nu_{SLPinTFSbyT_{circular}}$, IV – $\nu_{SLPinTFSbyT_{Toeplitz}}$, V – $\nu_{SLPinTFSbyT_{AR1}}$, VI – $\nu_{SLPinTFSbyT_{AR4}}$, VII – $\nu_{SLPinTFSbyT_{MA1}}$, VIII – $\nu_{SLPinTFSbyT_{MA2}}$, IX – $\nu_{SLPinTFSbyT_{ARMA1,1}}$, X – $\nu_{SLPinTFSbyT_{ARMA2,2}}$, XI – $\nu_{SLPinTFSbyT_{conv}}$)

Classif. method Data	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Vowels $N=9$	0.08	0.97	1.09	1.19	0.91	0.8	0.88	0.87	0.87	0.84	0.32
Vowels $N=14$	0.07	1.06	1.08	1.05	0.97	0.88	0.94	0.92	0.91	0.90	0.20
Vowels $N=54$	0.03	1.05	1.12	1.10	1.03	1.05	1.04	1.04	1.04	1.04	0.84
Mammo $N=10$	0.18	0.98	1.02	0.96	0.96	0.98	0.97	0.97	0.96	0.90	0.71
Mammo $N=20$	0.09	1.02	1.11	1.02	1.07	1.04	1.14	1.03	1.08	1.03	0.47
Mammo $N=25$	0.07	1.19	1.44	1.33	1.54	1.38	1.46	1.41	1.40	1.34	0.43
Sonar $N=20$	0.22	1.08	0.98	0.96	0.99	0.99	1.05	1.01	0.98	0.98	0.66
Sonar $N=30$	0.19	1.11	1.03	1.03	1.04	1.04	1.08	1.06	1.04	1.03	0.61
Sonar $N=80$	0.11	1.45	1.98	1.93	1.93	1.94	1.7	1.93	1.89	1.92	1.89
Ionosph $N=11$	0.16	1.00	1.01	0.99	1.05	1.13	1.07	1.01	1.06	1.04	0.54
Ionosph $N=16$	0.14	1.05	1.01	0.97	1.06	1.10	1.07	1.04	1.08	1.08	0.48
Ionosph $N=66$	0.10	1.31	1.35	1.32	1.29	1.33	1.27	1.30	1.29	1.30	1.23
Musk $N=55$	0.18	1.01	1.03	0.97	1.00	1.00	1.00	1.01	1.00	1.01	0.59
Musk $N=83$	0.14	0.97	1.00	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.52
Musk $N=180$	0.08	0.74	0.80	0.78	0.73	0.74	0.73	0.73	0.73	0.73	1.44
Satellite $N=12$	0.04	1.02	0.87	0.84	1.02	0.73	1.03	1.19	1.04	0.98	0.32
Satellite $N=18$	0.04	1.12	0.98	0.82	1.10	0.74	1.14	1.24	1.15	1.00	0.17
Satellite $N=72$	0.04	1.87	1.37	1.13	1.88	1.21	2.14	2.01	2.13	1.48	0.94
Thyroid $N=6$	0.05	1.02	0.90	0.79	0.97	1.00	1.02	1.11	1.09	1.06	0.27
Thyroid $N=9$	0.04	1.04	0.98	0.84	1.01	0.9	1.09	1.10	1.09	1.00	0.40
Thyroid $N=36$	0.03	1.57	1.55	1.45	1.54	1.52	1.72	1.67	1.67	1.60	1.14
Lung $N=22$	0.24	0.99	0.99	0.99	1.03	1.03	1.01	1.03	1.03	1.03	0.70
Lung $N=33$	0.23	1.07	1.09	1.10	1.10	1.15	1.12	1.13	1.12	1.14	0.64
Lung $N=132$	0.07	0.56	0.79	0.77	0.64	0.69	0.68	0.69	0.69	0.69	1.90
Phonetic $N=32$	0.02	1.03	0.73	0.63	1.08	0.78	0.92	0.8	0.89	0.76	0.38
Phonetic $N=48$	0.02	1.13	0.92	0.76	1.38	0.92	1.18	0.87	1.12	0.85	0.10
Phonetic $N=80$	0.02	1.03	0.83	0.88	1.35	0.95	1.14	0.95	1.03	0.95	0.46
Stock $N=31$	0.27	0.68	1.52	1.38	0.69	0.69	0.68	0.68	0.68	0.67	0.87
Stock $N=46$	0.22	0.62	1.57	1.43	0.62	0.62	0.62	0.62	0.62	0.60	0.68
Stock $N=184$	0.10	0.44	1.34	1.01	0.44	0.44	0.43	0.43	0.42	0.40	1.41

Table 5

Mean generalization error EP_N^{RDA} of standard RDA and the relative efficacies $\nu = \dfrac{P_V^{RDA}}{P_N^{classiifer}}$ of SLP, trained in the transformed feature space using CM structured by BD model of a general form, BD model with blocks structured by the process type models, BD Markov model (I – EP_{RDA}, II – $\nu_{SLPinTFSbyT_{BD}}$, III – $\nu_{SLPinTFSbyT_{BD\&circular}}$, IV – $\nu_{SLPinTFSbyT_{BD\&Toeplitz}}$, V – $\nu_{SLPinTFSbyT_{BD\&AR1}}$, VI – $\nu_{SLPinTFSbyT_{BD\&AR4}}$, VII – $\nu_{SLPinTFSbyT_{BD\&MA1}}$, VIII – $\nu_{SLPinTFSbyT_{BD\&MA2}}$, IX – $\nu_{SLPinTFSbyT_{BD\&ARMA1,1}}$, X – $\nu_{SLPinTFSbyT_{BD\&ARMA2,2}}$, XI – $\nu_{SLPinTFSbyT_{BD\&Markov}}$)

Classif. method Data	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Mammo $N=10$	0.18	0.68	0.85	0.78	0.89	0.85	0.90	0.88	0.90	0.88	–
Mammo $N=20$	0.09	0.56	1.16	1.03	1.23	1.23	1.12	1.12	1.08	1.14	–
Mammo $N=25$	0.07	0.92	1.29	1.18	1.36	1.26	1.32	1.29	1.29	1.30	–
Thyroid $N=6$	0.05	0.47	0.78	0.74	0.83	0.53	0.61	0.63	0.60	0.58	–
Thyroid $N=9$	0.04	0.68	0.90	0.77	0.85	0.54	0.66	0.64	0.67	0.59	–
Thyroid $N=36$	0.03	1.61	1.84	1.75	1.69	0.83	1.34	1.22	1.33	1.01	–
Lung $N=22$	0.24	1.36	1.05	1.02	1.06	1.02	1.04	1.03	1.03	1.01	1.38
Lung $N=33$	0.23	1.62	1.11	1.10	1.08	1.06	1.07	1.07	1.06	1.06	1.64
Lung $N=132$	0.07	2.05	0.80	0.84	0.69	0.69	0.72	0.71	0.67	0.60	1.90
Phonetic $N=32$	0.02	1.01	1.53	1.38	1.64	0.74	1.14	0.88	1.07	0.71	0.42
Phonetic $N=48$	0.02	0.85	1.22	0.95	1.21	0.65	1.05	0.90	0.95	0.75	0.47
Phonetic $N=80$	0.02	1.68	1.47	1.31	1.68	1.02	1.40	1.16	1.27	1.02	0.64
Stock $N=31$	0.27	0.91	0.72	0.73	0.69	0.69	0.68	0.69	0.69	0.69	0.55
Stock $N=46$	0.22	1.00	0.65	0.65	0.61	0.62	0.60	0.62	0.61	0.62	0.44
Stock $N=184$	0.10	1.30	0.50	0.50	0.45	0.45	0.43	0.45	0.43	0.44	0.21

In our experiments, we used the optimal RDA regularization parameter λ and the optimal stopping moment of SLP training which were determined using the test set. Therefore we have adjusted to the test set, and both RDA and SLP results presented in Tables 4 and 5 are optimistically biased. Nevertheless, we have adjustment for both competing methods, the RDA and SLP. Due to the significant reduction in the generalization error obtained for some data types in 25 independent experiments one may hope that the proposed method is an efficient tool to be used and to be studied in a further research.

5. Conclusions

In this paper we considered a usefulness of the process type and the Block-diagonal type covariance matrix structurization to reduce the generalization error of the statistical linear classifiers and the single layer perceptron. While applying the structurization methodology to solve ten multidimensional real world pattern classification problems by means of the modified liner discriminant analysis, in most cases we did not achieve a significant gain (except for one data set with one model). For some cases only slightly worse results were obtained in comparison with RDA with an optimal regularization parameter λ.

The efficacy of the structurization increases if the structurized matrix estimates are used to transform the data into the spherical one, and the optimally stopped SLP is applied afterwards. This is a new way to incorporate the statistical knowledge in SLP training process.

In the training process, the single layer perceptron can correct crude hypothesis on the data normality, common covariance matrix of the populations, the structurization model – the factors that cause the low generalization accuracy of LDA. Typically, the efficacy of the structurization increases with an increase in the learning-set size as the accuracy of the estimated model parameters becomes higher.

The standard problem that arises while applying the structurization methodology in the classifier design is the selection of the model to structurize the covariance matrix. The are several arguments which do not allow to use the simplified analytical expressions of the asymptotic PMC of Fisher LDA with the structured covariance matrix to select the right model. Main arguments are: the formulae are derived for Gaussian populations with common covariance matrix and the real parameters of the data are unknown to the researcher. In general case an expression of expected PMC, suitable for all models, is not available (yet). Moreover, it would be inaccurate for non-Gaussian case. Therefore, the most suitable structurization model can be chosen on the basis of a priori information on the data and the cross validation error estimation method. In reality, however, there exists data for which no one of 18 models considered in this paper results the improvement in the generalization accuracy. More models to structurize the covariance matrix should be developed and investigated.

One more problem is associated with the determination of the optimal stopping moment in the SLP training. In the present paper we used the test set for this purpose as well for the estimation of the optimal RDA regularization parameter λ. In practical work, an additional validation set should be chosen to decide when to stop to train the SLP classifier, and which structurization method to choose.

In this paper we used a simple, possibly statistically inefficient, statistical procedures to estimate unknown parameters of the process and block type models. In principle, more accurate optimal algorithms are possible, probably leading to an increase in the efficacy of the structurization, however demanding more computer time.

In principle our simulation studies show the applicability of the structurization technique in the classifier design to solve the real world pattern classification problems. A proper choice of the validation-set and an extension of the new methodology to multi-category case are subjects for future research.

Acknowledgments

I would like to thank my supervisor Prof. Šarūnas Raudys for inspiring the topic of this paper as well as for his stimulating critique, careful and constructing comments. I am grateful to Dr. Algimantas Rudžionis from Kaunas University of Technology, Prof. Bulent Sankur from Bogazici University, Istanbul, Prof. Mineichi Kudo from Hokaydo University, Prof. Jack Sklansky from UCA, Irvine, and Prof. Allen Long from City University, London, for providing the real world data sets.

References

Box, G.E.P., G.M. Jenkins (1974). *Time Series Analysis. Forecasting and Control*. Mir, Moscow (in Russian).

Bryant, J., and L.E.Jr. Gusseman (1979). Distance preserving linear feature selection. *Pattern Recognition*. **5/6**, 347–352.

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

Deev, A.D. (1974). Discriminant function designed on independent blocks of variables. *Proc. Acad. of Sci. of USSR, Eng. Cybernetics*, USSR J., **12**, 153–156 (in Russian).

Friedman J.M.(1989). Regularized discriminant analysis. *American Statistical Association*, **84**, 165–175.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Han, C.P. (1970). Distribution of discriminant function in circular models. *Ann. Inst. Stat. Mathematics*, **22**(1), 117–175.

Isermann, R. (1981). *Digital Control Systems*. Springer-Verlag, Berlin.

Kligienė, N. (1977). Asymptotic estimate of the probability of misclassification of autoregressive sequences. *Statistical Problems of Control*, **19**, 81–101 (in Russian).

Kligys, V. (1981). On the classification of multivariate Markov sequences. *Statistical Problems of Control*, **50**, 57–75 (in Russian).

Kligys, V. (1984). Investigation of the algorithms of classification by lots. *Statistical Problems of Control*, **68**, 95–106 (in Russian).

Meshalkin, L.D., and V.I. Serdobolskij (1978). Errors in classifying multivariate observations. *Theory of Probabilities and Applications*, **23**(4), 772–781 (in Russian).

Morgera, D., and D.B. Cooper (1977). Structurized estimation: Sample size reduction for adaptive pattern classification. *IEEE Trans. Information Theory*, **23**, 728–741.

Raudys, Š. (1967). On determining training sample size of linear classificator. *Computing Systems*, **28**, Novosibirsk, 79–87 (in Russian).

Raudys, Š. (1972). On the amount of a priori information in the designing the classification algorithm. *Proc. Acad. of Sci. of USSR, Eng. Cybernetics*, **14**, 168–174 (in Russian).

Raudys, Š. (1973). Estimation of probability of misclassification. *Statistical Problems of Control*, **5**, 10–45 (in Russian).

Raudys, Š. (1976). Limitation of sample size in classification problems, *Statistical Problems of Control*, **18**, 1–186.

Raudys, Š. (1991). Methods for overcoming dimensionality problems in statistical pattern recognition. A review. *Zavodskaya Laboratorya*, **57**(3), 45–55 (in Russian).

Raudys, Š. (1998). Evolution and generalization of a single neurone: I. Single-layer perceptron as seven statistical classifiers. *Neural Networks*, **11**, 283–296.

Raudys, Š., V. Pikelis (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. on PAMI,* **PAMI-2**(3), 242–252.

Raudys, Š., and A. Saudargienė (1998). Structures of the covariance matrices in the classifier design. *Advances in Pattern Recognition. Springer Lecture notes in computer science*. Vol. 1451 (Proc. joint IAPR int. workshops / SSPR'98 and SPR'98, August 11–13, 1998, Sydney, Australia), 583–592.

Sammon, J.W.Jr. (1970). An optimal discriminant plane. *IEEE Trans. Comput.,* **C-19**, 826–829.

A.Saudargienė graduated from Kaunas University of Technology in 1994. She is a Ph. D. program student at the Institute of Mathematics and Informatics, Vilnius. Her research interests include design of statistical classifiers and artificial feed-forward neural networks.

Kovariacinės matricos struktūrizavimas pagal procesų tipo ir blokinius-diagonalinius modelius

Aušra SAUDARGIENĖ

Straipsnyje nagrinėjama imties kovariacinės matricos struktūrizavimo įtaka statistinių tiesinių klasifikatorių ir optimaliai apmokyto vienasluoksnio perceptrono mažų mokymo imčių savybėms. Analizuojama struktūrizavimo modelio parinkimo problema. Aprašyti procesų tipo bei blokiniai diagonaliniai modeliai, pritaikyti dešimčiai realių klasifikavimo uždavinių spręsti. Nauja klasifikatorių sudarymo metodika – kovariacinės matricos struktūrizavimas, duomenų transformavimas į sferinius ir vienasluoksnio perceptrono panaudojimas – daugumoje atvejų leido gauti žymų klasifikavimo klaidos sumažėjimą.