PERGAMON

Pattern Recognition 33 (2000) 1989–1998

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Scaled rotation regularization

Šarūnas Raudys*

Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania

Received 15 December 1998; accepted 12 July 1999

Abstract

A new regularization method - a scaled rotation - is proposed and compared with the standard linear regularized discriminant analysis. A sense of the method consists in the singular value decomposition $\mathbf{S}=\mathbf{TDT}'$ of a sample covariance matrix $\mathbf{S}$ and a use of the following representation of an inverse of the covariance matrix $\mathbf{S}^{-1} = \mathbf{T}^{\alpha}(\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{T}^{\alpha\prime}$. For certain data structures the scaled rotation helps to reduce the generalization error in small learning-set and high dimensionality cases. Efficacy of the scaled rotation increases if one transforms the data by $\mathbf{y} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha\prime}\mathbf{x}$ and uses an optimally stopped single layer perceptron classifier afterwards. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Regularized discriminant analysis; Learning-set size; Dimensionality; Single-layer perceptron; Generalization; Scaled rotation

1. Introduction

An essential factor while designing any pattern recognition system is a learning-set size/dimensionality ratio. In standard linear and quadratic discrimianant analysis, one needs to estimate the population covariance matrix and invert it. When p, the dimensionality of the feature vector, exceeds n, the number of observations used to estimate the covariance matrix Σ, the estimate $\mathbf{S}$ of the matrix becomes singular and one cannot invert it. A similar problem arises when n is close to p.

There are a number of ways to overcome these kinds of difficulties. We can categorize these techniques into the following three groups [1]:

(a) dimensionality reduction by feature extraction or feature selection,
(b) structurization of the true covariance matrix Σ, and its description by a small number of parameters. Examples are diagonal structurization, block structurization, or Toeplitz matrix.
(c) regularization of the sample covariance matrix. The simplest and the most popular example is a use of a shrinkage (ridge) estimate [2,3]

$$\mathbf{S}^{\mathrm{RDA}} = \mathbf{S} + \lambda\mathbf{I}, \tag{1}$$

where $\mathbf{S}$ is the conventional maximum likelihood estimate of the covariance matrix Σ, $\mathbf{I}$ is a $p \times p$ identity matrix and λ is a positive regularization constant. In this paper we will analyse the third group more thoroughly.

The classical linear regularized discriminant analysis (RDA) can be obtained or explained from different approaches. It can be obtained from a predictive Bayes approach if one assumes a prior distribution of Σ^{-1} to be a Wishart $W_m(p, \mathbf{I})$, where m is a number of degrees of freedom [4,5]. Then λ, the regularization parameter, is uniquely related to m and also n, the sample size used to estimate the covariance matrix. The weight vector of the linear RDA

$$\mathbf{w}^{\mathrm{RDA}} = (\mathbf{S} + \lambda\mathbf{I})^{-1}\,(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})), \tag{2}$$

can also be obtained after an adaptive minimization of the conventional sum of squares cost function of the standard linear perceptron with a "weight decay" term [6]. In Eq. (2), $\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}$ are sample mean vectors. The weight vector (2) also can be obtained in the adaptive

*Fax: + 370-2-729-209.
E-mail address: raudys@das.mii.lt (Š. Raudys).

0031-3203/00/$20.00 © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(99)00183-1

linear (and even nonlinear SLP, if we start training from zero initial weights) SLP training after the first few iterations if the following conditions are satisfied: (E1) the centre of the data is moved to the zero point, (E2) if $N_2 = N_1 = N$, we use symmetrical targets, (E3) we start training from zero weights, and (E4) we use the total gradient training [7,8]. Then, after the first iteration, we obtain the classifier equivalent to the Euclidean distance classifier (EDC), and after t iterations, a weight vector that is equivalent to that resulting from the linear RDA. The regularization parameter λ changes during training; it decreases with an increase in the number of iterations t.

In a modification of the standard ridge estimate, instead of the term $\lambda\mathbf{I}$, some authors use $\lambda\,\mathrm{tr}(\mathbf{S})/p\mathbf{I}$. In our experiments reported in Section 3, we also tested the regularized estimate $\mathbf{S}^{\mathrm{JS}} + \lambda\mathbf{I}$, with James and Stein estimator $\mathbf{S}^{\mathrm{JS}} = \mathbf{T}\mathbf{D}^{\mathrm{JS}}\mathbf{T}'$, where $\mathbf{T}$ is a $p\times p$ eigenvectors matrix of $\mathbf{S}, \mathbf{D}^{\mathrm{JS}}$ is a $p\times p$ diagonal matrix composed of estimates $d_j^{\mathrm{JS}} = (n-1)/(n+p-2j)\,d_j$, and $d_1, \dots, d_p$ are eigenvalues of the matrix $\mathbf{S}$ in the singular-value decomposition $\mathbf{S}=\mathbf{T}\mathbf{D}\mathbf{T}'$ [9].

One more way to overcome numerical difficulties associated with the badly conditioned covariance matrix is to use pseudoinversion. Such approaches have been used in statistical pattern recognition [10–12]. In the standard linear RDA we add the constant λ to each singular value d_j of $\mathbf{S}$. In the pseudoinversion approach, one ignores directions with zero eigenvalues:

$$\mathbf{S}^{\mathrm{Pseudoinverse}} = \mathbf{T}_r(\mathbf{D}_r)^{-1}\mathbf{T}_r', \tag{3}$$

where $\mathbf{D}_r$ is an $r\times r$ diagonal matrix composed of r nonzero eigenvalues $d_1, d_2, \dots, d_r$, $\mathbf{T}_r$ is a $p\times r$ left-eigenvectors matrix corresponding to the eigenvalues $d_1, \dots, d_r$. It was noticed experimentally [12] and shown theoretically [13] that usage of the pseudoinversion helps to reduce the generalization error when $n < p$.

In the standard linear RDA, when the parameter λ decreases and approaches zero, the linear classifier formed using Eq. (1) tends to the standard Fisher linear discriminant function (DF).

When λ increases and tends to infinity, the sample covariance matrix $\mathbf{S}$ is increasingly ignored, and the linear classifier tends to the Euclidean distance classifier (EDC). In the multivariate Gaussian $\mathrm{N}(\mu_i, \Sigma)$ case (we will call it the GCCM class model of the data), the generalization error of EDC (the SLP classifier after the first iteration) [14–17] is

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

where

$$\Phi\{a\} = \int_{-\infty}^{a} (2\pi)^{-1/2}\sigma^{-1}\exp\{-t^2/(2\sigma^2)\}\,\mathrm{d}t,$$

$N = N_1 = N_2$.

$$T_\mu^* = 1 + \frac{2p^*}{\delta^{*2}N}, \quad \delta^* = \frac{\mu'\mu}{\sqrt{\mu'\Sigma\mu}}$$

and determines the separability of the pattern classes,

$$\mu = \mu_1 - \mu_2 \quad \text{and} \quad p^* = \frac{(\mu'\mu)^2(\mathrm{tr}\,\Sigma^2)}{(\mu'\Sigma\mu)^2}$$

which we call the *intrinsic dimensionality* of the GCCM data for the Euclidean distance classifier. From the definition, it is clear that $1 < p^* < \infty$.

The term T_μ^* indicates that small sample properties of this classifier are highly affected by the true distribution densities of the classes. When Σ is proportional to the identity matrix $p^* = p$, the classifier is relatively insensitive to the learning-set size. However, for data configurations with high intrinsic dimensionality p^*, EDC is very sensitive to the learning-set size.

The generalization error of the standard Fisher linear classifier (the SLP classifier after many iterations) [15–17] is

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

where δ is the Mahalanobis distance, the term $T_\mu = 1 + 2p/\delta^2N$ arises from an inexact sample estimation of the mean vectors of the classes, and the term $T_\Sigma = 1 + p/(2N-p)$ arises from an inexact sample estimation of the covariance matrix.

For very small samples, and many configurations of parameters of distributions of the pattern classes, EDC can outperform the standard Fisher classifier. Therefore a change of λ balances the classifier between the simple EDC and the more complex Fisher classifier, and can help to reduce the generalization error in small learning-set cases [6]. However configurations of the parameters exist (e.g. when Σ is close to singular, and/or $p^* \gg p$), where EDC is very sensitive to the finiteness of the learning set size and results in very high asymptotic errors [17]. *In such situations*, both the Fisher classifier and EDC perform poorly, and *the standard regularization is not effective*.

An objective of the present paper is to find a way to improve performance of the standard linear RDA.

2. A scaled rotation

Above we presented one possible reason for the low efficacy of the standard RDA in certain situations — the potentially high value of p^*, the "intrinsic dimensionality" of the data for EDC. A possible way to improve the estimate of the sample covariance matrix is to regularize the matrix $\mathbf{S}$ only in directions with nonzero eigenvalues.

Thus, we reduce the matrix **S** regularization in *directions associated with zero variability.*

In the pseudoinversion approach and different modifications of RDA reported in the introduction, one actually manipulates only the eigenvalues of the sample covariance matrix. *Nothing is done with the eigenvectors.* In the small learning-set case, however, the eigenvectors are estimated with error as well, and one may hope that a certain simplification of these estimates at times can be useful too.

Let us introduce a new "regularized" estimate of the eigenvectors matrix in the singular-value decomposition $\mathbf{S}=\mathbf{TDT}'$,

$$\mathbf{T}_{\text{new}} = \mathbf{T}^{\alpha}, \tag{6}$$

where a scalar parameter α controls a similarity of $\mathbf{T}_{\text{new}}$ to the conventional sample estimate $\mathbf{T}$ $(\alpha = 1)$, and to the identity matrix $\mathbf{I}$ $(\alpha = 0)$.

Then, we propose the following new *double regularized* estimate of the covariance matrix:

$$\mathbf{S}^{\text{SR}} = \mathbf{T}^{\alpha}(\mathbf{D} + \lambda\mathbf{I})\mathbf{T}^{\alpha\prime}. \tag{7}$$

In this estimate, we control the *rotation* of the sample covariance matrix by the parameter α. We control the scaling of the covariance matrix by the parameter λ: the modified diagonal matrix $\mathbf{D} + \lambda\mathbf{I}$ is composed of the components $d_1 + \lambda, d_2 + \lambda, \dots, d_p + \lambda$. We will call this regularization technique the "*scaled rotation*" (SR).

3. Simulation experiments with artificial data

The scaled rotation is controlled by two parameters, α and λ. In order to investigate the usefulness of different regularization methods one needs to obtain analytical expressions for the generalization error. This can be done by using Taylor series expansions (see e.g. an expansion with respect to λ in our earlier paper [6]). It is difficult, however, to obtain expansions accurate enough for a wide range of values of α and λ. Therefore, the analysis of the efficacy of the new technique was performed by means of simulation. In the simulation experiments that follow, we examine the two-class case of a linear discriminant anlaysis with artificial Gaussian $N(\mu_1, \Sigma)$, $N(\mu_2, \Sigma)$ data models. We estimated the weight vector $(w_0, \mathbf{w})$ of different modifications of the linear discriminant function by using different randomly chosen learning-sets, and calculated the generalization error P_N analytically:

$$P_N = \frac{1}{2}\Phi\left\{-\frac{w_0 + \mathbf{w}'\mu_1}{\sqrt{\mathbf{w}'\Sigma^{-1}\mathbf{w}}}\right\} + \frac{1}{2}\Phi\left\{\frac{w_0 + \mathbf{w}'\mu_2}{\sqrt{\mathbf{w}'\Sigma^{-1}\mathbf{w}}}\right\}. \tag{8}$$

The following linear classifiers were examined in this section:

- RDA — the standard linear regularized discriminant analysis with the optimal λ evaluated from minimum values of the generalization error,
- Pseudo — the pseudo-Fisher classifier with estimate (3),
- EDC — the Euclidean distance classifier,
- SR — the regularized linear discriminant analysis with the scaled rotation, where the optimal values of the parameters α and λ were chosen to minimize an estimate of the generalization error.

To determine the optimal values of the regularization parameters in each learning experiment with one learning-set, we used a grid formed by 50 values of $\lambda_1 = \lambda/(1 - \lambda)$ in an interval (0, 1), and $k_\alpha = 25$ values of α $(\alpha = 0, 1/16, 1/8, \dots, 23/16, 3/2)$. For calculations we used a Matlab package. E.g., to obtain $\mathbf{T}^\alpha$, we used commands: $\mathbf{T}_2 = \text{sqrtm}\,(\mathbf{T})$; $\mathbf{T}_4 = \text{sqrtm}\,(\mathbf{T}_2)$; $\mathbf{T}_8 = \text{sqrtm}\,(\mathbf{T}_4)$; $\mathbf{T}_{16} = \text{sqrtm}\,(\mathbf{T}_8)$. Then for $\alpha = 7/16$ we wrote: if $a = = 7\mathbf{T}^\alpha = \text{real}\,(\mathbf{T}_4{*}\mathbf{T}_{16}{*}\mathbf{T}_8)$; end. In principle, one can use more "natural" parameterizations of the covariance matrix, such as described in Pinheiro and Bates [18] and used in subsequent studies [19]. Incidentally, we also analysed two modifications of the standard linear RDA with $\lambda\text{tr}(\mathbf{S})/p\mathbf{I}$, and a James and Stein estimator [9] with $d_j^{\text{JS}} = (n - 1)/(n + p - 2j)\,dj$; however no gain was obtained.

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. We used 40-variate artificial Gaussian data vectors as Friedman [20] did in his analysis on RDA in the quadratic case, and we have chosen $N_1 = N_2 = N = 13$, the same dimensionality/ sample size ratio, as in Friedman's experiments. Data were generated according to the following configurations of eigenvalues and mean differences (with associated codings in parentheses):

(1) True eigenvalues set to $\lambda_j^{\text{true}} = (9(j - 1)/\ (p - 1) + 1)^2$, values used by Friedman (coded by $F\lambda$).

(2) True eigenvalues set to $\lambda_j^{\text{true}} = 100\,\mathrm{e}^{(j-1)/2} + 0.05$ (coded by expλ) along with

  (1) mean differences $\Delta\mu_j^{\text{true}} = 2.5\sqrt{d_j/p}\,(p - j)/(p/2 - 1)$ (that is, first features are most informative — coded as Fμfirst)

  (2) mean differences $\Delta\mu_j^{\text{true}} = 2.5\sqrt{d_j/p}\,(j - 1)/(p/2 - 1)$ (last features are most informative — coded as Fμlast).

  (3) $\Delta\mu_j^{\text{true}} = d_j$ for $j = 1, 2, \dots, p/2$, and $\Delta\mu_j^{\text{true}} = 0$ f or $j = p/2 + 1, 2, \dots, p$ (the last $p/2$ features are uninformative — coded as Myμ).

The specific combinations of the eignevalues/mean differences that we used are evident in the tables as the "product" of the corresponding codes. For example, in order to examine the influence of directions with zero discriminitive information, we combined the eigenvalues Fλ with the mean differences in Myμ to give Fλ Myμ.

The differences $\Delta\mu_j^{\text{true}}$ in the means of all artificial data sets were normalized in order to obtain the Mahalanobis distance $\delta = 3.76$ — this corresponds to the Bayes error $P_{\mathrm{B}} = 0.03$. Afterwards the data were rotated using *randomly* generated orthonormal matrices $\mathbf{T}_{\text{random}}$. A left/upper triangle of $\mathbf{T}_{\text{random}} = ((t_{ij}))$ was composed from $(p-1)p/2$ random $N(\delta_{ij}, 1)$ variables (δ_{ij} is the Kronecker symbol). The remaining $p\,(p+1)/2$ components of $\mathbf{T}_{\text{random}}$ were calculated in a way to fulfil the orthonormality condition $\mathbf{T}'_{\text{random}}\mathbf{T}_{\text{random}\cdot} = \mathbf{I}$. The orthonormal transformation of the data does not change the asymptotic classification error, however, it affects training peculiarities of the nonlinear single-layer perceptron used in our experiments. For each set of λ_j^{true} and $\Delta\mu_j^{\text{true}}(j = 40)$ we generated 100 random 40×40 rotation matrices $\mathbf{T}_{\text{random}}$, and a new random learning set composed of $n = 2N = 26$ 40-variate random vectors.

The results from this simulation are presented in the first five columns of Table 1. In the first column, we present the code for the configuration, the dimensionality of the feature vector p, and the learning-set size N. In the second column, we present mean values and standard deviations of the generalization error of the standard RDA with optimal value of λ. In the following columns, we present the relative efficacy of different classifiers $\gamma = P_n^{\text{RDA}}/P_n^{\text{classifier}}$. This is the generalization error of our benchmark method — the optimized standard RDA divided by the generalization error of the classifier under consideration in the particular column. Estimates presented in the last two columns correspond to additional use of the single-layer perceptron and will be discussed later in Section 5. The upper rows correspond to mean values and the lower ones to standard deviations.

The fifth row of the Table 1, "FλMyμT39", corresponds to one particular randomly generated transformation matrix, the 39th, $\mathbf{T}^{*39}_{\text{random}}$, for which with model F$\lambda$My$\mu$ we obtained a high efficacy for the scaled rotation regularization. This row of the table presents results where we generated 100 independent random learning sets using only one particular transformation matrix $\mathbf{T}^{*39}_{\text{random}}$. Mean values corresponding to one (the 71th) random learning-set with the highest efficacy of the SR method are displayed in the sixth row (coded $N = 13^*$).

The simulation experiments with the artificial Gaussian data show that *data structures exist where the scaled rotation increases the efficacy of the standard RDA*. The highest average gain in efficacy we obtained using 100 randomly chosen learning-sets was 1.64, and 2.65, using one particular learning-set. Thus, the generalization error was reduced from 0.212 for the standard RDA to 0.08 for the scaled rotation. Note that this comparison is made using one of the most effective [20] regularization techniques for statistical classification methods — the optimized RDA. However, the scaled rotation is not a universal method: for some data structures we obtained no or very insignificant gain. For the GCCM data model the success of the scaled rotation depends on $\Delta\mu$ and Σ: a mutual distribution of true eigenvalues δ_j(true) of the matrix Σ and the difference in the means $\Delta\mu_j$(true), as well as on the transformation matrix $\mathbf{T}_{\text{random}}$. Individual peculiarities of the learning-sets play a very important role too.

4. Simulation experiments with 10 real data sets

In order to see whether the data structures favourable for the scaled rotation exist in real-world pattern classification problems, we also used 10 real data sets. In Table 2, we present short characteristics of the data sets:

Table 1
The mean generalization error EP_n^{RDA} of the standard linear RDA and the relative efficacy of the pseudo-Fisher classifier, EDC, the scaled rotation, SLP, and SLP with scaled rotation (upper rows), and standard deviations (lower rows)

Classification method data type	EP_n RDA	γ Pseudo-Fisher	γ EDC	γ Scaled rotation	γ SLP	γ Scaled rotation and SLP
FλFμfirst	0.210	0.634	0.828	1.212	0.966	1.286
$p = 40, N = 13$	0.036	0.129	0.099	0.188	0.058	0.226
expλFμfirst	0.108	0.662	0.605	1.037	0.866	1.054
$p = 40, N = 13$	0.022	0.119	0.160	0.085	0.164	0.092
FλFμlast	0.057	0.344	0.988	1.043	1.017	1.055
$p = 40, N = 13$	0.007	0.101	0.034	0.066	0.044	0.069
FλMyμ	0.145	0.523	0.784	1.379	0.976	1.452
$p = 40, N = 13$	0.034	0.142	0.115	0.300	0.086	0.328
FλMyμT39	0.140	0.560	0.789	1.639	0.981	1.772
$p = 40, N = 13$	0.031	0.164	0.107	0.353	0.074	0.352
FλMyμT39, $N = 13^*$	0.212	0.753	0.863	2.649	1.039	2.855

Table 2
Short characteristics of the real-world data sets

No.	Data name	p	Features' characteristics	N_{g1}	N_{g2}	$P^{\mathrm{FIS}}_{R\infty}$	$P^{\mathrm{SLP}}_{R\infty}$
1. Sat	Satellite	36	Energy in 9 pixels × 4 bands	479	415	0.031	0.001
2. Chr	Chromosomes	30	Banding patterns	500	500	0.014	0.006
3. Vow	Vowels	28	Spectral and cepstral	400	400	0.013	0.003
4. Lung	Lung sounds	66	Spectral and cepstral	180	180	0.050	0.053
5. Stock	Stock	92	4 days history	610	770	0.056	0.010
6. Thyr	Thyroid	18	6 continuous and 12 binary	93	191	0.021	0.000
7. Musk	Musk	166	Shape of the molecule	207	269	0.050	0.057
8. Iono	Ionosphere	33	Autocorrelation of radar return	127	226	0.103	0.031
9. Mam	Mammograms	65	Shape, histogram, wavlets	57	29	0.000	0.000
10. Son	Sonar	60	Energy in frequency bands	111	97	0.087	0.024

the number and a code for each data set, the dimensionality of the feature vector p, brief characteristics of the feature vectors, the size of our general population: N_{g1}, N_{g2}, and the "resubstitution" error estimates of the linear Fisher and SLP classifiers $P^{\mathrm{FIS}}_{R\infty}$ and $P^{\mathrm{SLP}}_{R\infty}$. These estimates served as estimates of the asymptotic error and helped to characterize deviations of the real-world data sets from the GCCM model.

For training, we selected very small randomly chosen subsets of the data composed of N vectors from each pattern class. The generalization error of the linear classifiers formed using a particular random learning-set was estimated experimentally by classifying all available vectors (our general population). For each data set we randomly picked 25 learning sets with the same sizes each time, and formed the classifiers discussed in the artificial data experiments.

Our simulation studies confirmed the previous conclusion: the efficacy of each particular classification method depends on the data type. Moreover, for the real data, we noticed that the efficacy of different methods notably varies with N, the size of the randomly chosen learning-set. For 10 different randomly chosen learning-sets (the learning-set size $N = 18$), we present in the first three columns of Table 3 the generalization errors of the standard RDA (with optimal value of λ), the pseudo-Fisher classifier, and scaled rotation for the 36-variate Satellite data. (We will discuss estimates presented in the last three columns later in Section 5.) In the table we see that the scaled rotation sometimes helps to reduce the generalization error. The best classifier differs with each learning-set. In general, the pseudo-Fisher classifier loses against the standard RDA. However, for some data types and small learning-sets ($2N < p$) even this "unoptimal" classifier at times results in the smallest generalization error.

Average results obtained in 25 experiments with randomly chosen learning-sets are presented in the first five columns of Table 4. Presentation of results is similar to that used in Table 1. We will discuss estimates presented in the last three columns later in Section 5. Note that the upper rows correspond to mean values and the lower ones to standard deviations.

Table 3
The generalization errors of the standard RDA (with optimal λ), the pseudo-Fisher classifier, the scaled rotation, the optimally stopped SLP in the original feature space, SLP in an optimally rotated space (R & SLP), and SLP after the optimal scaled rotation (the best classifier is in bold). Ten rows in the table represent 10 randomly chosen learning-sets ($N = 72$)

RDA	**P-F**	**SR**	**SLP**	**R & SLP**	**SR & SLP**
0.0470	0.0626	0.0380	0.0291	**0.0268**	0.0380
0.0492	0.0559	0.0369	0.0503	0.0503	**0.0358**
0.0414	0.0783	0.0403	0.0302	**0.0257**	0.0336
0.0380	0.0447	**0.0380**	0.0414	0.0414	**0.0380**
0.0257	0.0671	0.0257	0.0246	**0.0235**	0.0257
0.0358	0.1029	0.0358	0.0380	0.0380	**0.0347**
0.0459	0.0660	**0.0369**	0.0414	0.0414	**0.0369**
0.0425	0.0604	0.0369	**0.0168**	**0.0168**	0.0302
0.0414	0.0526	0.0313	0.0336	0.0324	**0.0291**
0.0268	0.0570	0.0268	0.0268	**0.0257**	**0.0257**

As a rule, distributions in the real-world differ from the GCCM model. For the non-Gaussian data sets considered here, the scaled rotation is not very efficient. However, from 10 real-world problems analysed in the confines of this paper, only in two or three problems we have rather a notable gain in comparison with the optimal regularized dicriminant analysis. In the next section we will show that an additional use of the optimally stopped SLP increases the efficacy of the scaled rotation.

5. Data transformations, and the single-layer perceptron

We will assume that conditions E1–E4 are satisfied when training the SLP classifier. Then, after the first total gradient training iteration, one obtains the Euclidean

Table 4
The mean generalization error EP_n^{RDA} of the standard linear RDA and the relative efficacy of the pseudo-Fisher classifier, EDC, the scaled rotation, SLP, and SLP with the scaled rotation

Class. method data type	Learn. set size Data	EP_n of RDA	γ Pseudo-Fisher	γ Scaled rotation	γ SLP	γ Optimal rotation & SLP	γ Scaled rotation & SLP.	γ of the "best" classif.
Sat	18	0.043	0.200	1.093	1.212	1.224	1.158	1.328
		0.010	0.146	0.088	0.463	0.475	0.160	0.422
Sat	56	0.038	0.827	1.059	1.680	1.715	1.285	1.733
		0.006	0.111	0.064	0.581	0.583	0.360	0.560
Sat	72	0.034	0.841	1.026	1.767	1.785	1.204	1.794
		0.005	0.104	0.054	0.672	0.676	0.152	0.666
Chro	10	0.036	0.445	1.163	1.003	1.023	1.242	1.279
		0.012	0.256	0.268	0.305	0.322	0.315	0.344
Chro	20	0.026	0.249	1.079	0.913	0.922	1.128	1.131
		0.007	0.139	0.088	0.149	0.146	0.106	0.103
Vow	9	0.102	0.485	1.102	1.143	1.182	1.139	1.259
		0.056	0.252	0.116	0.375	0.472	0.129	0.452
Vow	14	0.073	0.274	1.112	1.116	1.126	1.171	1.254
		0.025	0.123	0.162	0.312	0.313	0.167	0.272
Vow	56	0.033	0.646	1.014	1.138	1.142	1.163	1.273
		0.010	0.216	0.027	0.275	0.272	0.145	0.193
Lung	22	0.260	0.622	1.072	0.993	0.996	1.086	1.089
		0.050	0.113	0.055	0.064	0.065	0.059	0.058
Lung	33	0.226	0.636	1.049	0.979	0.996	1.073	1.089
		0.027	0.155	0.050	0.072	0.075	0.067	0.067
Stock	31	0.326	0.821	1.001	0.793	0.821	1.124	1.146
		0.045	0.206	0.008	0.112	0.110	0.204	0.207
Stock	46	0.339	0.763	1.000	0.943	0.870	1.641	1.678
		0.033	0.106	0.001	0.086	0.062	0.294	0.239
Thyr	6	0.447	1.247	1.983	1.002	1.347	1.983	2.079
		0.200	0.780	0.929	0.009	0.653	0.929	0.921
Thyr	9	0.397	0.718	1.813	1.004	1.090	1.813	1.813
		0.205	0.278	0.727	0.017	0.240	0.727	0.727
Musk	55	0.247	0.688	1.008	1.061	1.125	1.041	1.130
		0.026	0.114	0.011	0.065	0.082	0.033	0.078
Musk	83	0.216	0.678	1.011	1.092	1.133	1.080	1.162
		0.023	0.119	0.015	0.099	0.086	0.057	0.075
Iono	11	0.230	0.601	1.037	1.019	1.038	1.059	1.120
		0.061	0.258	0.039	0.136	0.142	0.040	0.113
Iono	16	0.191	0.497	1.089	1.073	1.145	1.138	1.210
		0.057	0.187	0.141	0.114	0.174	0.152	0.184
Mamm	10	0.194	0.556	1.152	0.589	0.591	1.181	1.181
		0.050	0.186	0.314	0.201	0.200	0.315	0.315
Sonar	20	0.283	0.734	1.056	1.064	1.072	1.074	1.105
		0.041	0.125	0.058	0.062	0.062	0.058	0.055
Sonar	30	0.239	0.723	1.058	1.120	1.130	1.084	1.168
		0.030	0.139	0.046	0.125	0.135	0.059	0.106

distance classifier and moves further toward linear RDA and the Fisher linear DF with conventional (when $N_1 + N_2 > p$) or pseudo (when $N_1 + N_2 < p$) inversion of the covariance matrix. Thus, if the training is successful and we succeed in optimally stopping training, we can obtain the optimal RDA by using this iterative numerical method. With further training, the SLP classifier can approach the robust, the minimum empirical error and the maximal margin classifiers [7,8]. Therefore, if the data differ from Gaussian with common covariance matrix then, in principle, with further training one can expect to obtain a smaller generalization error. Thus, in our experiments with the real-world data sets, we included the SLP into the set of classification methods tested. Similar to optimal RDA and the scaled rotation, the training of the perceptron was stopped optimally according to estimates of the generalization error obtained while classifying all available vectors (our general population).

Iterative training of the single-layer perceptron becomes difficult when variances of the data are different in various directions (i.e., when eigenvalues of the covariance matrix Σ are essentially different) [21]. We can try to equalize the variances by transforming the data by means of rotation and scaling: $\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{T}'\mathbf{x}$, where the $p \times p$ matrices $\mathbf{D}$ and $\mathbf{T}$ are defined by the singular-value decomposition $\mathbf{S}$=$\mathbf{TDT}'$. Then the sample covariance matrix of the vector $\mathbf{y}$ will be the identity matrix. After the first learning iteration we obtain the discrimianant function

$$g(\mathbf{y}) = (\mathbf{y} - \tfrac{1}{2}(\bar{\mathbf{y}}^{(1)} + \bar{\mathbf{y}}^{(2)}))'(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}))\,k_E$$

$$= (\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}))'\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}))\,k_E, \qquad (9)$$

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.

This means, when training in the transformed ($\mathbf{y}$) space, after the first iteration we obtain the classifier that is equivalent to the Euclidean distance classifier in the $\mathbf{y}$ space, and the standard linear Fisher DF in the original ($\mathbf{x}$) feature space. Now suppose we transform the data by means of matrix $\mathbf{G}_{\mathrm{RDA}} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}$: $\mathbf{y} = \mathbf{G}_{\mathrm{RDA}}\mathbf{x}$. Then after the first iteration we obtain the classifier that is equivalent to the linear RDA in the $\mathbf{x}$ space, and move towards the standard linear Fisher classifier. When we transform the data by means of the matrix $\mathbf{G}_{\mathrm{SR}} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha\prime}$: $\mathbf{y} = \mathbf{G}_{\mathrm{SR}}\mathbf{x}$, after the first iteration we obtain the classifier that is equivalent to the scaled rotation regularization in the original ($\mathbf{x}$) space. When we transform the data by the matrix $\mathbf{T}^{\beta\prime}(\beta > 0)$: $\mathbf{y} = \mathbf{T}^{\beta\prime}\mathbf{x}$, after the first iteration we obtain EDC in both, the original ($\mathbf{x}$), and in the transformed ($\mathbf{y}$) spaces. In subsequent iterations, we have RDA, and move towards the Fisher classifier. When the data are Gaussian having a common class covariance matrix, there is then only a minor chance to reduce the generalization error in further training. For non-Gaussian data and for different class covariance matrices, however, one can expect a certain success.

Therefore, in the second part of our experimental work, we tested the nonlinear SLP. In simulation experiments, we translated the learning data centre $(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})/2$ into the zero point, initialized 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 our experiments with very small learning-sets, in order to obtain a large margin quickly, we increased the learning step η progressively with each iteration number t: $\eta = 0.2 * 1.03^t (t_{\max} = 500)$. The progressive increase in the learning step prevents a gradient of the SLP cost function to converge to zero and allows to obtain the wider gamma of statistical classifiers in the preceptron training [7,8]. We trained the SLP classifier in the following situations:

(1) in the original ($\mathbf{x}$) space,
(2) in the transformed space ($\mathbf{y} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha\prime}\mathbf{x}$), where the optimal values of α and λ were determined in the scaled rotation experiment (SR & SLP), and
(3) in the transformed space ($\mathbf{y} = \mathbf{T}^{\beta\prime}\mathbf{x}$), where we selected the optimal β value after $k_\alpha = 25$ SLP training sessions performed for 25 rotation matrices $\mathbf{T}_\beta$ (optimal rotation & SLP).

In each experiment, we determined t_{opt}, the optimal stopping point. For this we used experimental estimates of the generalization error obtained from the all available data or theoretical values calculated from Eq. (8). Finally, we selected the best result. The results are presented in the last two columns of Table 1 and the last four columns of Table 4. In both tables we give the average results. In addition, in Table 4, we have also standard deviations.

Results for GCCM data model in Table 1 indicate that for the Gaussian data, the efficacy of the SLP in the original space is almost the same as that of the optimal 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.

In experiments with the non-Gaussian real-world data (Table 4), we trained the perceptron for more (1200) iterations. On average, the SLP allowed better generalization than the RDA. In certain experiments, the reduction was high: e.g., 1.924% of errors for the SLP trained in the original feature space versus 3.4% of errors for RDA, and 3.3% for the scaled rotation (for the Satellite data, $N = 72$, average values). Training of the SLP in the space obtained after the optimal scaled rotation was less efficient — 2.8% of errors. However, for the Satellite data ($N = 72$) a joint optimization of the rotation (parameter α) and the number of iterations (t) resulted in the highest performance — 1.905% of classification errors, 1.785 times less than that obtained by the optimal RDA.

It is worthwhile to note that after the transformation $\mathbf{y} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha\prime}\mathbf{x}$, we obtain *a significant increase in the learning speed*: only a few training iterations were sufficient to obtain the smallest generalization error. This is an effect of the fact that succeeding this particular data transformation we have approximately spherical data, and after the first iteration we have EDC. The EDC classifier is suited to classify the spherical data very well. In the experiments with the real-world data, the type of the best classifier depends on the data type and the learning-set size, and to a certain extent — on a particular randomly chosen learning-set. For some data sets the scaled rotation was not efficient at all (e.g. for the stock data). The same can be said about the SLP classifier in the original feature space. For this data type, very good results were obtained after a simple transformation: $\mathbf{y} = (\mathbf{D} + 0.02 * \mathbf{I})^{-1/2}\mathbf{T}'\mathbf{x}$: $\gamma = 1.138$ for $N = 31$ and $\gamma = 1.678$ for $N = 46$ (compare with values in Table 4) This transformation, however, is not universal, and does not work for all types of the data. In most cases, the additional use of SLP, however, helped to reduce the generalization error substantially. A general conclusion follows that the data transformation, and subsequent training SLP is a very useful tool, however, the transformation type (parameters α and λ) should be chosen in the experimental way for each concrete pattern classification problem and each particular learning-set.

6. Concluding remarks

The singular-value decomposition $\mathbf{S}=\mathbf{TDT}'$ represents the sample covariance matrix $\mathbf{S}$ as a function of sample eigenvalues $\mathbf{D}$ and eigenvectors $\mathbf{T}$. In the conventional regularization methods, one tries to regularize the eigenvalues only. *In our approach, we also regularize the eigenvectors matrix.* The scaled rotation $\mathbf{S}^{\mathrm{SR}} = \mathbf{T}^{\alpha}(\mathbf{D} + \lambda\mathbf{I})\mathbf{T}^{\alpha'}$ uses two regularization parameters, α and λ, and is a mathematical model for simpler parametrization of the covariance matrix.

The efficacy of the joint regularization of the eigenvectors and the eigenvalues depends on characteristics of the parameters of the distribution of the data (the mutual distribution of the eigenvalues and the components of the difference in the mean vectors of the pattern classes, and the eigenvectors matrix) as well as on peculiarities of a particular learning-set. In experiments with the artificial data, the efficacy of the scaled rotation depends on the data model. In principle, one can construct a data where the scaled rotation is largely efficient. In experiments with the artificial data reported in this paper, the maximal average reduction in the generalization error of the scaled rotation approach over the optimal standard RDA was 1.64, and 1.77 when the scaled rotation data transformation $\mathbf{y} = (\mathbf{D} + \lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha'}\mathbf{x}$ and SLP were used in addition (in the experiment with 100 learning-sets). For some data types, we obtained zero or very insignificant gain. In experiments with the real-world data, the highest gain for one particular learning-set exceeded 3. It was noticed that the highest relative gain often is obtained when the generalization error of RDA is high, and when the learning-set size is relatively small. A general conclusion follows: the scaled rotation is not a universal method. It is efficient only for certain structures of the data.

In our comparative experiments, we used optimal values of λ (for the standard RDA), α and λ (for the scaled rotation), and the number of the learning iterations t_{opt} (for the SLP classifier) estimated from the test-set data or analytically for the Gaussian pattern classes (Eq. (8)). Therefore, our estimates for the optimal RDA, scaled rotation, and SLP are optimistically biased. In real applications, we will use a cross-validation method, and will obtain a smaller gain. In spite of the fact that our experiments do not reflect absolute efficacy of the scaled rotation regularization, nevertheless, they state that, for some types of the data and sizes of the learning-sets, in principle, an introduction of the additional regularization parameter (scalar α) together with usage of SLP in the original and/or the transformed feature space can be more powerful than the standard RDA. Additional research work should be done in order to find conditions *where* the new method is efficient and *how* to estimate optimal regularization parameters α and λ cheaply. To verify and to improve the new regularization method one needs to analyse a large number of the real-world pattern classification tasks, use real, finite-sized, validation-sets to determine the optimal regularization parameters.

A multivariate analysis advocates that in very high-dimensional and large design-set cases, variances of the classification error estimates are comparatively low (see Refs. [16,22]). Therefore, in addition to a standard (the cross-validation) way to determine the optimal values of λ, α, and t_{opt} and to choose the classification method, a leaving-one out or a rotation method can be applied. Special numerical calculation schemes that speed up the calculations should be developed.

In the present paper we analysed the efficacy of the scaled rotation in the linear discriminant analysis problem only. No doubt, this sample covariance matrix regularization technique can be used in a quadratic pattern classification task with different covariance matrices as well as in regression.

7. Summary

The singular-value decomposition $\mathbf{S}=\mathbf{TDT}'$ represents the sample covariance matrix $\mathbf{S}$ as a function of sample eigenvalues $\mathbf{D}$ and eigenvectors $\mathbf{T}$. In the conventional regularization methods, one tries to regularize the eigenvalues only. *In our approach, we regularize the eigenvectors matrix.* The scaled rotation $\mathbf{S}^{\mathrm{SR}} = \mathbf{T}^{\alpha}(\mathbf{D} + \lambda\mathbf{I})\mathbf{T}^{\alpha'}$ uses two regularization parameters, α and λ, and is a mathematical model for simple parametrization of the covariance matrix.

The efficacy of the joint regularization of the eigenvectors and the eigenvalues depends on a data structure (the mutual distribution of the eigenvectors and the components of a difference in the mean vectors of the pattern classes, and the eigenvectors matrix) as well as on peculiarities of specific learning-set. In experiments with the

artificial data, the maximal reduction in the generalization error of the scaled rotation approach over the optimal standard RDA was 1.64 times, and 1.77 times when the scaled rotation data transformation and SLP were used in addition (averages in the experiment with 100 randomly chosen learning-sets). Prior to training the SLP classifier we moved the centre of the data to the zero point; if $N_2 = N_1 = N$, we used symmetrical targets, started training from zero weights, and used the total gradient training. For some data types, we obtained zero or very insignificant gain. In experiments with the real-world data, the highest gain for one particular learning-set exceeded 3 times.

In comparative experiments, we used optimal values of λ (for the standard RDA), and α and λ (for the scaled rotation), and the number of the learning iterations t_{opt} (for the SLP classifier) estimated from the test-set data or analytically for the Gaussian pattern classes. Therefore, our estimates do not reflect absolute efficacy of the scaled rotation regularization, but, nevertheless, state that, for some data structures and sizes of the learning-sets, in principle, the new regularization method together with usage of SLP in the original and/or the transformed feature space can be more powerful than the standard RDA. No doubt, this sample covariance matrix regularization technique can be used in pattern classification tasks with different covariance matrices, as well as in regression.

Acknowledgements

The author is thankful to Dr. Algimantas Rudzionis from the Speech Analysis Laboratory of Kaunas University of Technology, Dr. Bulent Sankur from Istanbul Bogazici University, Prof. Mineichi Kudo from Hokaydo University, Prof. Jack Sklansky from UCI, Irvine, Prof. Allen Long from City University, London and others for providing the real-world data sets for the experiment. The author express his gratitude to Laura Thompson from Baylor University, TX for useful remarks and aid in a final revision of the paper.

References

[1] S. Raudys, Methods to overcome dimensionality problems in statistical pattern recognition, A review paper, Zavodskaya Laboratorya (USSR J.) N.3 (1991) 45, 49–55 (in Russian).

[2] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for orthogonal problems, Technometrics 12 (1970) 55–67.

[3] P.J. Di Pillo, Biased discriminant analysis: evaluation of the optimum probability of misclassification, Commun. Statist. - Theory Methods A 8 (14) (1979) 1447–1457.

[4] S. Geisser, Posterior odds for multivariate normal classifications, J. Roy. Statist. Soc. Ser. B 21 (1) (1964) 69–76.

[5] D. Keehn, A note on learning for Gaussian properties, IEEE Trans. Inform. Theory IT- 11 (1965) 126–131.

[6] S. Raudys, M. Skurichina, Small sample properties of ridge estimate of the covariance matrix in statistical and neural net classification, in: New Trends in Probability and Statistics, Multivariate Statistics and Matrices in Statistics, Proceedings of the fifth Tartu Conference, Tartu - Puhajarve, Estonia, 23–28 May Vol. 3, VSP/TEV. Vilnius, 1994, pp. 237–245.

[7] S. Raudys, Linear classifiers in perceptron design, ICPR13, Proceedings of 13th International Conference on Pattern Recognition, Vienna, Austria, August 25–29, Vol. 4, Track D: Parallel and Connectionist Systems, IEEE Computer Society Press, Los Alamitos, 1996, pp. 763–767.

[8] S. Raudys, Evolution and generalization of a single neurone. Part I, SLP as seven statistical classifiers, Neural Networks 11 (2) (1998) 283–296.

[9] W. James, C. Stein, Estimation with quadratic loss, in: J. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1961, pp. 361–379.

[10] J. Schurmann, Polynomklassifikatoren fur Zeichenerkennung, R.Oldenbourg Verlag, Munchen and Wien, 1977.

[11] L.G. Malinovskij, Hypotheses on Subspaces in the Problem of Discriminant Analysis of Normal Populations, Nauka, Moscow, 1979, pp. 195–206 (in Russian).

[12] R.P.W. Duin, Small sample size generalization, Proceedings of 9th Scandinavian Conference on Image Analysis, June 6–9, Uppsala, Sweden, 1995.

[13] S. Raudys, R.P.W. Duin, On expected classification error of the Fisher classifier with pseudo-inverse covariance matrix, Pattern Recognition Lett. 19 (1998) 385–392.

[14] S. Raudys, On determining training sample size of a linear classificator, Computing Systems 28 (1967) 79–87 (1967, in Russian).

[15] S. Raudys, On the amount of a priori information in designing the classification algorithm, Proc. Acad. Sci. USSR, Engng. Cybernet. 14 (1972) 168–174 (in Russian).

[16] S. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-13 (1991) 252–264.

[17] S. Raudys, Evolution and generalization of a single neurone. II. Complexity of statistical classifiers and sample size considerations, Neural Networks 11 (2) (1998) 297–313.

[18] J. Pinheiro, U. Bates, Unconstrained parameterizations for variance-covariance matrices, Statist. Comput. 6 (1995) 289–296.

[19] L. Thompson, A personal communication (1999).

[20] J.M. Friedman, Regularized discriminant analysis, J. Amer. Statist. Assoc. 84 (1989) 165–175.

[21] Y. Le Cun, I. Kanter, S. Solla, Eigenvalues of covariance matrices: application to neural-network learning, Phys. Rev. Lett. 66 (18) (1991) 2396–2399.

[22] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.

About the Author—ŠARŪNAS RAUDYS was born in Kaunas, Lithuania, in 1941. He received the M.Sc. degree in electrical and computer engineering from the Kaunas University of technology in 1963, a Soviet candidate of science (Ph.D) degree from the Institute of physics and mathematics, Vilnius, 1969, and a Soviet Doctor of science degree from the Institute of electronics and computer science, Riga in 1978. Currently he is the Head of Data analysis department at the Institute of mathematics and informatics in Vilnius. His research interests include multivariate analysis, statistical pattern recognition, artificial neural networks, data mining methods and biological information processing systems. He is the author of a monograph "Small sample problems in statistical pattern recognition", and has over hundred research papers. He is an associate editor of international journals Informatica, Pattern Recognition and Image Analysis, Pattern Recognition. He was a member of Program committee and an invited speaker in 20 international conferences, a guest scientist in several research laboratories in Russia, the Netherlands, France, USA and Japan.