Pattern Analysis & Applications (1999)2:44–52

Regularisation of Linear Classifiers by Adding Redundant Features*

Marina Skurichina and Robert P. W. Duin

Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, Delft, The Netherlands

Abstract: The Pseudo Fisher Linear Discriminant (PFLD) based on a pseudo-inverse technique shows a peaking behaviour of the generalisation error for training sample sizes that are about the feature size: with an increase in the training sample size, the generalisation error first decreases, reaching a minimum, then increases, reaching a maximum at the point where the training sample size is equal to the data dimensionality, and afterwards begins again to decrease. A number of ways exist to solve this problem. In this paper, it is shown that noise injection by adding redundant features to the data is similar to other regularisation techniques, and helps to improve the generalisation error of this classifier for critical training sample sizes.

Keywords: Critical sample size; Generalisation error; Noise injection; Peaking behaviour; Pseudo Fisher Linear Discriminant; Regularisation; Ridge estimate

1. INTRODUCTION

The main problem in estimating classifiers by small training sets is that they require the inverse of the covariance matrix, which is impossible to perform when the number of training objects N is less than the data dimensionality p. One of the ways in which to overcome the small sample size problem is to modify the standard classifiers in one way or another. However, even modified classifiers, such as the Pseudo-Fisher Linear Discriminant (PFLD) [1], may become very unstable, and have a peaking effect of the generalisation error when the training sample size is comparable with the data dimensionality [2–4].

In the past, the following ways have been studied to solve this problem:

1. Removing features (decreasing p) by some feature selection method.
2. Adding objects (increasing N), either by using larger training sets, or if this is not possible, by generating additional objects (noise injection [5]).
3. Removing objects (decreasing N) brings the classifier out of the instable region. This method has been studied by us [2,3], and is also effectively used in the Support Vector Classifier [6].

In this paper we will show by some examples that the fourth way can also be effective:

4. Adding redundant features (increasing p) [7]. Like the third method, this brings the classifier out of the instable region, but now by enlarging the dimensionality by noise.

We will concentrate on the injection of noise by adding redundant features to the data and its effect on the performance of the PFLD. The data used in our simulation study are presented in Section 2. The PFLD is discussed in Section 3. The use and performance of noise injection in the data feature space is considered in Section 4. The effect of adding redundant features on the PFLD, and its similarity to other regularisation techniques, are discussed in Section 5. Conclusions can be found in Section 6.

2. THE DATA

Two artificial data sets and one real data set are used for our experimental investigations. These data sets have a high dimension, because we are interested in critical situations where the PFLD has a bad performance.

The first set is a 30-dimensional correlated Gaussian data

Received: 10 November 1998
Received in revised form: 7 January 1999
Accepted: 7 January 1999
** Presented at SPR '98*

set constituted by two classes with equal covariance matrices. Each class consists of 500 vectors. The mean of the first class is zero for all features. The mean of the second class is equal to 3 for the first two features, and equal to 0 for all other features. The common covariance matrix is a diagonal matrix with a variance of 40 for the second feature and a unit variance for all other features. The intrinsic class overlap (Bayes error) is 0.064. This data set is rotated using a 30×30 rotation matrix which is $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ for the first two features and the identity matrix for all other features. We call these data further 'Gaussian correlated data'. Its first two features are presented in Fig. 1.

The second data set consists of two 30-dimensional Gaussian distributed data classes with unequal covariance matrices. Each data class contains 500 vectors. The first data class is distributed spherically with unit covariance matrix and with zero mean. The mean of the second class is equal to 4.5 for the first feature and equal to 0 for all other features. The covariance matrix of the second class is a diagonal matrix with a variance of 3 for the first two features and a unit variance for all other features. We call these data further 'Gaussian spherical data with unequal covariance matrices'. Its first two features are presented in Fig. 2.

The last data set consists of real data collected through spot counting in interphase cell nuclei (see, for instance, Netten et al [8] and Hoekstra et al [9]). Spot counting is a technique to detect numerical chromosome abnormalities. By counting the number of coloured chromosomes ('spots'), it is possible to detect whether the cell has an aberration that indicates a serious disease. A FISH (Fluorescence In Situ Hybridization) specimen of cell nuclei was scanned using a fluorescence microscope system, resulting in computer images of the single cell nuclei. From these single cell images, 16×16 pixel regions of interest were selected. These regions contain either background spots (noise), single spots or touching spots. From these regions, we constructed two classes of data: the noisy background and single spots, omitting the regions with touching spots. The samples of size 16×16 were considered as a feature vector of size 256. The first class of data (the noisy background) consists of 575 256-dimensional vectors, and the second class (single spots) – of 571 256-dimensional vectors. We call these data 'cell data' in the experiments.

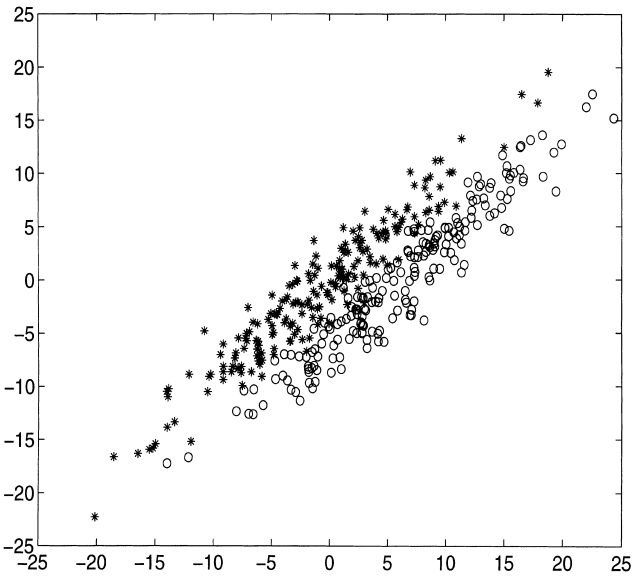

Fig. 1. Scatter plot of a two-dimensional projection of the 30-dimensional Gaussian correlated data.

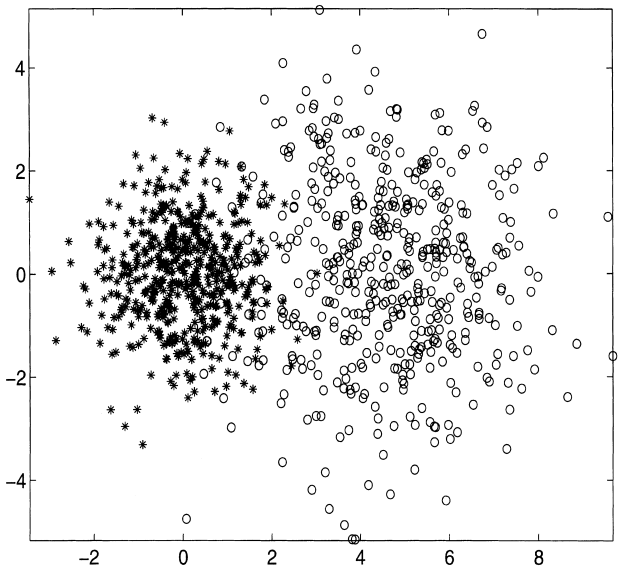

Fig. 2. Scatter plot of a two-dimensional projection of the 30-dimensional Gaussian spherical data with unequal covariance matrices.

Training data sets with 3–200 (with 3–300 for cell data) samples per class are chosen randomly from the total set. The remaining data are used for testing. These and all other experiments are repeated 10 times for independent training sample sets. In all figures, the averaged results over 10 repetitions are presented.

3. PSEUDO FISHER LINEAR DISCRIMINANT

The most popular and commonly used linear classifier is the Fisher Linear Discriminant (FLD) [10,11]

$$g_F(x) = [x - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)})]' \, S^{-1} \, (\overline{X}^{(1)} - \overline{X}^{(2)}) \qquad (1)$$

where S is the standard maximum likelihood estimation of the $p \times p$ common covariance matrix Σ, x is a p-variate vector to be classified and $\overline{X}^{(i)}$ is the sample mean vector of the ith class, $i = 1,2$.

Notice that Eq. (1) is the mean squared error solution for the linear coefficients (w, w_0) in

$$g_F(x) = w \bullet x + w_0 = L \qquad (2)$$

with $x \in X$ and with L being the corresponding desired outcomes, 1 for class 1 and -1 for class 2. When the number of data features p exceeds the total number of training vectors N, the estimate matrix S becomes singular and the direct inverse becomes impossible [12]. The expected probability of misclassification rises dramatically [13].

The modification of the FLD, which allows us to avoid the inverse of an ill-conditioned covariance matrix, is the so-called Pseudo Fisher Linear Discriminant [1]. In the PFLD a direct solution of Eq. (2) is obtained by (using augmented vectors):

$$g_{PF}(x) = (w, w_0) \bullet (x,1) = (x,1)(X,I)^{-1} L \qquad (3)$$

where $(x,1)$ is the augmented vector to be classified and (X,I) is the augmented training set. The inverse $(X,I)^{-1}$ is the Moore–Penrose Pseudo Inverse, which gives the minimum norm solution. Before the inversion, the data are shifted such that they have zero mean. This method is closely related to singular value decomposition.

For values $N \geq p$ the PFLD, maximising the distance to all given samples, is equivalent to the FLD (1). For values $N < p$, however, the Pseudo Fisher rule finds a linear subspace, which covers all the data samples. On this plane the PFLD estimates the data means and the covariance matrix, and builds a linear discriminant perpendicular to this subspace in all other directions for which no samples are given.

The behaviour of the PFLD as a function of the sample size is studied elsewhere [2,4]. For one sample per class this method is equivalent to the Nearest Mean and to the Nearest Neighbour methods. If the total sample size is equal to or larger than the dimensionality, $N \geq p$, the method is equivalent to the FLD. In between, the generalisation error shows a minimum and a maximum at the point $N = p$ (see Fig. 3). This can be understood from the observation that the PFLD succeeds in finding hyperplanes with equal distances to all training samples until $N = p$. In Raudys and Duin [14], an asymptotic expression for the generalisation error of the PFLD is derived, which explains theoretically the behaviour of the PFLD.

4. PERFORMANCE OF NOISE INJECTION BY ADDING REDUNDANT FEATURES

To improve the generalisation error of the PFLD for critical values of the training sample size ($N = p$), a number of techniques could be used (see the Introduction). One of the ways to solve this problem involves generating more training objects by noise injection into the training data. Usually, spherical Gaussian distributed noise is generated around each training object. However, this method requires us to know quite precisely the optimal variance of the noise in order to get good results. The optimal value of the noise variance depends upon many factors such as the training sample size, the data dimensionality and the data distribution [5]. It could vary dramatically for different data. As a rule it is computationally expensive to find the optimal value of the noise variance.

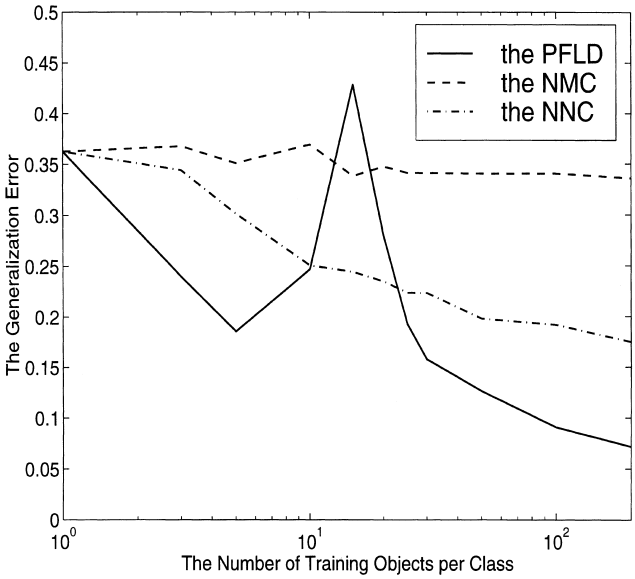

Fig. 3. Generalisation error of the Pseudo Fisher Linear Discriminant (PFLD), the Nearest Mean Classifier (NMC) and the Nearest Neighbour Classifier (NNC) versus the training sample size for 30-dimensional Gaussian correlated data.

To demonstrate the influence of the noise variance λ^2 on the generalisation error of the PFLD, we considered the 30-dimensional Gaussian correlated data. The averaged results for some values of λ^2 are presented in Fig. 4. We can see that the performance of the PFLD strongly depends upon the variance of the noise.

Considering the small sample size properties (a learning

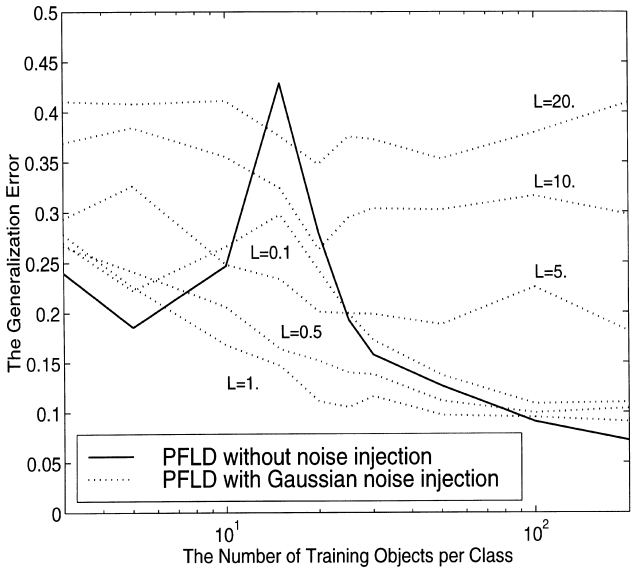

Fig. 4. Generalisation error of the PFLD with and without noise injection to the training objects with different values of the noise variance $\lambda^2 = L$ versus the training sample size for 30-dimensional Gaussian correlated data.

curve) of the PFLD, one can reach another solution: decrease the number of training objects in order to avoid the critical training sample size problem. It could be also performed by noise injection into the data feature space, instead of adding noise to the training objects. In this case, the data dimensionality is enlarged by adding Gaussian distributed features $N(0, \lambda^2)$. When increasing the data dimensionality p, the training sample size N relatively decreases, leaving a critical area $N = p$, where the PFLD has a high high generalisation error. For values $N < p$ the PFLD performs much better than for the critical sizes of the training set.

Let us now investigate the effectiveness of noise injection by adding redundant features for the three examples of the data described in Section 2. To study the influence of the injection of 'noisy' features to the data, for each considered data, r additional redundant 'noisy' features were generated having Gaussian distributions with zero mean and unit variance $N(0, 1)$ for both classes. This enlarges the data dimensionality from p to $p + r$. The generalisation error of the PFLD for 30-dimensional Gaussian correlated data and 30-dimensional Gaussian spherical data with unequal covariance matrices without noise injection in the feature space and with 20, 70 and 170 additional redundant 'noisy' features is presented in Figs 5 and 6, respectively. The generalisation error of the PFLD obtained on the cell data without noise injection in the feature space and on the cell data with 44, 100, 144 and 200 redundant features is presented in Fig. 7.

For all data the PFLD shows a critical behaviour with a high maximum of the generalisation error around the critical training sample size $N = p$. Figures 5–7 demonstrate nicely that noise injection into the data feature space helps to avoid the peaking effect of the generalisation error of the

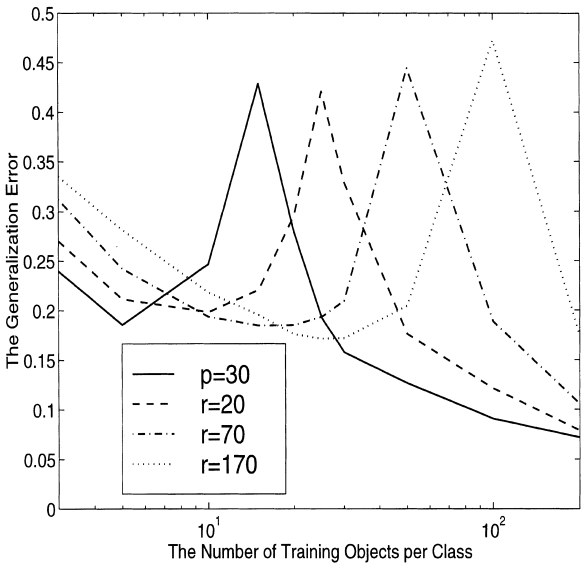

Fig. 5. Generalisation error of the PFLD versus the training sample size for Gaussian correlated data without noise injection in the feature space ($p = 30$) and with 20, 70 and 170 additional redundant features ($p = 50$, 100, 200).

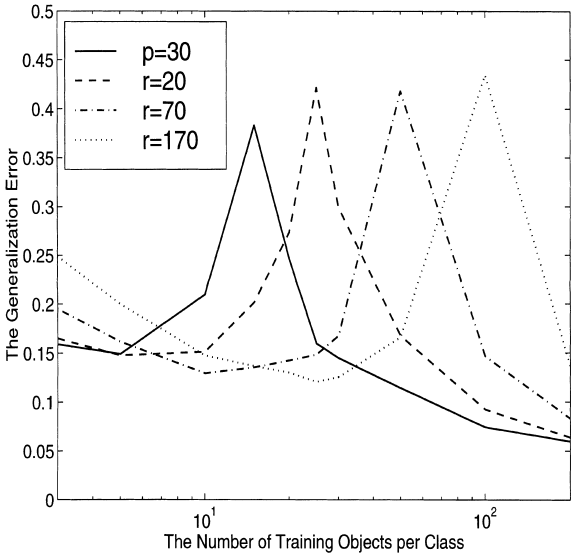

Fig. 6. Generalisation error of the PFLD versus the training sample size for Gaussian spherical data with unequal covariance matrices without noise injection in the feature space ($p = 30$) and with 20, 70 and 170 additional redundant 'noisy' features ($p = 50$, 100, 200).

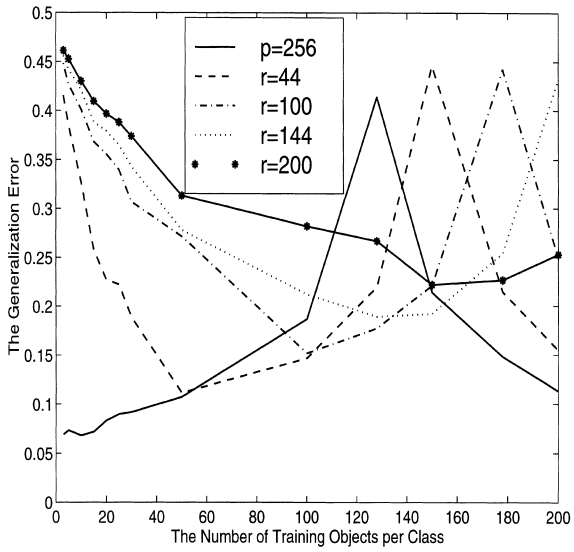

Fig. 7. Generalisation error of the PFLD versus the training sample size for 256-dimensional cell data without noise injection ($p = 256$) and with 44, 100, 144 and 200 additional redundant 'noisy' features ($p = 300$, 356, 400, 456).

PFLD for a given number of training objects. We can see that the redoubling of the data dimensionality by adding 'noisy' features has already doubled the performance of the classifier at the point $N = p$. For cell data, it was enough to add 44–100 'noisy' features for the same improvement. When the number of added 'noisy' features was 4–5 times larger than the original dimensionality of the data, the peak of the generalisation error was smoothed almost completely:

the generalisation error was reduced in a whole region around the critical training sample size. However, for very small training sample sets, adding redundant features was useless. The reason could be following. Adding noise with a quite large variance ($\lambda^2 = 1$) to a highly dimensional feature space with only a few objects makes the training data set too 'noisy' to represent the entire data set correctly. In this case, it becomes difficult or even impossible to build a good discriminant function. All the data considered demonstrate nicely that the more noise that is added to the data by adding redundant features, the larger the generalisation error obtained in the case of very small training sample sizes. A smaller noise variance should probably be used to get better results for small training set sizes. For critical training data sizes, adding redundant features helps to avoid the peaking effect of the generalisation error of the PFLD.

One can notice that the improvement obtained in the generalisation error depends upon the number r of redundant features used. It is also reasonable to suppose that the generalisation error depends upon the noise variance in redundant features. A mathematical attempt to understand this relation is made in next section. Figures 5–7 suggest that the relationship between the number of redundant features r and the noise variance in redundant features λ^2 depends upon the training sample size, and may also depend upon the intrinsic data dimensionality. Obviously, this question requires a more careful investigation in future. Nevertheless, our simulation study completely proves the possible usefulness of noise injection in the data feature space in order to reduce the generalisation error of the PFLD for critical training sample sizes.

5. REGULARISATION BY ADDING REDUNDANT FEATURES IN THE PFLD

In this section we make two attempts to understand how the addition of redundant features to the data affects the performance of the PFLD. It is well known that when the data dimensionality p is larger than the number of training objects N, the PFLD constructs the linear discriminant in the linear subspace, in which the training data are located, and perpendicularly to all other dimensions where no training data are presented. Therefore, first we try to understand what happens with the training data set in the linear subspace found by the PFLD when adding redundant features to the data. On the other hand, considering that adding redundant features is actually noise injection in the feature space, it is logical to suppose that it should be similar to other regularisation techniques. In Section 5.2 we try to show this by some mathematical analysis and a simulation study.

5.1. The Inter-Data Dependency by Redundant Features

Let $\{x_1, x_2, \ldots, x_N\}$ be the original training data set of N objects x_i, $i = \overline{1,N}$, where each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is a p-dimensional vector, and $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$ is the mean of this training data set. Suppose that r redundant features y_{ij}, $j = \overline{1, r}$, having a normal distribution $N(0, \lambda^2)$, are added to each training vector x_i. These features y_{ij} constitute an r-dimensional vector of redundant features $y_i = (y_{i1}, y_{i2}, \ldots, y_{ir})'$ for each training vector x_i, resulting in a set of $(p + r)$-dimensional vectors $\{z_1, z_2, \ldots, z_N\}$ with $z_i = (x_i\ y_i)'$, $i = \overline{1,N}$. Now $X = (x_1\ x_2\ \ldots\ x_N)$, $Y = (y_1\ y_2\ \ldots\ y_N)$ and $Z = (z_1\ z_2\ \ldots\ z_N)$ are data matrices with sizes of $p \times N$, $r \times N$ and $(p + r) \times N$.

Adding redundant features provides additional information, and transforms the data in multidimensional space. To see how the data are transformed, we will consider the covariance matrix $G(Z) = Cov(Z',Z) = E\{(Z - M)'(Z - M)\}$ of the extended data, and compare it with the covariance matrix $G(X) = Cov(X',X) = E\{(X - \mu)'(X - \mu)\}$ of the original data X. $G(X)$ will be called the *data dependency matrix*, as it shows the inter-dependency between the data objects (multidimensional vectors). They counterbalance the usual *covariance matrix* $Cov(X,X')$, which shows the dependency between data features.

The inner product $(Z - M)'(Z - M)$ can be written as a sum of two inner products

$$(Z - M)'(Z - M) = ((X - \mu)'\ Y') \begin{pmatrix} X - \mu \\ Y \end{pmatrix} = (X - \mu)'(X - \mu) + Y'Y$$

Since X and Y are statistically independent, the data dependency matrix of Z can be expressed as follows:

$$G(Z) = Cov(Z',Z) = E\{(Z - M)'(Z - M)\} = E\{(X - \mu)'(X - \mu) + Y'Y\} = G(X) + G(Y)$$

Considering that components y_{ij} of the matrix Y are statistically independent variables with normal distribution $N(0, \lambda^2)$, the matrix $G(Y)$ is a diagonal matrix with diagonal elements having $\lambda^4\chi_r^2$ distribution. Therefore, $G(Y) = Cov(Y',Y) = E(Y'Y) = 2\lambda^4 rI$, where I is the $r \times r$ identity matrix. That gives us

$$G(Z) = G(X) + 2\lambda^4 rI \tag{4}$$

Formula (4) shows that the training data set is actually decorrelated by adding redundant features to the data as variances of $G(Z)$ become larger. The distance between training objects increases and tends to be equal.

When the training samples size N is smaller than the data dimensionality p, the Pseudo Fisher Linear Discriminant finds the linear subspace of dimensionality $N - 1$, which covers all training samples, estimates the data distribution parameters there and builds a discriminant function in this linear subspace. If the original data are enlarged by redundant features, noise is added in the feature space. According to formula (4), by adding noisy features, the distances between training objects increase and the data are decorrelated. The larger the variance λ^2, and the more noise in the feature space that is added, then the more the data decorrelated. For very large values of λ^2, however, the information in the original data is lost. The classifier trained on such data is thus bad.

Formula (4) also shows that the speed of the data decorrelation depends thus upon two parameters: the number of redundant features r; and the noise variance λ^2. When r is small and λ^2 is large, the decorrelation of data objects goes faster than with the large r and the small λ^2. To illustrate the dependence of the generalisation error of the PFLD on the number of redundant features r and the variance of the noise λ^2, we considered the 30-dimensional Gaussian correlated data with the critical training sample size $N = 15 + 15 = p$. The averaged results are presented in Figs 8 and 9, and show that adding redundant features to the data affects the generalisation error of the PFLD. One can see clearly that the influence of the noise variance λ^2 on the generalisation error of the PFLD is stronger than the influence of the number of redundant features r. Figures 8 and 9 also show that for each value of the noise variance, an optimal value of the number of redundant features exists, and *vice versa*. Obviously, more study is required in this direction. However, it seems that to get a smaller generalisation error by adding redundant features, it is preferable to add many redundant features with a small variance than a few redundant features with a large variance of noise.

5.2. Regularisation by Noise Injection

To understand better why adding redundant features can improve the performance of the PFLD, let us consider the sample covariance matrix and its decomposition used in the PFLD. As mentioned before, in the PFLD the pseudo inverse of the sample covariance matrix S is used. A sense of the pseudo-inverse consists in a singular value decomposition of S: $TST' = D$, where T is an orthogonal matrix. Then the pseudo inverse of matrix S

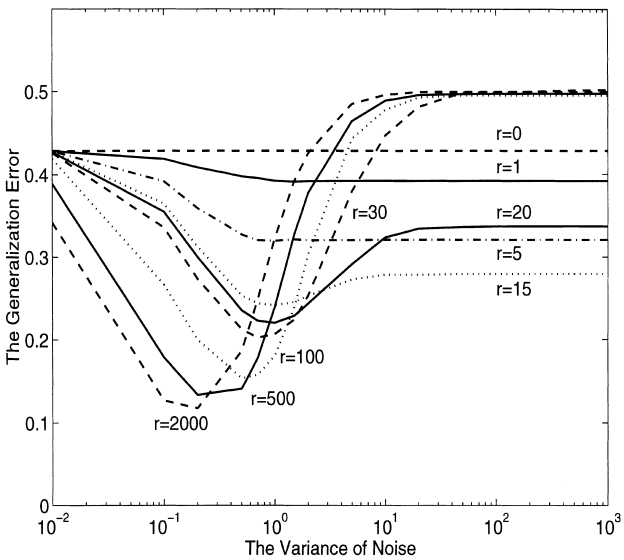

Fig. 8. Generalisation error of the PFLD versus the noise variance $\lambda^2 = L$ for different numbers of redundant features r for 30-dimensional Gaussian correlated data with a training sample size $N = 15 + 15 = p$.

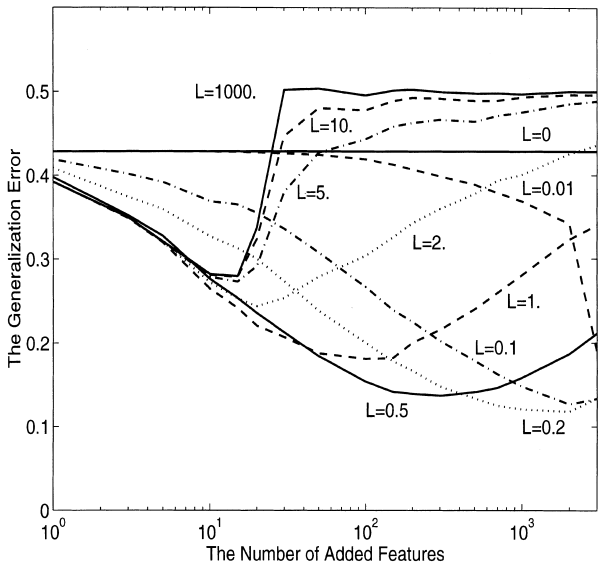

Fig. 9. Generalisation error of the PFLD versus the number of redundant features r for different values of the noise variance $\lambda^2 = L$ for 30-dimensional Gaussian correlated data with a training sample size $N = 15 + 15 = p$.

$$S^{-1} = T'D^{-1}T$$

is used instead of the direct inverse of S.

We now consider what happens with the sample covariance matrix S in the PFLD when one adds redundant features to the data.

Let us keep the same definitions as above, and let $S = Cov(X,X')$ be the $p \times p$ sample covariance matrix of the training data set X. Since X and Y are statistically independent, that gives the $(p + r) \times (p + r)$ covariance matrix of Z

$$C = \begin{bmatrix} S & 0 \\ 0 & \lambda^2 I \end{bmatrix}$$

Let

$$T = \begin{bmatrix} T_1 & T_2 \\ T'_2 & T_3 \end{bmatrix}$$

be the $(p + r) \times (p + r)$ orthogonal transformation matrix, which diagonalises the covariance matrix C: $TCT' = D$. Notice that T_1 and T_3 are $p \times p$ and $r \times r$ symmetrical matrices, respectively. The matrix T_2 is a $p \times r$ matrix, and is not symmetrical.

As T is an orthogonal matrix satisfying the condition

$$TT' = \begin{bmatrix} T_1T'_1 + T_2T'_2 & T_1T_2 + T_2T'_3 \\ T'_2T'_1 + T_3T'_2 & T'_2T_2 + T_3T'_3 \end{bmatrix} = I$$

the following equations should hold:

$$T_1T'_1 + T_2T'_2 = I$$
$$T'_2T_2 + T_3T'_3 = I$$

That leads to the expressions

$$T_2 T'_2 = I - T_1 T'_1 \tag{5}$$

$$T_3 T'_3 = I - T'_2 T_2 \tag{6}$$

Therefore

$$D = TCT' = \begin{bmatrix} T_1 & T_2 \\ T'_2 & T_3 \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \lambda^2 I \end{bmatrix} \begin{bmatrix} T_1 & T_2 \\ T'_2 & T_3 \end{bmatrix},$$

$$= \begin{bmatrix} T_1 S T'_1 + T_2 \lambda^2 I T'_2 & T_1 S T_2 + T_2 \lambda^2 I T'_3 \\ T'_2 S T'_1 + T'_3 \lambda^2 I T'_2 & T'_2 S T_2 + T_3 \lambda^2 I T'_3 \end{bmatrix}$$

As D is a diagonal matrix, it should be

$$D = \begin{bmatrix} T_1 S T'_1 + \lambda^2 T_2 T'_2 & 0 \\ 0 & T'_2 S T_2 + \lambda^2 T_3 T'_3 \end{bmatrix}$$

Substituting Eqs (5) and (6) into the equation above yields

$$D = \begin{bmatrix} D_1 + \lambda^2 I & 0 \\ 0 & D_2 + \lambda^2 I \end{bmatrix} \tag{7}$$

where $D_1 = T_1(S - \lambda^2 I)T'_1$ and $D_2 = T'_2(S - \lambda^2 I)T_2$.

Let us now consider the ridge estimate of the $p \times p$ sample covariance matrix $S^* = S + \lambda^2 I$. It is known [5] that regularisation by the ridge estimate of the covariance matrix is equivalent to Gaussian noise injection to the training objects in the FLD. Let $\tilde{T}$ be the $p \times p$ orthogonal matrix which diagonalises the sample covariance matrix S. Then by applying the orthogonal transformation to the ridge estimate S^*, we obtain

$$D^* = \tilde{T} S^* \tilde{T}' = \tilde{T}(S + \lambda^2 I)\tilde{T}'$$
$$= \tilde{T} S \tilde{T}' + \tilde{T} \lambda^2 I \tilde{T}' = \tilde{D} + \lambda^2 I$$

where $\tilde{D} = \tilde{T} S \tilde{T}'$. Thus, the ridge estimate of the sample covariance matrix S is also presented in the diagonal matrix

$$D^* = \tilde{D} + \lambda^2 I \tag{8}$$

obtained by singular value decomposition in the PFLD.

Comparing Eqs (7) and (8), one can see that adding redundant features to the data in the PFLD is similar (but not equivalent) to the ridge estimate of the sample covariance matrix S. To illustrate the similarity of adding redundant features to regularisation techniques, such as noise injection to training objects and ridge estimate of the covariance matrix, we considered all the data described in Section 2 with the critical training sample size. A hundred redundant features with different values of noise variance (but the same for each redundant feature) were added to each data. The averaged results for the generalisation error of the PFLD without regularisation, with Gaussian noise injection to the training objects, with ridge estimate of the covariance matrix and when adding redundant features, are presented in Figs 10–12 for 30-dimensional Gaussian correlated data, 30-dimensional Gaussian spherical data with unequal covariance matrices and 256-dimensional cell data, respectively. The results obtained for all data sets are similar. One should take into account that cell data have the large dimensionality. Therefore, the optimal values of regularisation parameters for different types of regularisation differ more for this data set than for other two data sets. For the same reason, for the cell data the generalisation error of the PFLD with regularisation and when adding redundant features increases more slowly with an increase in the value of the noise variance than for other data sets. Nevertheless, in all of the figures, one can see that the generalisation error of the PFLD with Gaussian noise injection to the training objects and the generalisation error of the PFLD with 100 redundant features added to the data behave similarly. The generalisation error of the PFLD with ridge estimate of the sample covariance matrix also behaves in a similar way for small values of the regularisation parameter, and different for large values. However, the simulation study performed demonstrates nicely that adding redundant features to the data is similar to other regularisation techniques.

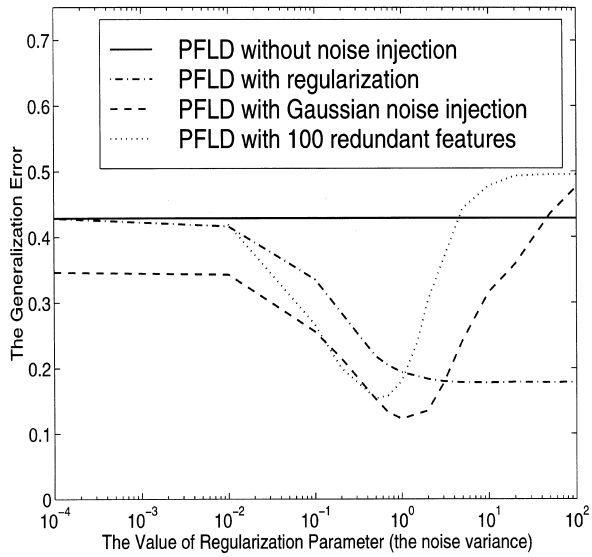

Fig. 10. Generalisation error of the PFLD with different types of regularisation *versus* the value of the regularisation parameter (the noise variance $\lambda^2 = L$) for 30-dimensional Gaussian correlated data with a training sample size $N = 15 + 15 = p$.

6. CONCLUSIONS

The PFLD may have a peaking behaviour of the generalisation error for training sample sizes that are about the feature size. Based on the small sample size properties of the PFLD, in this paper it has been suggested that injecting noise into the data feature space improves the generalisation error of the PFLD for critical training sample sizes. This approach was studied for two artificial data sets and one example of real data. Simulation results have shown that adding redundant 'noisy' features to the data allows us to dramatically reduce the generalisation error of the PFLD in the region of critical training sample sizes.

Mathematical analysis and simulation studies have shown that adding noise by redundant features is similar to other regularisation techniques, such as Gaussian noise injection to the training data and ridge estimate of the covariance matrix. It was noticed that there exists an optimal relation-

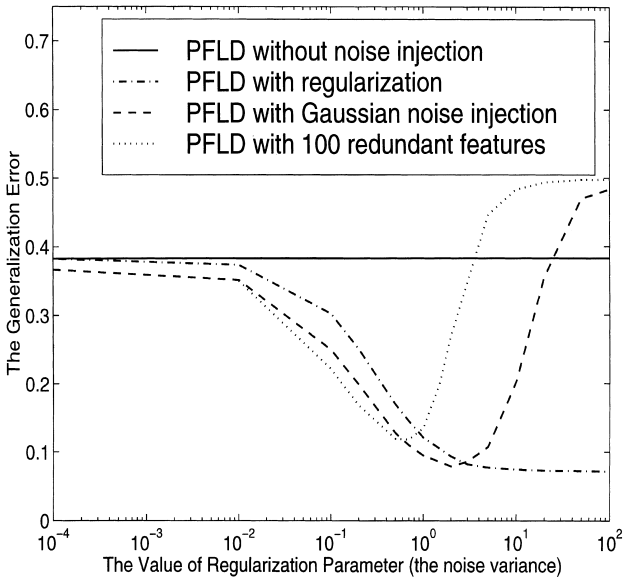

Fig. 11. Generalisation error of the PFLD with different types of regularisation *versus* the value of the regularisation parameter (the noise variance $\lambda^2 = L$) for 30-dimensional Gaussian spherical data with unequal covariance matrices and with a training sample size $N = 15 + 15 = p$.

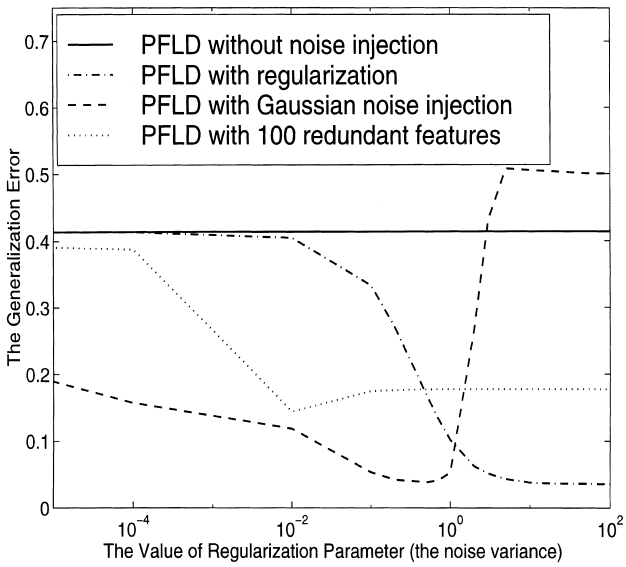

Fig. 12. Generalisation error of the PFLD with different types of regularisation *versus* the value of the regularisation parameter (the noise variance $\lambda^2 = L$) for 256-dimensional cell data with a training sample size $N = 128 + 128 = p$.

ship between the number of redundant features and the noise variance. This optimal relation might depend upon the size of the training data set and the intrinsic data dimensionality. However, it still needs more investigation to find an explicit expression.

Finally, let us note that some non-linear classifiers, (e.g. the quadratic classifier) may also have a peaking behaviour of the generalisation error in the region of critical training sample sizes. Therefore, it could be expected that adding redundant features could help to improve the performance of such classifiers constructed on critical training sample sizes.

Acknowledgements

This work was supported by the Foundation for Applied Sciences (STW) and the Dutch Organization for Scientific Research (NWO).

References

1. Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press, 1990, pp 400–407
2. Duin RPW. Small sample size generalization. Proceedings of 9th Scandinavian Conference on Image Analysis, vol 2, Uppsala, Sweden, 1995, pp 957–964
3. Skurichina M, Duin RPW. Stabilizing classifiers for very small sample sizes. Proceedings of ICPR, Vienna, Austria, 1996, pp 891–896
4. Skurichina M, Duin RPW. Bagging for linear classifiers. Pattern Recognition 1998;31(7):909–930
5. Raudys Š, Skurichina M, Cibas T, Gallinari P. Optimal regularization of neural networks and ridge estimates of the covariance matrix in statistical classification. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications (Int Journal of Russian Academy of Sciences) 1995;5(4):633–650
6. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273–297
7. Skurichina M, Duin RPW. Regularization by adding redundant features. Advances in Pattern Recognition. Proceedings of Joint International Workshops SSPR'98 and SPR'98, Sydney, Australia, 1998, pp 564–572
8. Netten H, Young IT, Prins M, van Vliet LJ, Tanke HJ, Vrolijk J, Sloos W. Automation of fluorescent dot counting in cell nuclei. Proceedings of the 12th Int. Conference on Pattern Recognition, Vol 1, Jerusalem, 1994, pp 84–87
9. Hoekstra A, Netten H, de Ridder D. A neural network applied to spot counting. Proceedings of ACSI'96, the Second Annual Conference of the Advanced School for Computing and Imaging, Lommel, Belgium, 1996, pp 224–229
10. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936;7(2):179–188
11. Fisher RA. The precision of discriminant functions. Annals of Eugenics 1940;10(4)
12. Rao R. On some problems arising of discrimination with multiple characters. Sankya 1949;9:343–365
13. Raudys Š, Pikelis V. On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence 1980;2(3):242–252
14. Raudys Š, Duin RPW. On expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recognition Letters 1998;19(5–6):385–392

Robert P. W. Duin studied applied physics at Delft University of Technology in the Netherlands. In 1978 he received a PhD degree for a thesis on the accuracy of statistical pattern recognisers. In his research he included various aspects of the automatic interpretation of measurements, learning systems and

classifiers. Between 1980 and 1990 he developed and studied hardware architectures and software configurations for interactive image analysis. At present he is an Associate Professor of the Faculty of Applied Sciences of Delft University of Technology. His main research interest is in the design and evaluation of learning algorithms for pattern recognition applications. This includes in particular neural network classifiers, support vector classifiers and classifier combining strategies.

Marina Skurichina studied applied mathematics and graduated from Vilnius University in 1989. From 1989 to 1996 she worked as a research fellow, and later as a PhD student, in the Department of Data Analysis at the Institute of Mathematics and Informatics in Vilnius, Lithuania. Now she is working on her PhD thesis in the Pattern Recognition Group of the Faculty of Applied Sciences of Delft University of Technology, in the Netherlands. She is the author of about 10 scientific papers. Her scientific interests include artificial neural networks in pattern recognition, the training of artificial neural networks, noise injection and regularisation methods.

Correspondence and offprint requests to: M. Sturichina, Pattern Recognition Group, Department of Applied Physics, Faculty of Applied Sciences, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands. Email: duin@ph.tn.tudelft.nl