ELSEVIER

Pattern Recognition Letters 19 (1998) 385–392

Pattern Recognition Letters

# Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix

Sarunas Raudys [a,*], Robert P.W. Duin [b,1]

[a] *Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania*
[b] *Faculty of Applied Physics, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, Netherlands*

Received 22 January 1997; revised 22 January 1998

**Abstract**

The pseudo-Fisher linear classifier is considered as the ''diagonal'' Fisher linear classifier applied to the principal components corresponding to non-zero eigenvalues of the sample covariance matrix. An asymptotic formula for the expected (generalization) error of the Fisher classifier with the pseudo-inversion is derived which explains the peaking behaviour: with an increasing number of learning observations from one up to the number of features, the generalization error first decreases, and then starts to increase. 

*Keywords:* Statistical classification; Fisher linear discriminant; Pseudo-inversion; Generalization error; Scissors effect; Sample size; Dimensionality

## 1. Introduction. Fisher linear discriminant function and its modifications

A classical learning-set based statistical classifier designed to allocate an unknown $p$-variate vector $\boldsymbol{x}$ to one of two multivariate Gaussian populations differing in mean vectors $\mu_1$, $\mu_2$, but sharing the same covariance matrix $\Sigma$, is a linear discriminant function (DF)

$$g(\boldsymbol{x}) = \left[\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\right]' \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) = \boldsymbol{w}^{\mathrm{F}\prime}\boldsymbol{x} + w_o^{\mathrm{F}}, \tag{1}$$

where

$$\boldsymbol{w}^{\mathrm{F}} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}), \quad w_o^{\mathrm{F}} = -\tfrac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})'\boldsymbol{w}, \tag{2}$$

and $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$, $\boldsymbol{S}$ are sample maximum likelihood estimates of mean vectors $\mu_1$, $\mu_2$ and the covariance matrix $\Sigma$ (see, e.g., Anderson, 1958; Fukunaga, 1990).

The DF (1) is a plug-in type allocation (classification) rule, and in a general case, it is not an optimal sample-based decision rule. The only one approach which strictly generates the optimal allocation rules is a

* Corresponding author. E-mail: raudys@ktl.mii.lt.
[1] E-mail: duin@ph.tn.tudelft.nl.

0167-8655/98/$19.00 
PII: S0167-8655(98)00016-6

Bayes approach. It generates decision rules with *the minimal classification error for a set of allocation problems* defined by a prior distribution of parameters. It is known (Gupta, 1977) that the linear DF (1) coincides with the optimal Bayes predictive rule for a certain uniform prior distribution of $\mu_1$, $\mu_2$ and $\Sigma$ if prior probabilities of the pattern classes are equal among themselves, and for learning-set sizes $N_2 = N_1 = N$.

When the number of learning observations $n = N_1 + N_2 < p + 2$, the sample estimate of the covariance matrix $\boldsymbol{S}$ becomes singular. To enable allocation in such situations, different approaches were developed. In a part of the approaches, some assumptions on the structure of the covariance matrices are utilised, e.g., it is assumed that the covariance matrix $\Sigma$ is proportional to the identity matrix ($\Sigma = \sigma^2 \boldsymbol{I}$; then we have the Euclidean distance, or the nearest means classifier), or it is diagonal, or has a block structure. In the ''diagonal'' classifier, it is assumed that the variables are independent, i.e., $\Sigma$ is a diagonal matrix. Then we have following weights of the linear classifier:

$$\boldsymbol{w}^{\mathrm{D}} = \boldsymbol{D}^{-1}\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}\right), \quad w_o^{\mathrm{D}} = -\tfrac{1}{2}\left(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)}\right)' \boldsymbol{w}^{\mathrm{D}}, \tag{3}$$

where $\boldsymbol{D}$ is a diagonal matrix composed of the diagonal elements of matrix $\boldsymbol{S}$.

Another intermediate case between the Euclidean distance classifier and the linear Fisher classifier is the regularized linear discriminant analysis (DA). There, instead of the conventional sample estimate $\boldsymbol{S}$, one uses the shrinkage (ridge) estimate of the covariance matrix $\boldsymbol{S}_{\mathrm{R}} = \boldsymbol{S} + \lambda \boldsymbol{I}$ (Di Pillo, 1979; McLachlan, 1992).

One more alternative to the Fisher linear DF is to use a pseudo-inverse $\boldsymbol{S}*$ (see, e.g. Fukunaga, 1990) instead of $\boldsymbol{S}^{-1}$. A possible pseudo-inverse approach consists of a singular value decomposition of matrix $\boldsymbol{S}$:

$$\boldsymbol{T}\boldsymbol{S}\boldsymbol{T}' = \begin{bmatrix} \boldsymbol{d} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \end{bmatrix}$$

is an orthogonal matrix such that

$$\boldsymbol{T}\boldsymbol{S}\boldsymbol{T}' = \begin{bmatrix} \boldsymbol{d} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

$\boldsymbol{d} = \boldsymbol{t}_1 \boldsymbol{S} \boldsymbol{t}_1'$ is the $r \times r$ diagonal matrix composed of the $r = 2N - 2$ non-zero eigenvalues $d_1, d_2, \ldots, d_r$ of $\boldsymbol{S}$. Then the pseudo-inverse of matrix $\boldsymbol{S}$ is

$$\boldsymbol{S}^{*} = \boldsymbol{T}' \begin{bmatrix} \boldsymbol{d}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{T}. \tag{4}$$

This approach to find the weight vector $\boldsymbol{w}$ of the linear discriminant function in the small learning-set case was used by Schürmann (1977), Malinovskij (1979), Duin (1995).

The existence of several alternative classification rules raises a problem to choose one of them in each particular practical problem. The choice of the best type of the classifier depends on the data, complexity of the classifier and the learning-set size. Usually in the small learning-set case, the generalization error of a simple structured classification rule is lower than that of a complex classifier. Vice versa, in large learning-set cases, the generalization error of the simple structured classification rule is higher than that of complex classifiers. Typically, learning curves – plots of the generalization error with the learning-set size – of two classifiers of different complexity intersect and portray scissors (Raudys, 1970; Kanal and Chandrasekaran, 1971; Duin, 1978; Jain and Chandrasekaran, 1982; Raudys and Jain, 1991). This phenomenon is called a ''scissors effect''.

In the literature, a number of research papers has been published concerning the learning curves of the standard Fisher linear DF, the Euclidean distance classifiers, the ''diagonal'' classifiers, regularized linear discriminant analysis and other allocation rules (Deev, 1970, 1972; Raudys, 1972; Raudys and Jain, 1991;

Raudys and Skurichina, 1994; Amari et al., 1992; see also Wyman et al. (1990) and Chap. 4 in (McLachlan, 1992)). Unfortunately we do not have such results for the pseudo-Fisher linear classifier.

The importance of the analysis of the pseudo-Fisher linear classifier arises also in connection with an extensive use of artificial neural networks. It was shown that under certain conditions, the nonlinear single layer perceptron (SLP) can realise decision boundaries of the Euclidean distance classifier, the regularized linear DA, the standard Fisher classifier and the Fisher linear classifier with pseudo-inversion of the sample covariance matrix (Raudys, 1998). For the latter classifier *an unexpected behaviour was noticed*: with an increase in $N$, the generalization error first decreases, then increases, has a maximum at $N = p/2$ and then decreases again (Duin, 1995; Skurichina and Duin, 1996). An understanding of the reasons for this ''strange'' behaviour of the pseudo-Fisher classifier can help to choose proper parameters in the neural network training algorithms. An objective of the present paper is to explain this behaviour theoretically, and to obtain an equation for the generalization error of the pseudo-Fisher linear classifier.

The remaining of the paper is organised as follows. In the next section, we derive an asymptotic formula for the expected error of the Fisher DF with the pseudo-inverse which explains the peaking behaviour. The third section contains simulation results and a discussion.

## 2. The expected error of the pseudo-Fisher linear classifier

When the number of learning observations is small, the estimates $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ and $\boldsymbol{S}$ ($\boldsymbol{S}*$) become inexact and result in an increase in the classification error of observation vectors which do not participate in the design of the classification rule. We call this error rate an expected probability of misclassification or simply – a mean generalization error.

We consider the pseudo-Fisher linear classifier as the ''diagonal'' Fisher linear classifier applied to the principal components corresponding to $r$ non-zero eigenvalues of the sample covariance matrix $\boldsymbol{S}$. In the other $p - r$ directions it is orthogonal to the subspace of principal $r$ directions. Note, when $n = N_1 + N_2 < p + 2$, then $r = N_1 + N_2 - 2$. This means that *the dimensionality of the new feature space changes with n, the learning-set size*. Therefore, in the analysis of the learning curve ''the generalization error versus the learning-set size'', the increasing dimensionality plays an important role.

To simplify analytical work we consider the case in which $N_2 = N_1 = N$, the prior probabilities of the pattern classes are equal to $1/2$, and true covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}$ ($p \times p$ identity matrix). When $n < p + 2$, the rank of the sample covariance matrix $\boldsymbol{S}$ is $r = n - 2$. Denote

$$\boldsymbol{y} = \boldsymbol{t}_1 \boldsymbol{x}, \qquad \bar{\boldsymbol{y}}^{(i)} = \boldsymbol{t}_1 \bar{\boldsymbol{x}}^{(i)}, \qquad \boldsymbol{d} = \boldsymbol{t}_1 \boldsymbol{S} \boldsymbol{t}_1',$$

where $\boldsymbol{t}_1$ is a *random* $r \times p$ *orthonormal matrix* as defined above. Then the discriminant function in the new $r$-variate space is

$$g(\boldsymbol{y}) = \left[ \boldsymbol{y} - \tfrac{1}{2}\left( \bar{\boldsymbol{y}}^{(1)} + \bar{\boldsymbol{y}}^{(2)} \right) \right]' \boldsymbol{d}^{-1} \left( \bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)} \right), \tag{5}$$

where according to the above definition, $\boldsymbol{d}$ is the diagonal $r \times r$ matrix with the eigenvalues $d_1, d_2, \ldots, d_r$ in its diagonal. If the learning-set vectors are considered to be random, then the discriminant function (1) should be considered as a random variable which depends on the 3 independent $p$-variate random vectors $\boldsymbol{x}$, $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ and the random $p \times p$ matrix, $\boldsymbol{S}*$.

It is a known fact, that for independent learning-set observations also $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ and $\boldsymbol{S}$ ($\boldsymbol{S}*$) are statistically independent. Thus, the transformation matrix $\boldsymbol{t}_1$ and the vectors $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ are statistically independent too. Consequently, for a model of the spherical Gaussian distribution $N(\mu_i, \boldsymbol{I})$ of $\boldsymbol{x}$, for *any* fixed orthonormal $\boldsymbol{t}_1$ the vectors $\bar{\boldsymbol{y}}^{(i)} = \boldsymbol{t}_1 \bar{\boldsymbol{x}}^{(i)}$, have a spherical Gaussian distribution $N(\boldsymbol{t}_1 \mu_i, (1/N)\boldsymbol{I})$. Moreover, when $N_2 = N_1$ for *any* fixed orthonormal $\boldsymbol{t}_1$

$$\boldsymbol{u} = \left( \bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)} \right) - \boldsymbol{m} \sim N\left( \boldsymbol{0}, \frac{2}{N} \boldsymbol{I} \right), \qquad \boldsymbol{v} = \boldsymbol{y} - \frac{1}{2}\left( \bar{\boldsymbol{y}}^{(1)} + \bar{\boldsymbol{y}}^{(2)} \right) - \tfrac{1}{2}\boldsymbol{m} \sim N\left( \boldsymbol{0}, \left( 1 + \frac{1}{2N} \right) \boldsymbol{I} \right),$$

and both vectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, are statistically independent. In the above equations, we denoted $\boldsymbol{m} = \boldsymbol{t}_1(\mu_1 - \mu_2) = (m_1, m_2, \ldots, m_p)'$. Thus, $g(\boldsymbol{y}) = (1/2\boldsymbol{m} + \boldsymbol{v})'\boldsymbol{d}^{-1}(\boldsymbol{m} + \boldsymbol{u})$ is a function of two independent $r$-variate Gaussian random vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and of the $r \times r$ diagonal random matrix $\boldsymbol{d}$.

To obtain an asymptotic expression for the mean generalization error of the pseudo-Fisher classifier we go about in the same way as in the analysis of the diagonal classifier (Raudys, 1972). A *key point* in this analysis is the assumption that $p$, the dimensionality, and $n$, the sample size, are large. When $r$, the number of dimensions, is high, the distribution of the linear discriminant function (5) – the sum of $r$ components – approaches the Gaussian distribution. Therefore the mean generalization error approaches the following expression:

$$\frac{1}{2}\Phi\left\{-\frac{E\left[g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d})|\boldsymbol{x}\in\pi_1\right]}{\sqrt{V\left[g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d})|\boldsymbol{x}\in\pi_1\right]}}\right\} + \frac{1}{2}\Phi\left\{\frac{E\left[g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d})|\boldsymbol{x}\in\pi_2\right]}{\sqrt{V\left[g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d})|\boldsymbol{x}\in\pi_2\right]}}\right\}. \tag{6}$$

where the function $\Phi\{a\} = \int_{-\infty}^{\alpha}(2\pi)^{-1/2}\exp\{-t^2/2\}\mathrm{d}t$ is a standard cumulative Gaussian distribution function, $E$ and $V$ denote expectation and variance of the discriminant function with respect to the random vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and the random matrix $\boldsymbol{d}$.

In order to calculate the moments we make the following remarks and assumptions:

(1) We take into account that for $n < p + 2$, the transformation $\boldsymbol{y} = \boldsymbol{t}_1\boldsymbol{x}$, is a random one. Thus, it is reasonable to *assume* that in the new $r$-variate space, the components of the vector $\boldsymbol{t}_1(\mu_1 - \mu_2) = (m_1, \ldots, m_p)'$ are random Gaussian zero mean variables: $m_j \sim N(0, \delta^2/p)$, where $\delta^2 = (\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)$ is the Mahalanobis distance in the original $p$-variate space. From this assumptions it follows that a real Mahalanobis distance $\delta_{\mathrm{d}}^2 = \sum_{j=1}^{r} m_j^2$ in the $r$-variate space is a scaled chi-square random variable with a mean value $\delta^2 r/p$, and a standard deviation $\delta^2(\sqrt{2r})/(p)$. For large $p$ and $r$ *we ignore the standard deviation* and *assume* $\delta_{\mathrm{d}}^2 \approx \delta^2 r/p$. The expectation and the standard deviation of the fourth statistical moment of the Gaussian random variable are $\mu_4 = \sum_{j=1}^{r} m_j^4 = 3\delta^4 r/p^2$, and $\delta^4(\sqrt{7r})/(p^2)$. Again, for large $p$ and $r$ *we ignore the standard deviation*, and *assume* $\mu_4 \approx 3\delta^4 r/p^2$.

(2) When $\boldsymbol{x}$ is multivariate spherical Gaussian $N(\mu_1, \boldsymbol{I})$, $N(\mu_2, \boldsymbol{I})$, then the sample covariance matrix $\boldsymbol{S}$ has the Wishart $W(n-2, \boldsymbol{I})$ density (see, e.g., Anderson (1958)) with $n-2$ degrees of freedom. We *consider* that, in the transformed $r$-variate space, the variables $d_1, d_2, \ldots, d_r$, the components of the matrix $\boldsymbol{d}$, are chosen at random with respect to the components $m_1, m_2, \ldots, m_r$. In this context, it follows that randomly chosen eigenvalues $d_1, d_2, \ldots, d_r$ are statistically independent and identically distributed. Later in our analysis we use $E_{\mathrm{d}}$ and $V_{\mathrm{d}}$, the expectation and the variance of the arbitrarily chosen component $1/d_j$.

Taking into account the above estimates of $d_j$, $\delta_{\mathrm{d}}^2$ and $\mu_4$ we have

$$Eg(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d}|\boldsymbol{x}\in\pi_i) = \frac{1}{2}(-1)^{i+1}\boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{m} = \frac{1}{2}(-1)^{i+1}\delta_{\mathrm{d}}^2 E_{\mathrm{d}} = \frac{1}{2}(-1)^{i+1}\delta^2\frac{r}{p}E_{\mathrm{d}}, \tag{7}$$

$$\begin{aligned} Vg(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d}|\boldsymbol{x}\in\pi_i) &= E\left[g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d}|\boldsymbol{x}\in\pi_i)\right]^2 - \left[Eg(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d}|\boldsymbol{x}\in\pi_i)\right]^2 \\ &= \frac{1}{4}\left[\boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{m}\boldsymbol{m}'\boldsymbol{d}^{-1}\boldsymbol{m} - \boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{m}\boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{m}\right] + \boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{d}^{-1}\boldsymbol{m} \\ &\quad + \frac{1}{4}\boldsymbol{m}'E\boldsymbol{d}^{-1}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{d}^{-1}\boldsymbol{m} + E\mathrm{tr}\left[\boldsymbol{d}^{-1}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{d}^{-1}\boldsymbol{u}\boldsymbol{u}'\right] \\ &= \frac{1}{4}\mu_4 V_{\mathrm{d}} + \left[\left(\delta_{\mathrm{d}}^2\left(1+\frac{1}{N}\right) + \frac{2p}{N}\right) + \frac{p}{N^2}\right]\left(V_{\mathrm{d}} + E_{\mathrm{d}}^2\right). \end{aligned} \tag{8}$$

Substitution of Eq. (7) and (8) into Eq. (6) results in

$$EP_N^{(\mathrm{PF})} \approx \Phi\left\{ -\frac{\delta_\mathrm{d}}{2} \frac{1}{\sqrt{\left(1+\frac{V_\mathrm{d}}{E_\mathrm{d}^2}\right)\left(1+\frac{1}{N}\left(1+\frac{2p}{\delta_\mathrm{d}^2}\right)+\frac{p}{\delta_\mathrm{d}^2 N^2}\right)+\frac{\mu_4}{4\delta_\mathrm{d}^2}\frac{V_\mathrm{d}}{E_\mathrm{d}^2}}} \right\}. \tag{9}$$

In order to apply this result to evaluate the expected error of the Fisher linear classifier with the pseudo-inverse covariance matrix we need to know theoretical values of the coefficient of variation $\gamma = \sqrt{V_\mathrm{d}}/E_\mathrm{d}$ of the random variable $1/d_j$. With an increase in the learning-set size $n$ from 1 to $p$, the smallest eigenvalues become underestimated: some of them become extremely small. Therefore, the coefficient of variation $\gamma$ increases with an increase in the learning-set size $n$ from 1 to $p$. Unfortunately, analytical expressions of $\gamma$ are very complex. Therefore, in the present paper, values for $\gamma$ when $\boldsymbol{S}$ has a Wishart $W(n,\boldsymbol{I})$ distribution have been found using numerical methods. In effect, we use table values obtained by means of statistical modelling of the Wishart density and its moments (Table 1).

Use of $\delta_\mathrm{d}^2$, $\mu_4$ and $\gamma = E\sqrt{V_\mathrm{d}}/E_\mathrm{d}$ in Eq. (9) leads to theoretical values of the generalization error

$$EP_N^{(\mathrm{PF})} \approx \Phi\left\{ -\frac{\delta}{2}\sqrt{\frac{N}{2p}} \frac{1}{\sqrt{(1+\gamma^2)\left(1+\frac{1}{N}+\frac{4p^2}{\delta^2}\frac{1}{N^2}+\frac{2p^2}{\delta^2}\frac{1}{N^3}\right)+\gamma^2\frac{3\delta^2}{8p^2}N}} \right\}, \tag{10}$$

In Eq. (10), the term

$$T_\mu = 1+\frac{1}{N}+\frac{4p^2}{\delta^2}\frac{1}{N^2}+\frac{2p^2}{\delta^2}\frac{1}{N^3}$$

arises from inexact sample estimation of the mean vectors of the classes, the term $T_r = \sqrt{N/(2p)}$ arises from the reduction in the number of features, and the terms $\gamma = \sqrt{V_\mathrm{d}}/E_\mathrm{d}$ and $T_\mathrm{eig} = 3\delta^2 N/(8p^2)$ arise from the inexact estimation of the eigenvalues $d_1, d_2, \ldots, d_r$.

Note that with an increase in the learning set size $n$ from 1 up to $p$, the terms $T_\mu$, and $T_r$ tend to decrease the generalization error – the number of the dimensions used for the classification increases. The terms $\gamma$, and $T_\mathrm{eig}$ tend to increase the generalization error – the smallest eigenvalues become underestimated. Consequently, large weighting factors $1/d_j$ in Eq. (5) increase, underestimating important directions. The smallest eigenvalues overestimate useless directions. Numerical calculations show an interesting and unexpected behaviour of the classification error: with an increase of the learning-set size $N$, the generalization error first decreases, reaches a minimum and afterwards begins to increase (see theoretical graphs 2 in Fig. 1a–b). The minimal error is obtained when $N = p/4$ $(n = p/2)$, and the maximal errors are obtained when $N = p/2$ $(n = p)$. For $n > p$, we

Table 1
Coefficient of variation $\gamma = \sqrt{V_\mathrm{d}}/E_\mathrm{d}$ of the inverse eigenvalues of the Wishart $W_{p,n}$ matrix

| $p$ | $n/p$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.00 |
| 30 | 0.38 | 0.54 | 0.71 | 0.89 | 1.12 | 1.40 | 1.74 | 2.41 | 4.15 | 54.8 |
| 50 | 0.37 | 0.53 | 0.70 | 0.86 | 1.09 | 1.31 | 1.66 | 2.35 | 3.70 | 37.7 |
| 100 | 0.36 | 0.52 | 0.69 | 0.84 | 1.02 | 1.25 | 1.60 | 2.20 | 3.15 | 11.5 |

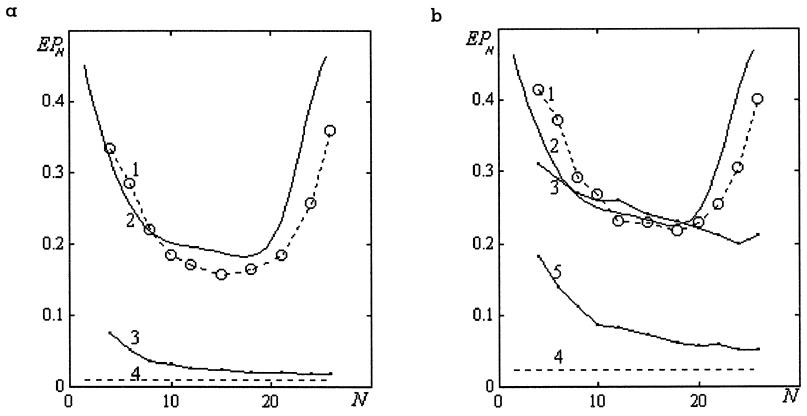


Fig. 1. Generalization error versus $N$, the number of learning samples. (a) Spherical uncorrelated data, $p = 50$, $P_\infty^{(F)} = 0.01$. (b) Gaussian correlated data, with common $\Sigma$, $p = 50$, $\rho = 0.45$, $P_\infty^{(F)} = 0.0238$. (1) pseudo-Fisher linear DF (average of 25 or 50 simulation experiments); (2) pseudo-Fisher DF (asymptotic Eq. (13)); (3) Euclidean distance classifier; (4) asymptotic error $P_\infty^{(F)}$; (5) regularized linear DA for optimal $\lambda$, the regularization parameter.

have the Fisher linear DF. Then the expected error regularly decreases with an increase in $N$ (Deev, 1970, 1972; Raudys, 1972; Wyman et al., 1990)

$$EP_N^{(F)} \approx \Phi\left\{ -\frac{\delta}{2}\frac{1}{\sqrt{T_\mu T_\Sigma}} \right\}, \tag{11}$$

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

## 3. Numerical analysis and discussion

To evaluate the accuracy of the asymptotic analytical formula derived to calculate the expected classification error of the Fisher classifier with the pseudo-inverse, we performed a number of simulation experiments with artificial Gaussian data. In a first series of experiments, we used two pattern classes of 50-variate spherical Gaussian data. While generating the data, the mean vectors of the separate features were generated by a Gaussian distribution and were normalised in such a way that the Mahalanobis distance $\delta = 4.65$ ($P_\infty^{(F)} = 0.01$). The pseudo-Fisher classifier was trained 50 times on 50 independent randomly chosen learning-sets composed of $N$ samples from each class. For each learning-set, the generalization error (a ''test-set error'') was calculated analytically

$$P_N = \frac{1}{2}\Phi\left\{ -\frac{w_0 + \boldsymbol{w}'\mu_1}{\sqrt{\boldsymbol{w}'\Sigma^{-1}\boldsymbol{w}}} \right\} + \frac{1}{2}\Phi\left\{ \frac{w_0 + \boldsymbol{w}'\mu_2}{\sqrt{\boldsymbol{w}'\Sigma^{-1}\boldsymbol{w}}} \right\}, \tag{12}$$

where $w_0$, and $\boldsymbol{w}$ are the weights of the linear classifier.

The average values of 50 experiments are presented in Fig. 1a.

The second series of experiments differs in the correlation between the features, which was chosen $\rho = 0.45$. Mean values from 25 independent experiments are presented in Fig. 1b. In order to demonstrate potential possibilities of opposing algorithms of the linear discriminant analysis in this experiment, we tested the regularized discriminant analysis too. Graph 5 in Fig. 1b corresponds to the generalization error of the linear regularized DA when we used the optimal value $\lambda_{\text{opt}}$ of the smoothing parameter for each particular

learning-set. To find $\lambda_{opt}$ we examined a number of values of $\lambda$, and on the basis of Eq. (12) (estimates) the best one was selected.

Simulation experiments indicate that in spite of the assumptions, our simple asymptotic formula (Eq. (10)) is comparatively accurate to describe the small sample behaviour of the pseudo-Fisher linear classifier. With an increase in the learning-set size $N$, the expected classification error first decreases. This is caused by the fact that the ''effective'' Mahalanobis distance $\delta_d^2 = \boldsymbol{m}'\boldsymbol{m} = r/\delta^2$ increases with an increase in $r = 2N - 2$, the number of directions used to perform the classification procedure.

When the total number of learning samples $n = 2N$ approaches the dimensionality, $p$, some of the eigenvalues of the sample covariance matrix become extremely large while the others become *extremely small*. A high variability of the sample estimates of the eigenvalues causes that some of the directions (essentially these are the *random* directions) become highly ''overestimated'' by an excessively large ''weighting factor'' $1/d_j$ in the new transformed space. The regularization of the covariance matrix $\boldsymbol{S}_R = \boldsymbol{S} + \lambda \boldsymbol{I}$ reduces this ''weighting factor'', i.e., makes it $1/(d_j + \lambda)$, and thus reduces the generalization error dramatically (curve 5 in Fig. 1b). Nevertheless, the influences of these directions (with zero eigenvalues) remain the highest. The pseudo-Fisher classifier plainly ignores the directions with zero eigenvalues.

Both, in the pseudo-Fisher classifier and the regularized linear discriminant analysis, some ''ad hoc'' heuristical procedures are utilised in order to reduce the negative influence of zero and/or smallest eigenvalues of the sample covariance matrix $\boldsymbol{S}$. The regularized DA can be explained mathematically by the Bayesian statistics when instead of the uniform prior distribution of $\Sigma$, some other type of prior distribution is chosen. There, the regularization constant $\lambda$ is a parameter of the prior distribution. Possibly, special choices of the prior distribution for the eigenvalues of the matrix $\Sigma$ can lead to more effective modifications of the classifier.

Our analysis shows that in standard pattern classification problems, the present version of *the pseudo-Fisher linear classifier is not a good choice in the very small learning-set case*. We exposed the factors that cause the multiple peaking behaviour of the learning curve: the pseudo-Fisher linear classifier is the ''diagonal'' Fisher linear classifier in the subspace of the principal components corresponding to non-zero eigenvalues of the sample covariance matrix $\boldsymbol{S}$. In spite of the fact that with an increase in the learning-set size $n$ from 1 up to $p$, the number of the principal directions increases, the random weighting factors $1/d_j$ in the ''diagonal'' classifier increase, underestimate important directions and overestimate worthless ones. The addition of the regularization parameter $\lambda$ to all eigenvalues $d_j$ of the covariance matrix in the regularized DA, diminishes this latter negative effect. Further analysis can lead to more modifications of the pseudo-Fisher classifier and the regularized linear DA, as well as unveil data models where a limited number of directions in the transformed feature space by the eigenvectors of the matrix $\boldsymbol{S}$ can be favourable to perform the classification.

## References


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

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis. Willey, New York.

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

Deev, A.D., 1972. Asymptotic expansions for distributions of statistics W, M, W∗ in discriminant analysis. Statistical Methods of Classification, vol. 31. Moscow University Press, Moscow, pp. 6–57 (in Russian).

Di Pillo, P.J., 1979. Biased discriminant analysis: evaluation of the optimum probability of misclassification. Commun. Stat. Theor. Meth. A 8 (14), 1447–1457.

Duin, R.P.W., 1978. On Accuracy of Statistical Pattern Recognizers, Duch Efficiency Bureau, Pijnacker.

Duin, R.P.W., 1995. Small sample size generalization. In: Borgefors, G. (Ed.), SCIA'95, Proc. 9th Scandinavian Conf. on Image Analysis, vol. 2. Uppsala, June 6–9, 1995, pp. 957–964.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Gupta, A.K., 1977. On the equivalence of two classification rules. Biometrical J. 19 (5), 365–367.

Jain, A., Chandrasekaran, B., 1982. Dimensionality and sample size considerations in pattern recognition practice. Handbook of Statistics, vol. 2. North-Holland, Amsterdam, pp. 835–855.

Kanal, L., Chandrasekaran, B., 1971. On dimensionality and sample size in statistical pattern classification. Pattern Recognition 3, 238–255.

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

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

Raudys, S., 1970. On the problems of sample size in pattern recognition. Proc. 2nd All-Union Conf. Statist. Methods in Control Theory, vol. 2. Nauka, Moscow, pp. 64–67 (in Russian).

Raudys, S., 1972. On the amount of a priori information in designing the classification algorithm. Proc. Acad. of Sciences of the USSR, Technical. Cybernetics, vol. 10. Nauka, Moscow, No. 4, pp. 168–174 (in Russian).

Raudys, S., 1998. Evolution and generalization of a single neurone. Part I. SLP as seven statistical classifiers. Neural Networks, in press.

Raudys, S., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans. Pattern Analysis Machine Intelligence 13, pp. 252–264.

Raudys, S., Skurichina, M., 1994. 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, Proc. 5th Tartu Conf., vol. 3. Tartu-Puhajarve, Estonia, 23–28 May 1994, pp. 237–245.

Schürmann, J., 1977. Polynomklassifikatoren für Zeichenerkennung, R. Oldenbourg Verlag, Munich.

Skurichina, M., Duin, R., 1996. Stabilizing classifiers for very small sample sizes, Proc. 13th Int. Conf. on Pattern Recognition, Vienna, August 25–29, 1996, vol. 2, track B: Pattern Recognition and Signal Analysis. IEEE Computer Society Press, Los Alamitos, 1996, pp. 891–896.

Wyman, F., Young, D., Turner, D., 1990. A comparison of asymptotic error rate expansions for the sample linear discriminant function. Pattern Recognition 23, 775–783.