Structures of the Covariance Matrices in the Classifier Design

Šarūnas Raudys and Aušra Saudargienė

Institute of Mathematics and Informatics
Akademijos 4, Vilnius 2600, Lithuania
e-mail: raudys@klt.mii.lt

Abstract *- Structurization of the covariance matrices helps to reduce a number of parameters to be estimated. When assumptions on the structure of the matrix are correct the structurization of the covariance matrix helps to reduce the generalization error in small learning-set cases. Efficacy of the matrix structurization increases if one decorrelates and scales the data, and uses the optimally stopped single layer perceptron classifier afterwards.*

Index terms*: regularized discriminant analysis, learning-set size, generalization, dimensionality, covariance matrix, parameters reduction, single layer perceptron.*

1. Introduction.

An essential factor while designing any pattern recognition system is a learning-set size / dimensionality ratio. In a standard linear and quadratic discrimianant analysis, one needs to estimate populations covariance matrices and invert them. When p, the dimensionality of the feature vector, exceeds n, the number of observations used to estimate the covariance matrix S, this matrix becomes singular and one can not invert it. Similar problem arise when n is close to p.

There is a number of ways to overcome this kind of difficulties (see e.g. Raudys, 1991). We'll categorise these techniques into following three groups:

a) dimensionality reduction by feature extraction or feature selection,

b) regularization of the sample covariance matrix. The simplest and the most popular example is a use of the shrinkage (ridge) estimate

$$\mathbf{S}^{RDA}=\mathbf{S}+\lambda\mathbf{I}, \qquad (1)$$

where S is the conventional maximum likelihood estimate of the covariance matrix Σ, I is a pxp identity matrix and λ is a positive regularization constant (Friedman, 1989, McLachlan, 1992),

c) structurization of a true covariance matrix Σ, and its description by a small number of parameters. Examples are assumption that Σ is a diagonal matrix, it has a block structure, Σ is a Toeplitz matrix of general form, Σ is circular, Σ describes an autoregression process, e.t.c.

From theoretical studies it is known small sample properties of the statistical pattern classifiers depend on a number of parameters r used to characterise the covariance matrix. When the number of parameters r is proportional to dimensionality p, asymptotically for large n , r and p, estimation of the *common* for both pattern classes covariance matrix (CM) Σ does not affect the increase in the

generalization error (Raudys, 1972, Deev, 1974, Meshalkin & Serdobolskij, 1978, see also Raudys & Pikelis, 1980; Raudys & Jain, 1991). It means, for large n, r and p, one can expect high efficacy of structurization of the covariance matrix .

In this paper we'll analyse the third group of techniques more thoroughly. We consider several popular and unpopular structures of CM discussed in Raudys (1991) and show that when assumptions on the structure of the matrix are correct the structurization of CM helps to reduce the generalization error in small learning-set cases. Efficacy of the matrix structurization increases if one uses the information about the structure in order to decorrelate and normalise the data, and uses the optimally stopped single layer perceptron classifier afterwards.

Notations and abbreviations:

N - number of learning vectors in one class, p - number of the features (dimensionality),
RDA - regularized discriminant analysis, LDA - linear discriminant analysis,
EDC - Euclidean distance classifier, AR - autoregression model,
OFS - original feature space, TFS - transformed feature space,
SLP - single layer perceptron, CM - covariance matrix,

2. Structures of the covariance matrix.

Use of the Gaussian model to describe the distribution density of the pattern classes leads to the quadratic discriminant function. In two category case, one needs to estimate two p-variate mean vectors and two pxp covariance matrices, altogether $2p+p(p+1)$ parameters. An assumption that the pattern classes share the common CM Σ leads to the standard Fisher linear discriminant function (DF) with smaller number of parameters to be estimated from the learning-set

$$g(\mathbf{x}) = (\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}))' \mathbf{S}^{-1} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}), \qquad (2)$$

where

$\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ are sample means,
S is the sample estimate of the covariance matrix, and
$\mathbf{x} = (x_1, ..., x_p)'$ is a p-variate vector to be classified.

When one assumes that the features are independent, instead of S one uses a diagonal variance matrix **D** composed from diagonal elements of S. In certain applications, the features can be grouped into blocks of the features, and one of possible ways to reduce the number of parameters is to assume that the blocks are statistically independent. In the high dimensional cases, the number of parameters required to characterise CM

$$\Sigma^{BL} = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Sigma_h \end{bmatrix} \tag{3}$$

is reduced immensely.

An interesting and useful model which requires a small number of parameters to describe a joint dependence between all p variables is a model based on an approximation of a joint probability distribution by the first order *tree dependence*

$$f(x_1, x_2, ..., x_p) = f(x_1) f(x_2 | x_1) f(x_3 | x_2, x_1) ... f(x_p | x_{p-1}, ..., x_2, x_1). \tag{4}$$

In this model, it is assumed that each variable is conditioned upon, at most, one another variable. Then probability density function (4) can be written in the following form:

$$f(x_1, x_2, ..., x_p) = \prod_{j=1}^{p} f(x_j | x_{m_j}) \quad (0 \le m_j \le p) \tag{5}$$

where a sequence m_2, ..., m_p constitutes a graph of connections (an unknown permutation of the integers 1, 2,..., p) and $f(x_j | x_0)$, by definition is equal to $f(x_j)$. In a general case, the covariance matrix have $p * p$ non-zero elements. An inverse of this matrix Σ^{-1} which has to be used to design the classifier, however, has only $2p$ different non-zero elements. It is a result of the assumption that each component of the vector x depends only on one another component. This important model lacked to be described in all pattern recognition textbooks and remained practically unnoticed in pattern recognition community.

There is a number of models that allow to evaluate temporal or spatial dependence between the data points. The recognition of the temporal and spatial objects and phenomena requires to work with vectors of very high dimensionality. It is an advantageous area of application of the constrained statistical classifiers. Therefore an expertise gained in this knowledge area probably can be useful in ANN design.

Let the components x_1, x_2, ... , x_{p-1}, x_p of the multivariate vector x are measurements differing in time or in space, and assume they are stationary random process. Then the covariance matrix has Toeplitz structure. Only p parameters (δ_1, δ_2,... , δ_p) describe the structure of dependence between the variables. A number of special models, such as circular, autoregression, moving average, ARMA, and others, allow to reduce the number of parameters even more. In the circular model, we have a symmetry: $\delta_2 = \delta_p$, $\delta_3 = \delta_{p-1}$, $\delta_4 = \delta_{p-2}$, e.t.c.. This model is competent to estimate random periodical processes. A number of parameters q in the autorgression model, depends of the model's order, and typically $q << p$.

3. Simulation experiments.

The efficacy of structurization of the covariance matrix was performed by means of simulation. In the experiments, we estimated parameters of different modifications of the linear discriminant function by using different randomly chosen learning-sets and estimated the generalization error on a large test-set (for a real world data) or calculated it analytically (for an artificial Gaussian data). Two category case was analysed.

Following linear classifiers were analysed in this section:

- RDA - the standard linear regularized discriminant analysis (RDA) with the optimal λ evaluated from 50 estimates of the generalization error,
- the standard linear Fisher classifier (when $n < p$, we used a pseudoinverse of the covariance matrix),
- EDC - the Euclidean distance classifier, where it is assumed $\Sigma=\sigma^2 \mathbf{I}$,
- Tree - LDA with the tree type dependence structured covariance matrix,
- Toep - LDA with the Toeplitz structured covariance matrix,
- Circ - LDA with the circular structured covariance matrix,
- AR1 - LDA with the first order autoregression structured covariance matrix,
- AR4 - LDA with the fourth order autoregression structured covariance matrix.

We concentrated our analysis mainly on a case when the number of learning examples $n= N_1+N_2 = 2N$ is smaller than the number of dimensions p of the feature vector. Therefore in our investigation, we used 40-variate artificial Gaussian data similar to two 40-variate Friedman (1989) Gaussian data sets used in analysis of the linear RDA where the first or last features were most informative. In our analysis we used also eleven 40-variate Gaussian data sets where various types of assumptions on the structure of the covariance matrices were fulfilled.

In addition to 13 artificial data types, in our research, we also used *five real world data sets.* In 28-variate (spectral and cepstral features) vowels data, we had 400 vowels in one class pronounced by 20 speakers, and in 66-variate (spectral and cepstral features) lung noise data, we had 180 vectors measured on 18 patients in one class. 65-variate (shape, size, histogram statistics, Gabor wavelet response, etc.) mammogram data set consists of 57 bening and 29 malignant mammograms, and 60-variate (energy within a particular frequency over a certain period of time) sonar data set represents two classes that describe sonar signals bounced off a metal cylinder and those bounced off a cylindrical rock (111 and 97 patterns, respectively). Two class 33-variate ionosphere data set contains 127 and 226 patterns. While training the classifiers, we have chosen N vectors from each class randomly, and tested the classifiers on all vectors. Asymptotic error rates obtained by using different structurization methods are summarized in Table 1.

In all experiments with artificial Gaussian data we have chosen $N_1=N_2 = N=13$, the same dimensionality/ sample size ratio as in the Friedman's experiments with the LDA. All experiments were performed 25 times with each size of the learning-set.

Table 1. Asymptotic error rates of the data.

No	Model	Data	LDA	Circ	Toepl	Tree	AR1	AR4
1	Circ. 1	C30	0,03	0,03	0,03	0,05	0,28	0,21
2	Circ. 2	C15	0,03	0,03	0,03	0,05	0,19	0,14
3	Circ. 3	C05	0,03	0,03	0,03	0,04	0,06	0,05
4	Toepl 1	T1105	0,03	0,05	0,03	0,04	0,06	0,05
5	Toepl 2	T1054	0,03	0,05	0,03	0,04	0,04	0,04
6	Toepl 3	T31010	0,03	0,04	0,03	0,03	0,03	0,03
7	Tree 1	Tree1	0,03	0,09	0,09	0,03	0,06	0,09
8	Tree 2	Tree2	0,03	0,28	0,29	0,03	0,15	0,28
9	Tree 3	Tree3	0,03	0,10	0,10	0,03	0,07	0,10
10	AR 1	AR046	0,03	0,03	0,03	0,03	0,03	0,03
11	AR 4	AR034	0,03	0,04	0,03	0,03	0,07	0,03
12	Fried f	FλM f	0,03	0,10	0,10	0,09	0,30	0,10
13	Fried l	FλM l	0,03	0,04	0,04	0,04	0,30	0,04
14	Wowel	Wovels	0,01	0,06	0,05	0,05	0,06	0,07
15	Lung	Lung	0,05	0,23	0,23	0,26	0,29	0,30
16	Mamm.	Mamm.	0.01	0.01	0.01	0.01	0.01	0.01
17	Sonar	Sonar	0.087	0.164	0.149	0.221	0.188	0.173
18	Ionosph	Ionosph	0.103	0.145	0108	0.137	0.177	0.145

Results are presented in Table 2. In the first column, we present the code of the data and N, the learning set size (for the real world data). In the second one, we present mean values of the generalization error of the standard RDA with optimal value of λ. In following columns, we present relative eficacies of different classifiers: the generalization error of our bench-mark method - the standard RDA divided by the generalization error of the classifier under consideration in this column: $\gamma = P_n^{RDA}/P_n^{classifier}$. Gamma ($\gamma$) values are presented in bold for these Gaussian models where assumptions of the covariance matrices structure are correct.

From the simulation experiments we see the RDA always outperform the standard methods: the Fisher LDA and EDC. When assumptions on the structure of the matrix are correct the structurization of the covariance matrix often helps to reduce the generalization error in small learning-set cases. In some cases, the gain can be very high. E.g. for data AR034 we obtained a gain - 4.59 times on average for 25 randomly chosen learning sets. For some configurations of the data and learning-set sizes, even for cases when the postulated dependence model was correct, the structurization resulted no gain in comparison with the regularized discriminant analysis. Examples are all Gaussian data sets generated according the Toeplitz model. For these three data models even simple EDC performed better than the linear DF with the structured covariance matrix.

Structurization by the first order tree dependence, the Toeplitc, AR and circular models resulted no gain while applied to all real data sets. For the lung data the representation of 66x66 covariance matrix in a block form (3) composed from six 11x11 blocks in the diagonal resulted the asymptotic error 0.13, and helped to obtain a certain gain: in 25 experiments with learning sets of size N=22 the

generalization error was decreased 1.20 times ($EP_n^{RDA} = 0.24$), for $N = 33$ - 1.31 times ($EP_n^{RDA} = 0.23$), however for N=132, $\gamma = 0.52$ ($EP_n^{RDA} = 0.07$).

Table 2. The mean generalization error EP_n^{RDA} of the standard linear RDA and the relative efficacies $\gamma = P_n^{RDA}/P_n^{classifier}$ of the Fisher classifier, EDC, and LDA with use of different structured covariance matrices.

Classif. Method Data	EP^{RDA}	γEuclid	γF&Circ	γF&Tp	γF&Tree	γF&AR1	γF&AR4	γFisher
C30	0,12	0,43	**1,95**	0,55	0,69	0,43	0,43	0,63
C15	0,11	0,62	**1,76**	0,75	0,8	0,62	0,62	0,70
C05	0,05	0,95	**0,92**	0,72	0,73	0,87	0,81	0,43
T1105	0,05	0,9	0,59	**0,67**	0,81	0,89	0,83	0,41
T1054	0,05	0,96	0,6	**0,69**	0,67	0,83	0,45	0,38
T31010	0,04	0,98	0,55	**0,55**	0,51	0,71	0,56	0,32
Tree1	0,15	0,72	0,91	0,88	**1,1**	0,76	0,92	0,58
Tree2	0,2	0,77	0,95	0,92	**1,6**	0,74	1,00	0,58
Tree3	0,3	0,62	0,67	0,67	**1,23**	0,62	0,64	0,75
AR046	0,16	0,76	2,26	2,15	1,15	**2,65**	**2,9**	0,55
AR034	0,26	0,60	3,08	3,70	1,02	**0,63**	**4,59**	0,62
FλM f	0,21	0,84	0,83	0,83	1,03	0,84	0,86	0,67
FλM 1	0,05	0,99	0,97	0,91	0,67	0,99	0,99	0,31
Wov.9	0,08	0,70	0,90	0,99	0,91	0,74	0,62	0,49
Wov.14	0,07	0,7	0,89	0,83	0,82	0,77	0,65	0,27
Wov.56	0,03	0,35	0,44	0,53	0,48	0,39	0,35	0,63
Ma 10	0,18	0,55	0,65	0,63	0,97	0,61	0,62	0,66
Ma 20	0,09	0,37	0,42	0,43	0,84	0,39	0,41	0,54
Ma 25	0,07	0,26	0,32	0,33	0,69	0,32	0,32	0,51
Son20	0,22	0,70	0,80	0,74	0,83	0,80	0,78	0,8
Son 30	0,19	0,68	0,79	0,74	0,76	0,77	0,76	0,65
Son 80	0,11	0,41	0,66	0,65	0,55	0,60	0,58	0,92
Iono11	0,16	0,72	0,86	0,74	0,88	0,78	0,89	0,64
Iono16	0,14	0,67	0,79	0,73	0,82	0,72	0,81	0,52
Iono66	0,10	0,57	0,75	0,91	0,79	0,60	0,76	0,90

4. Data transformations and the single layer perceptron.

Recently it became known, that after the first total gradient training iteration of the single layer perceptron (SLP), one can obtain the Euclidean distance classifier and move further towards RDA and the standard Fisher linear DF if certain conditions are satisfied (the centre of the data is moved to the zero point, for N_2=N_1 one uses symmetrical targets for both pattern classes, starts training from zero weights, and uses the total gradient training). Thus, if the training is successful and one succeeds to stop training optimally, one can obtain the optimal RDA by using this iterative numerical method. In further training, the SLP classifier can move towards the minimum empirical error and the support vector (maximum margin)

classifiers (Raudys, 1996, 1998). Therefore, if the data differs from Gaussian with common covariance matrix then, in principle, in further training one can expect to obtain smaller generalization error.

Iterative training of the single layer perceptron becomes difficult when variances of the data are different in various directions; it means when eigenvalues of the covariance matrix Σ are essentially different. We can try to equalise the variances by transforming the data by means of rotation and scaling: $\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{T}'\mathbf{x}$, where $\mathbf{D}$ and $\mathbf{T}$ are $p \times p$ diagonal eigenvalue and $p \times p$ eigenvectors matrix of the sample covariance matrix S. Then sample covariance matrix $\mathbf{S}_y$ of the vector $\mathbf{y}$ will be the identity matrix. After the first learning iteration we obtain the discrimianant function

$$g(\mathbf{y}) = (\mathbf{y} - \frac{1}{2}(\bar{\mathbf{y}}^{(1)} + \bar{\mathbf{y}}^{(2)}))'(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}))k_E = (\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}))'\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}))k_E, \quad (6)$$

where $\bar{\mathbf{y}}^{(1)} = \mathbf{D}^{-1/2}\mathbf{T}'\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{y}}^{(2)} = \mathbf{D}^{-1/2}\mathbf{T}'\bar{\mathbf{x}}^{(2)}$, and k_E is a constant.

It means, after the training in the y space, after the first iteration we obtain the classifier that is equivalent to the Euclidean distance classifier in the transformed (y) space (TFS), and the standard linear Fisher DF in the original (x) feature space. Let now transform the data by means of the matrix $\mathbf{G}_{RDA} = (\mathbf{D}+\lambda\mathbf{I})^{-1/2}\mathbf{T}$: $\mathbf{y} = \mathbf{G}_{RDA}\mathbf{x}$. Then after the first iteration we obtain the classifier that is equivalent to RDA analysis in the x space. When we transform the data by means of matrix $\mathbf{G}_{TREE} = \mathbf{D}_{TREE}^{-1/2}\mathbf{T}_{TREE}'$: $\mathbf{y} = \mathbf{G}_{TREE}\mathbf{x}$, then after the first iteration we obtain the classifier that is equivalent to the LDA with tree structured covariance matrix in the original feature space (OFS). In above equation, $\mathbf{D}_{TREE}$ and $\mathbf{T}_{TREE}$ are are $p \times p$ diagonal eigenvalue and $p \times p$ eigenvectors matrix of the tree structurized estimate of the covariance matrix. The same considerations are valid for other structurization methods. When the data is Gaussian with the common for all classes covariance matrix, and the postulated feature dependence model is correct, then there is a small chance to reduce the generalization error in further training. For a non-Gaussian data, models with different CM in both pattern classes, and cases when one postulates the structure of CM incorrectly, however, an additional use of SLP can result a certain success.

Therefore, in a second part of our experimental work, we tested the nonlinear SLP. In simulation experiments we translated the learning data centre $0.5(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ into the zero point, initialised the SLP with the zero weight vector, used the sigmoid activation function and trained the perceptron in a batch-mode with a standard back propagation algorithm using targets 0 and 1. In experiments with small learning sets, in order to obtain a large margin quickly, we increased the learning step η progressively with each iteration number t: $\eta = 0.2*1.03^t$; $t_{max} = 500$. We trained the SLP in the original (x) space and afterwards in the transformed space ($\mathbf{y} = \mathbf{T}\mathbf{x}$). In each experiment, we determined an optimal stopping moment t_{opt} from estimates of the generalization error obtained either from the test-set data (for the real world data) or calculated analytically (for Gaussian data). Results are presented in Table 3.

Table 3. The mean generalization error EP_n^{RDA} of the standard linear RDA and the relative eficacies $\gamma = P_n^{RDA}/P_n^{classifier}$ of the SLP classifier without and after the data transformation performed by use of differently structured covariance matrices.

Class. Method Data	EP^{RDA}	γ SLP in OFS	γ SLP in TFS by $\mathbf{T}_{Circ}$	γ SLP in TFS by $\mathbf{T}_{Top}$	γ SLP in TFS by $\mathbf{T}_{Tree}$	γ SLP in TFS by $\mathbf{T}_{AR1}$	γ SLP in TFS by $\mathbf{T}_{AR4}$	γ SLP in TFS by $\mathbf{T}_{conv}$
C30	0,12	0,9	**1,97**	0,78	1,08	0,5	0,49	0,38
C15	0,11	0,91	**1,77**	0,98	0,88	0,67	0,66	0,36
C05	0,05	0,99	**0,92**	0,74	0,78	0,87	0,82	0,2
T1105	0,05	1,00	0,64	**0,71**	0,89	0,94	0,85	0,19
T1054	0,05	1,00	0,62	**0,7**	0,65	0,84	0,48	0,16
T31010	0,04	1,00	0,56	**0,56**	0,62	0,72	0,57	0,14
Tree1	0,15	1,00	1,13	1,09	**1,21**	1,05	1,18	0,47
Tree2	0,2	0,98	1,12	1,08	**1,68**	0,95	1,18	0,57
Tree3	0,3	0,94	1,07	0,99	**1,54**	0,9	1,09	0,78
AR046	0,16	1,00	2,29	2,17	1,22	**2,67**	**2,92**	0,45
AR034	0,26	0,89	3,45	3,84	1,23	**1,05**	**4,62**	0,67
FλM f	0,21	0,96	0,96	0,96	1,09	0,97	0,98	0,63
Fλ M 1	0,05	1,01	0,98	0,92	0,69	1,00	1,00	0,21
Wov.9	0,08	0,97	1,09	**1,19**	0,99	0,91	0,8	0,32
Wov.14	0,07	1,06	1,08	**1,05**	0,99	0,97	0,88	0,2
Wov.56	0,03	1,05	1,12	**1,1**	1,08	1,03	1,05	0,84
Ma 10	0,18	0,98	1,02	0,96	**1,2**	0,96	0,98	0,71
Ma 20	0,09	1,02	1,11	1,02	**1,64**	1,07	1,04	0,47
Ma 25	0,07	1,19	1,44	1,33	**1,41**	1,54	1,38	0,43
Son20	0,22	**1,08**	0,98	0,96	0,96	0,99	0,99	0,66
Son 30	0,19	**1,11**	1,03	1,03	0,99	1,04	1,04	0,61
Son 80	0,11	**1,45**	1,98	1,93	1,93	1,93	1,94	1,89
Iono11	0,16	1,00	1,01	0,99	0,96	1,05	**1,13**	0,54
Iono16	0,14	1,05	1,01	0,97	0,94	1,06	**1,1**	0,48
Iono66	0,1	1,31	1,35	1,32	1,36	1,29	**1,33**	1,23

In the Table we have the average results. The results advocate, the efficacy of the optimally stopped SLP in OFS is almost the same as that of the RDA. Higher values of the generalization error of the SLP typically were associated with these few cases when 500 iterations were not sufficient to train the perceptron. It is worth to note that after the transformation $\mathbf{y} = \mathbf{Tx}$, we obtained a significant increase in the learning speed: typically only few training iterations were sufficient to obtain the smallest generalization error.

Practically in all experiments, the transformations of the data according the structured estimate of the sample CM and subsequent use of the optimally stopped SLP helped to improve efficacy of the covariance matrix structurization. E.g. for the Gaussian data C30 the additional use of the data transformation based circular model and SLP helped to reduce the generalization error 1.97 times in comparison with the best statistical method - RDA. For some models additional use of SLP leads to a significant gain: e.g. with tree type dependence model we had athe gain γ_{Tree}=1.23,

and now $\gamma_{\text{Tree\&SLP}} = 1.56$. For Toeplitz model, however, possibly, we have chosen too "difficult" model and estimated model's parameters inefficiently and did not obtained any gain: e.g. $\gamma_{\text{F\&Toep}} = 0.67$, and γ SLP in TS by $\mathbf{T}_{\text{Toeplitz}} = 0.71$, however for other syntetic data (AR034) the Toeplitz model structurization resulted a significant gain: γ SLP in TFS by $\mathbf{T}_{\text{Toeplitz}} = 3.84$.

Similar observations we have for the real world data too. For the wovels data and learning-set sizes 9, 14 and 56 for the Toeplitz model we have had γLDA&Toep = 0.99, 0.83, and 0.53, i.e. no improvement in comparison with RDA. Now , with the data transformations and subsequent SLP training, we have γ SLP in TS by $\mathbf{T}_{\text{Toeplitz}} = 1.19$, 1.05, and 1.10. For the lung data and learning-set sizes 22, 33 and 132 for the block matrix model we have had γF&Toep = 1.20, 1.31 and 0.52. After the transformations and SLP we have γ SLP in TS by $\mathbf{T}_{\text{Toeplitz}} = 1.36$, 1.62, and 2.05, i.e., an obvious improvement.

5. Concluding remarks.

The efficacy of the structurization of the covariance matrix depends both on the data as well as on peculiarities of the particular learning-set. In a part of the experiments, we obtained no or very insignificant gain.

In our comparative experiments, we used optimal values of λ for the standard RDA estimated from the test-set, or analytically for the Gaussian pattern classes. Therefore our experiments result optimistic estimates of the efficacy of the regularized discriminant analysis and pessimistic estimates of our efficacy coefficients $\gamma = P_n^{\text{RDA}} / P_n^{\text{classifier}}$. Nevertheless, we see that for certain Gaussian models use of correct assumptions on the structure covariance matrix result an obvious gain. In most cases, structurization outperform the standard linear discriminant analysis, and in some cases, it loses against the regularized discriminant analysis. Therefore in practical applications, one obligatory must use an additional validation-set in order to decide which classification method to use. In extremely high-dimensional cases, however, one can expect these parameters can be estimated by the leaving-one out or rotation methods. The situation is analogous to RDA where we have to choose the optimal value of the regularization parameter λ.

Our experiments with the artificial and real world data sets have demonstrated that the efficacy of the matrix structurization increases if one uses the information about the structure in order to decorrelate and scale the data, and uses the optimally stopped single layer perceptron classifier afterwards. It is *a new way to incorporate an additional, the statistical, information in the perceptron training process.* We see, that it is useful to structurize the covariance matrix. More statistical models and more real world data sets should be analysed in future research work. Special numerical calculation schemes that speed up the calculations in the model validation stage should be developed.

In present paper we analysed the efficacy of the structurization of the covariance matrix and the joint use of the data transformation and the SLP in the

linear discriminant analysis problem only. No doubt, this structurization of the sample covariance matrix can be used in pattern classification with different covariance matrices as well as in regression tasks.

Acknowledgement.

The authors are thankful to Dr. Algimantas Rudzionis from Kaunas University of Technology, Dr. Bulent Sankur from Istanbul Bogazici University, Prof. Mineichi Kudo from Hokaydo University, and Professor Jack Sklansky from UCA, Irvine for providing the real world data sets for the experiments.

References.

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. *J. American Statistical Association*, **84**, 165-175.

McLachlan, G.J.(1992). *Discriminant Analysis and Statistical Pattern Recognition*. Willey.

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

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

Raudys, S. (1991). Methods for overcoming dimensionality problems in statistical pattern recognition. A review. *Zavodskaya Laboratorya* (*Factory Lab.*, USSR Journal), Moscow: Nauka, **3**, 45 & 49-55 (in Russian).

Raudys, S. (1996). Linear classifiers in perceptron design, ICPR13, *Proc. 13th Int. Conf. on Pattern Recognition* (Vienna, Austria, Aug.25-29) Vol. 4, Track D: Parallel and Connectionist Systems, IEEE Computer Society Press, Los Alamitos, 1996, 763-767.

Raudys, S. (1998). Evolution and generalization of a single neurone. Part I. SLP as Seven statistical classifiers. *Neural Networks* (accepted).

Raudys, S. & Jain, A.K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners- *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **PAMI-13**, 252-264.

Raudys, S. & Pikelis, V.(1980). On dimensionality, sample size, classification error and complexity of the classification algorithm in pattern recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **PAMI-2** (3), 242-252.