Pergamon

Pattern Recognition, Vol. 27, No. 8, pp. 1065–1077, 1994
Elsevier Science Ltd
Copyright © 1994 Pattern Recognition Society
Printed in Great Britain. All rights reserved
0031-3203/94 $7.00 + .00

0031-3203(93)E0033-4

COMPARATIVE ANALYSIS OF STATISTICAL PATTERN RECOGNITION METHODS IN HIGH DIMENSIONAL SETTINGS

STEFAN AEBERHARD,† DANNY COOMANS‡ and OLIVIER DE VEL†§

† Department of Computer Science, James Cook University, 4811 Australia
‡ Department of Mathematics and Statistics, James Cook University, 4811 Australia

(*Received* 8 *April* 1993; *in revised form* 7 *December* 1993; *received for publication* 20 *December* 1993)

Abstract—An extensive simulation study is reported comparing eight statistical classification methods, focusing on problems where the number of observations is less than the number of variables. Using a wide range of artificial and real data sets, two types of classifiers are contrasted; methods that classify using all variables, and methods that first reduce the number of dimensions to two or three. The simulations identified regularized discriminant analysis as the overall clearly most powerful classifier, and show that in most cases, a reduction of the dimensionality to two or three dimensions prior to classification increases the error in allocating test observations.

Discriminant analysis   High dimensionality   Classifier evaluation   Simulation
Dimensionality   Reduction

INTRODUCTION

Classification is the assignment of objects to one of K a priori defined classes based on a set of d variables. Each object is characterized by d measurements obtained for the d variables, usually represented by the measurement vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$. If each variable corresponds to an axis in a metric space, the observations $\mathbf{x}_i$ are points in d-dimensional *feature space*. Different class populations ideally occupy different regions in the feature space, allowing classification methods to allocate test observations based on their location in the space. However, the regions corresponding to two or more classes may overlap, in which case 100% correct classification is no longer possible.

In practice, the distributions of the class populations are unknown and have to be estimated from exemplar objects available from each class, called *training samples* or *training objects*. Based on how the training samples are used for the purpose of classification, we may distinguish two different types of classifiers. *Parametric methods* model the classes based on assumptions relating to the underlying class distributions, and the training samples are used to estimate the parameters in these models. Examples are quadratic and linear discriminant analysis. *Nonparametric methods* make no such assumptions, and classify a test object based on the training samples in the neighbourhood of the object in the feature space. Examples are the K-nearest neighbour and potential methods.[1]

The number of variables measured on the samples is called the *dimensionality d* of a classification problem. One would expect that an increase in the dimensionality of a particular problem would lead to increased classification performance as additional information becomes available. Contrary to this expectation, it has been found in practice that beyond a certain point, the addition of further variables leads to a decrease in the classification performance.[2,3]

The reasons for what seems at first a paradox are different for parametric and nonparametric methods. With parametric methods, the number of parameters in the models increases with the dimensionality d. Unless the number of training objects N_k is increased correspondingly, the increase in the dimensionality will lead to *poorly-posed settings*, i.e. the number of training samples N_k in each class is comparable to d. In poorly-posed settings, parameter estimates have very large variances.[3] Even worse, in *ill-posed settings*, when the number of N_k of training samples is less than the dimensionality, not all parameters can be estimated.

Nonparametric methods on the other hand rely on densely populated feature spaces for reliable classification. In high-dimensional settings, this can only be achieved with very large numbers of training samples,[4] hence the classification accuracy is decreased.

New technology in the form of improved instrumentation has resulted in classification problems which are often ill- or poorly-posed, and traditional classifiers with good performance in well-posed settings deteriorate. Two possible solutions to this dilemma are dimensionality reduction and regularization. Dimensionality reduction may be performed by *feature selection* or by transforming the full feature space into a

§ Author to whom all correspondence should be addressed (olivier @ cs. jcu. edu. au).

lower dimensional space, here denoted as the *reduced space*. Regularization is the procedure of biasing parameters towards what are thought to be more plausible values. This reduces the variance of the estimates at the cost of introducing additional bias.

DESCRIPTION OF THE CLASSIFIERS

Classification in the full feature space

Quadratic Discriminant Analysis (QDA)

Among the rules used to assign objects to one of several classes, the *Bayes minimum error rule* is theoretically optimal, in that a test object $\mathbf{x}_i$ is classified into the class with the largest posterior probability. Applying the Bayes theorem, the rule may be formulated as:

Assign object $\mathbf{x}_i$ to class ω_k if

$$p(\mathbf{x}_i|\omega_k)P(\omega_k) > p(\mathbf{x}_i|\omega_j)P(\omega_j), \quad \text{for all } j \neq k. \tag{1}$$

Here, $p(\mathbf{x}_i|\omega_k)$ are the class probability densities, and $P(\omega_k)$ is the a priori probability of class ω_k.

The densities $p(\mathbf{x}_i|\omega_k)$ in equation (1) are usually unknown and have to be estimated from the training samples. Quadratic Discriminant Analysis (QDA) assumes that the distribution of the data is multivariate normal. Hence substituting the equation for the multivariate normal distribution and taking the logarithm, the Bayes minimum error rule (equation (1)) leads to the following classification score $c_k(\mathbf{x}_i)$ for QDA:

$$c_k(\mathbf{x}_i) = (\mathbf{x}_i - \mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}_i - \mu_k) + \ln|\Sigma_k| - 2\ln(P(\omega_k)). \tag{2}$$

Σ_k is the population covariance matrix of class ω_k and μ_k is the corresponding class mean. QDA classifies a test object into the class that minimizes c_k. In practice, the population parameters are usually unknown and are replaced by the sample estimates $\hat{\Sigma}_k$ and $\hat{\mu}_k$

$$\hat{\Sigma}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}(\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^{\mathrm{T}}, \quad \hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{x}_i. \tag{3}$$

Here, N_k is the number of training samples in class k. QDA, however, has one major drawback. In order to obtain reliable estimates $\hat{\Sigma}_k$ of the class covariance matrices, very large numbers of training samples are required for large dimensionalities.[3] Further, $\hat{\Sigma}_k$ cannot be inverted in ill-posed cases. For the simulations reported here, singular value perturbation[5] was applied to the covariance matrices to enable inversion in such situations.

Linear Discriminant Analysis (LDA)

In the same way as QDA, LDA also assumes that the class populations follow a multivariate normal distribution. However, LDA makes the extra assumption that the classes have identical covariance matrices

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K. \tag{4}$$

The first consequence of this assumption is that the quadratic term in the classification score (equation (2)) is the same for all classes and can therefore be omitted. Consequently, the resulting boundaries between the classes are linear, hence the name Linear Discriminant Analysis. Secondly, the *pooled* covariance matrix $\hat{\Sigma}$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{K} N_k\hat{\Sigma}_k, \quad N = \sum_{k=1}^{K} N_k \tag{5}$$

is used instead of the class specific covariance matrices. N is the total number of training samples. Performing these changes, the classification score for LDA becomes

$$c_k(\mathbf{x}_i) = -2\mathbf{x}_i^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mu}_k + \hat{\mu}_k^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mu}_k - 2\ln(P(\omega_k)). \tag{6}$$

Since fewer parameters need to be estimated, LDA has often been preferred over QDA in high dimensional settings. However, because of its linear boundaries and because these are at half the distance between the class centroids, LDA has severe limitations in cases where the class covariance matrices differ significantly.

Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis (RDA)[6,7] was especially developed for ill- and poorly-posed problems and can be seen as an adaption of QDA. As QDA, RDA classifies a test object $\mathbf{x}_i$ into the class that minimizes the classification score given by equation (2). However, while QDA uses an unbiased estimate of Σ_k, RDA makes use of a regularized covariance matrix $\hat{\Sigma}_k(\lambda, \gamma)$ instead. By regularizing the class covariance matrices in two separate steps towards physically more plausible values, RDA tries to overcome the problems of QDA due to the large variance in the unbiased estimates of the covariance matrices.

With limited training samples, the unbiased estimates of the class covariance matrices may not adequately reflect the similarities that may exist between the class structures. In these cases, a linear combination of the class covariance matrices and the pooled covariance matrix will yield a better estimate of the underlying class structures and hence lead to better classification results. This is addressed in a first regularization step, where $\hat{\Sigma}_k$ is replaced by a linear combination $\hat{\Sigma}_k(\lambda)$ of the sample class covariance matrix $\hat{\Sigma}_k$ and the pooled covariance matrix $\hat{\Sigma}$:

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)\hat{\mathbf{Q}}_k + \lambda\hat{\mathbf{Q}}}{N_k(\lambda)}, \quad 0 \le \lambda \le 1 \tag{7}$$

with

$$N_k(\lambda) = (1-\lambda)N_k + \lambda N \tag{8}$$

and

$$\hat{\mathbf{Q}} = \sum_{k=1}^{K}\hat{\mathbf{Q}}_k, \quad \hat{\mathbf{Q}}_k = N_k\hat{\Sigma}_k \tag{9}$$

where $\hat{\Sigma}_k$ is defined by equation (3). Equation (8) adjusts the number of objects associated with the regularized covariance matrix as a linear combination of the number of training samples in class ω_k and the total number of training samples. The parameter $\lambda (0 \le \lambda \le 1)$ controls the degree of the regularization. A value $\lambda = 0$ results in quadratic discrimination, whereas for $\lambda = 1$, the discrimination is linear. Intermediate values result in a combination of the two.

It is well known that, for limited training samples, the estimates of the eigenvalues are biased.[8] The smallest eigenvalues are estimated too small, and the largest eigenvalue estimates are too large. The second regularization step (equation (10)) tries to counter this bias by scaling the covariance matrices towards the identity matrix multiplied by the average eigenvalue of $\hat{\Sigma}_k(\lambda)$.

$$\hat{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \frac{\gamma}{d}\mathrm{tr}\,[\hat{\Sigma}_k(\lambda)]\mathbf{I}, \quad 0 \le \gamma \le 1 \tag{10}$$

where tr $[\hat{\Sigma}_k(\lambda)]$ is the *trace* of $\hat{\Sigma}_k(\lambda)$, d the dimensionality and $\gamma (0 \le \gamma \le 1)$ a parameter controlling the second regularization. After applying both regularizations and using $\hat{\Sigma}_k(\lambda,\gamma)$ instead of $\hat{\Sigma}_k$, the classification score c_k (equation (2)) becomes

$$c_k(\mathbf{x}_i) = (\mathbf{x}_i - \hat{\mu}_k)^{\mathrm{T}}\hat{\Sigma}_k^{-1}(\lambda,\gamma)(\mathbf{x}_i - \hat{\mu}_k) + \ln|\hat{\Sigma}_k(\lambda,\gamma)| - 2\ln(P(\omega_k)). \tag{11}$$

The optimal values for parameters λ and γ are not known in advance and are jointly chosen such that the probability of correct classification (PCC) is maximized. This is achieved by evaluating the resulting classifier for a number of λ and γ pairs and choosing the values that result in the best performance (see references (6, 7) for details).

K-Nearest Neighbour method (KNN)

One of the most popular nonparametric methods is the *voting* KNN classifier. The KNN method classifies a test object into the class that supplies the largest number of neighbours among the K-nearest neighbours of the object. The simplicity, and the fact that KNN does not make any assumptions about the class distributions, are the main strong points of this method. However, being a nonparametric method, KNN has the drawback that it requires large numbers of training samples in order to perform well in high dimensional settings.[4]

Classification after reducing the dimensionality

Regularization is one way of coping with the problem of insufficient numbers of training samples in high dimensional settings. Another possible approach is to reduce the dimensionality prior to classification. This latter approach is taken by the four classifiers described in this section. More specifically, they are combinations of four different dimensionality reduction techniques followed by a classifier in the reduced space. Among many possible dimensionality reduction schemes, four methods which transform the data into two or three dimensions were chosen. The reasons for this choice were two-fold. Firstly, these transforms are popular because they enable the practitioner to visualize the data. Secondly, the inclusion of dimensionality reduction techniques which allow arbitrary numbers of dimensions to be retained would have considerably increased the complexity of the simulation. This is because the performance of such methods (such as variable selection or principal components) greatly depends on the criteria used to determine the resulting dimensionality. Often, no general best criterion exists, and either several are investigated or "rules of thumbs" are used for their selection.

These four classification techniques are referred to here as *two-stage* methods. The first stage is concerned with the dimensionality reduction and the second stage with classifying the test objects in the reduced space. The four methods investigated differ only in the first stage, i.e. the technique used to reduce the dimensionality. The classifiers used to assign the test data to classes in the second stage were 1NN and RDA, both described above. The rationale for not using LDA and QDA was that RDA includes these as special cases, and little would be gained since the data are well posed at the second stage.

In order to be able to correctly classify in the reduced space, the ability of the dimensionality reduction technique to separate the classes in the reduced space is of main importance. Therefore, the following descriptions of the two-stage methods will concentrate on the four different dimensionality reduction schemes.

Fisher's discriminant plane

Fisher's discriminant plane is based on two successive maximizations of the *Fisher criterion*, a measure of the class separability in the reduced space:

$$J(\vec{\phi}_i) = \frac{\vec{\phi}_i^{\mathrm{T}}\hat{\Sigma}_b\vec{\phi}_i}{\vec{\phi}_i^{\mathrm{T}}\hat{\Sigma}\vec{\phi}_i} \tag{12}$$

where $\hat{\Sigma}_b$ is the between-class scatter matrix and $\hat{\Sigma}$ the pooled covariance matrix. The between-class scatter matrix is defined as

$$\hat{\Sigma}_b = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_k - \hat{\mu}_0)(\hat{\mu}_k - \hat{\mu}_0)^{\mathrm{T}} \tag{13}$$

where $\hat{\mu}_0$ is the overall mean estimate, i.e. the mean of the set of all training samples. The d' vectors $\vec{\phi}_i$ that maximize $J(\vec{\phi})$ are the d' (where $d' \le \min(K-1, d)$) eigenvectors of $\hat{\Sigma}^{-1}\hat{\Sigma}_b$ corresponding to the d' largest eigenvalues.

The Fisher discriminant plane transform (FDP) maps the original data onto the plane spanned by two orthogonal vectors satisfying the Fisher criterion (equation (12)). These are the two first eigenvectors of $\hat{\Sigma}^{-1}\hat{\Sigma}_b$.

The second vector $\vec{\phi}_2$ can only be found if the pooled covariance matrix $\hat{\Sigma}$ estimated from the training samples can be inverted.[5] While this is not possible for ill-posed problems, Hong and Yang[5] have shown that if singular value perturbation is applied to the pooled covariance matrix, the inverse may be found and the resulting vector $\vec{\phi}_2$ is a good estimate of the theoretically optimal vector. The implementation used for the simulations follows this approach.

Fisher–Fukunaga–Koonz transform

Longstaff[9] has proposed the Fisher–Fukunaga–Koonz transform, for which he states that it yields a two-dimensional representation with usually more and never less discriminatory power than Fisher's discriminant plane. This new method, similar to FDP, uses the first Fisher vector as one of the vectors to span the resulting plane. The second vector is the Fukunaga–Koonz vector.

The Fukunaga–Koonz vector is the best discriminant vector in the case where the two means of a *two-class problem* coincide, and is defined by the direction in which the variances for the two classes differ the most. The vector is obtained by first applying the Fukunaga–Koonz transform, which rotates and scales the data such that the pooled covariance matrix becomes the identity **I**. It can be shown that after the transform, the two classes have the same eigenvectors and any two corresponding eigenvalues add to unity. Hence the Fukunaga–Koonz vector is the eigenvector associated with the two eigenvalues that differ the most. Furthermore, after performing the Fukunaga–Koonz transform, the first Fisher vector is collinear to the vector defined by the two class means.[9]

This method as initially proposed is only defined for the two-class case. However, we extended the technique such that it may be applied to classification problems with any number of classes. The extension is straightforward, one class is set aside, the remaining $K - 1$ classes are pooled and considered as one class only. The problem is now reduced to a two class situation, and the Fisher–Fukunaga–Koonz transform is applied. In the resulting Fisher–Fukunaga–Koonz plane, a probabilistic classifier is applied and the probability of belonging to the class singled out is recorded for each test object. After repeating this procedure K times for the K classes, each test object is classified into the class which, when singled out, scored the largest probability.

Fisher-radius plane transform (FR)

The Fisher-radius transform, also proposed by Longstaff,[9] transforms the data from the full feature space onto two or three dimensions using a *nonlinear* mapping. The advantage of nonlinear mappings is that these are able to make use of characteristics of the data inaccessible to linear projections. This may in certain situations lead to superior discrimination results. On the other hand, a major benefit of *linear* mappings onto two dimensions is that a visual appreciation of the class distributions may be gained. This advantage is likely to be lost with a nonlinear transform, since the nonlinearity in the mapping may make the interpretation of the class structures difficult.

As with the above transform, the Fukunaga–Koonz transform is initially performed, transforming the pooled covariance matrix into the identity matrix **I**. The first vector spanning the reduced space is again chosen to be the first Fisher vector, defined by the class means after the standardizing transform.

In order to illustrate the nonlinear part of the mapping, suppose that in a two class problem, all eigenvalues of one class covariance matrix are less than those associated with the other class, and that the two class means coincide. In this setting, the main feature discriminating a sample belonging to the first class from a sample belonging to the second class is its distance in the feature space from the coinciding means.

If not all eigenvalues of one class are smaller than those of the other class, the eigenvectors may be partitioned into two sets. The first set consisting of the eigenvectors associated with those eigenvalues which are larger for the first class, the second set corresponding to eigenvalues which are larger for the second class. Now, two distances of a test object to the coinciding means may be calculated. The first is the distance in the subspace spanned by the first set of eigenvectors, the second is the distance in the subset of the second set.

As the first discriminant vector for the Fisher-radius method after applying the Fukunaga–Koonz transform is the first Fisher vector, colinear to the axis passing through the two means, the two class means coincide in the subspace perpendicular to this axis. Hence the two distances as described above can be calculated in the subspace perpendicular to the axis. These distances become the second and third coordinate, the first being the projection onto the Fisher vector.

This method as initially proposed by Longstaff is only applicable to two class problems. However, the transform was extended in the same way as the Fukunaga–Koonz method enabling its use for situations where $K > 2$.

Fisher-variance plane

Duchene and Leclercq[10] proposed a simple modification to the Fisher discriminant plane approach described above. As with all transforms described so far, the first vector spanning the resulting two-dimensional space is the first Fisher vector. The second vector to span the Fisher-variance plane is the first principal component of the pooled data in the subspace perpendicular to the first Fisher vector.

Simulation study

A large simulation study was undertaken in order to evaluate the performance of the eight classification methods in a wide range of settings. A previous simulation study showed that RDA performed better than LDA and QDA in ill- and poorly-posed settings.[6] However, that study only used normally distributed data and more importantly did not include classification methods which are based on an initial reduction of the dimensionality.

It was the aim of this simulation to gain insight into the relative classification performance of the full feature

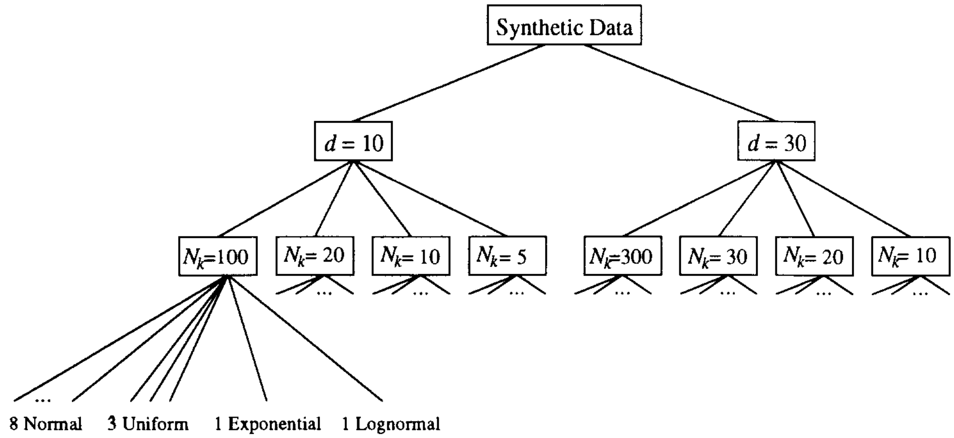

Fig. 1. Graphical representation of the 104 different settings of random data generated. Note that ten replicates were generated for each setting.

space methods and the two-stage classifiers. More specifically, the performance of the eight methods was studied with respect to the dimensionality d, the number of training samples per class N_k and the class distributions. Much is known in this respect for QDA and LDA,[2,3] but little about the behaviour of the three more recently proposed transformations investigated here.

Note that we restricted our comparisons to study the classification performance. In practice, other characteristics such as the computational efficiency, memory use and whether or not a classifier gives a measure of confidence with a particular decision also have to be taken into account. The classifiers were evaluated using nine real data sets and a wide range of synthetic data.

Synthetic data

The synthetic data were generated in order to gain insight into the relation between the characteristics of a classification problem and the performance of each of the eight classifiers. In this context, the term *setting* is used to denote classification problems which differ in one or several of the following; the dimensionality d, the number of training samples per class N_k, the shape of the class populations and the separation of the class means. The various combinations of shapes and class mean vectors used were motivated by the random data used in empirical studies by Friedman[6] and Rayens and Green.[11] For each setting, ten replicates were generated, each with its own training and test set. Furthermore, all settings were three-class problems and the size of the all test sets was 100 objects per class.

The first characteristic defining a specific setting is the dimensionality d. Settings with dimensionalities $d = 30$ and 10 were used. The second characteristic was the number of training samples N_k per class. For each of the two dimensionalities, four different sizes N_k of training sets were generated. For $d = 30$, these were $N_k = 300$, 30, 20 and 10, for $d = 10$, $N_k = 100$, 20, 10 and 5 were used. For each of the eight combinations of d and N_k, 13 different sets of three class populations were generated from four different distributions. Of these 13 settings, eight were generated using normal distributions, three using uniform distributions, one using an exponential and one a lognormal distribution. Combining all instances of the three parameters gave a total of 104 (2 × 4 × 13) settings. Figure 1 graphically depicts the 104 settings.

Details of the 13 different distributions are given below. As in reference (6), the covariance matrices are all diagonal and may be written for class ω_k as $\Sigma_k = \mathrm{D}(\sigma_i)$, $i = 1, \ldots, d$. Four different types of diagonal class covariance matrices have been used, denoted by $\mathrm{D1}_i(a,b)$, $\mathrm{D2}_i(a,b)$, $\mathrm{D3}_i(a,b)$ and $\mathrm{D4}_i(a)$. Functions D1 to D4 specify the ith diagonal element $\sigma_i (1 \leq i \leq d)$ as a function of a, b, i and d, where a and b specify the exact shape of the functions.

The four types are specified as follows:

$$\mathrm{D1}_i(a,b) \equiv \sigma_i^{(1)} = a + (b-a)\frac{(i-1)}{(d-1)}, \quad a < b, \qquad i = 1, \ldots, d, \text{ for } d = 10, 30 \quad (14)$$

$$\mathrm{D2}_i(a,b) \equiv \sigma_i^{(2)} = b + (a-b)\frac{(d-i)}{(d-1)}, \quad b < a, \qquad i = 1, \ldots, d, \text{ for } d = 10, 30 \quad (15)$$

$$\mathrm{D3}_i(a,b) \equiv \sigma_i^{(3)} = a + (b-a)\min\left(i, \frac{(d-i)}{(d/2)}\right), \quad a < b, \qquad i = 1, \ldots, d, \text{ for } d = 10, 30 \quad (16)$$

$$\mathrm{D4}_i(a) \equiv \sigma_i^{(4)} = a, \quad i = 1, \ldots, d, \text{ for } d = 10, 30. \quad (17)$$

If a class covariance matrix is of the type $\mathrm{D1}_i(a,b)$, then the associated distribution has the large variances in the dimensions corresponding to large i. Covariance matrices of type $\mathrm{D2}_i(a,b)$ have large variances in the low dimensions (small i), and classes corresponding to $\mathrm{D3}_i(a,b)$ in the medium dimensions. Class populations with covariance matrices of type $\mathrm{D4}_i(a)$ are circular. For populations with uniform distribution, the four functions define the range of each dimension rather than the variance.

Table 1. Specification of the 13 different combinations of class populations. Σ_k is the class covariance matrix of the kth class, $\mu_k(10)$ is the corresponding mean for $d = 10$, $\mu_k(30)$ for $d = 30$. See text for a specification of the functions D1–D4

		Class 1			Class 2			Class 3		
	Type	Σ_1	$\mu_1(10)$	$\mu_1(30)$	Σ_2	$\mu_2(10)$	$\mu_2(30)$	Σ_3	$\mu_3(10)$	$\mu_3(30)$
n1	normal	D4(1)	$\vec{0}$	$\vec{0}$	D4(2)	$\vec{0}$	$\vec{0}$	D4(3)	$\vec{0}$	$\vec{0}$
n2	normal	D1(1, 4)	$\vec{0}$	$\vec{0}$	D2(1, 4)	$\vec{0}$	$\vec{0}$	D3(1, 4)	$\vec{0}$	$\vec{0}$
n3	normal	D4(1)	$\vec{0}$	$\vec{0}$	D4(1)	$\vec{4}$	$\vec{2}$	D4(1)	$\pm\vec{4}$	$\pm\vec{2}$
n4	normal	D4(4)	$\vec{0}$	$\vec{0}$	D4(7)	$\vec{4}$	$\vec{2}$	D4(10)	$\pm\vec{4}$	$\pm\vec{2}$
n5	normal	D1(1, 10)	$\vec{0}$	$\vec{0}$	D1(1, 10)	$\vec{4}$	$\vec{2}$	D1(1, 10)	$\pm\vec{4}$	$\pm\vec{2}$
n6	normal	D1(1, 10)	$\vec{0}$	$\vec{0}$	D2(1, 10)	$\vec{4}$	$\vec{2}$	D3(1, 10)	$\pm\vec{4}$	$\pm\vec{2}$
n7	normal	D1(1, 5)	$\vec{0}$	$\vec{0}$	D2(1, 5)	$\vec{4}$	$\vec{2}$	D3(1, 5)	$\pm\vec{4}$	$\pm\vec{2}$
n8	normal	D1(1, 20)	$\vec{0}$	$\vec{0}$	D2(1, 20)	$\vec{4}$	$\vec{2}$	D3(1, 20)	$\pm\vec{4}$	$\pm\vec{2}$
u4	uniform	D4(4)	$\vec{0}$	$\vec{0}$	D4(7)	$\vec{4}$	$\vec{2}$	D4(10)	$\pm\vec{4}$	$\pm\vec{2}$
u5	uniform	D1(1, 10)	$\vec{0}$	$\vec{0}$	D1(1, 10)	$\vec{4}$	$\vec{2}$	D1(1, 10)	$\pm\vec{4}$	$\pm\vec{2}$
u6	uniform	D1(1, 10)	$\vec{0}$	$\vec{0}$	D2(1, 10)	$\vec{4}$	$\vec{2}$	D3(1, 10)	$\pm\vec{4}$	$\pm\vec{2}$
e6	exponential	D1(1, 10)	—	—	D2(1, 10)	—	—	D3(1, 10)	—	—
l6	lognormal	D1(0, 2)	$\vec{0}$	$\vec{0}$	D2(0, 2)	$\vec{0}$	$\vec{0}$	D3(0, 2)	$\pm\vec{0}$	$\pm\vec{2}$

Each of the 13 distributions generated is specified in Table 1. Note that when specifying the means, the notation $\vec{k}$ specifies the vector $(k, k, \ldots, k)^T$, the notation $\pm\vec{k}$ specifies the vector $(k, -k, k, -k, \ldots, -k)^T$, where the number of elements in the vector equals the dimensionality (10 or 30).

Two of the distributions are asymmetric. Each variable of data objects corresponding to settings **e6** is exponentially distributed with probability density function $f(x) = (1/\beta)e^{-x/\beta}$. Here, functions D1–D4 specify the parameter β as a function of the dimension. The second asymmetric distribution is lognormal (**l6**). For settings **l6**, instances of each variable are generated as $y = e^x$, where x follows a normal distribution. In this case, the functions D1–D4 specify the distribution of x.

Real data sets

To verify and supplement the results obtained using the generated data, the classifiers were also compared on the basis of nine real data sets. A brief description for each is given below.

UNIX command summary data. These data summarizes the daily use of UNIX commands by users of a set of networked workstations. The actual data extracted for the classification was the following. For seven users, the relative frequencies of use of five UNIX commands in a day was collected over 12 days. The seven different users define the seven classes and each day for which data is available for a particular user constitutes a sample in the class corresponding to this user. This results in a five-dimensional problem with seven classes and 8–12 samples per class. The commands used were "finger", "man", "who", "date" and "more". The relative use of these commands is different for different people and the aim was to tell which of the seven users a particular sample of frequencies belongs to.

Wine recognition data. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.[12] The analysis determined the quantities of 13 constituents found in each of the three types of wines. Hence the data is 13-dimensional with three classes defined by the three cultivars. The training sets were large with 59, 71 and 48 training samples per class, making this a well-posed problem.

Tertiary institutions data. This data set summarizes the responses of academics asked to evaluate the Australian institution they worked at.[13] Responses from a total of 32 institutions were averaged for each of the institutions, such that each institution represented one sample in the database. The institutions were divided into four classes: Old Universities (6), Young Universities (10), Colleges (11) and Institutes of Technology (5), where the number in brackets indicates the number of institutions in the class. For the purpose of classification, a total of 53 socioeconomic variables were included. Since the numbers of training samples are very small compared to the relatively high dimensionality, this is an ill-posed classification problem.

Cancer data. This data set relates to three types of pathological lung cancer.[5] Each type is described by 56 variables, the variables taking on integer values 1–4. The numbers of training samples are very small, 9 from the first, 13 from the second and 10 from the third type of cancer, rendering the problem very ill-posed.

Nuclear Magnetic Resonance (NMR) data. This data set arose from the search of new chemical compounds in corals. Data were obtained from 26 different molecules using NMR. The molecules all have a common substructure, except for the subtlety of a particular bond being either "up" or "down", defining the two classes of 13 molecules each. Each of the 19 carbons in the common substructure has a characteristic resonant frequency easily obtainable using NMR, constituting the 19 variables measured. With 19 dimensions and 13 training samples per class, this problem is ill-posed.

Glass types data. This data set is from reference (14). It summarizes a chemical analysis done on two types (classes) of glass. Glass which was float-processed and such which was not. The data has a dimensionality equal to 10 and is well-posed, with 87 training samples for the float-processed and 76 training samples for the second, non float-processed type.

Low Resolution Spectrometry (LRS) data. This data is from the Infra-Red Astronomy Satellite (IRAS) which was the first to attempt to map the full sky at infra-red wavelengths (data obtained from reference (14). The Low Resolution Observation (IRAS-LRS) programme observed high intensity sources over two continuous spectral bands. This database contains a total of 508 spectra derived from the IRAS-LRS database. There are four classes, corresponding to four types of stellar objects with $N_k = 90, 276, 39$ and 103. There are a total of 44 blue band and 49 red band channels of flux measurements, resulting in a total of 93 variables. This setting is very high dimensional, and ill-posed since one of the class sizes is very small compared to the number of dimensions.

Soya bean disease data. This is an extensive database on 15 different diseases of soya beans (data obtained from reference (14)). The sizes of the training samples are small; 10 for nine of the diseases, 20 for two and 40 for the remaining four diseases. The data is 35-dimensional, hence the problem is ill-posed with a large number of classes.

Fisher's Iris flower data. This is one of the most well-known databases in pattern recognition.[15] For three types of Iris flowers, measurements of four variables were taken. There are 50 training samples for each class, making this a well-defined problem.

METHODOLOGY

In order to evaluate and compare the classifiers on the basis of unbiased estimates of the relevant probability of correct classification (PCC), each of the eight classifiers was evaluated using independent test sets with the synthetic data, and the leave-one-out method with the real data. For each of the 104 settings of the synthetic data, the results obtained for the 10 replicated were averaged.

Of interest is the way the leave-one-out procedure was implemented for the four two-stage classifiers. Often, this type of classifier has been evaluated by applying the leave-one-out procedure in the reduced space only. However, this may lead to a strongly biased estimate in high dimensional settings. In order to obtain a truly unbiased estimate, the object left out also needs to be excluded from the dimensionality reduction scheme implemented by the classifier. This approach was followed in the comparisons presented here. The PCC recorded for the two-stage classifiers was the better of the ones achieved by RDA and 1NN, both of which were applied in the reduced space.

For computational reasons and in order to avoid introducing bias by evaluating several values for K (number of nearest neighbours), K = 1 was used throughout the simulations. The real data were standardized before applying the 1NN method. Because the simulation included a wide range of data sets, the use of the most commonly applied distance measure, the Euclidean distance, seemed most appropriate.

RESULTS

For reasons of clarity, we first discuss the performance of each method individually. The next main section (Discussion) will then relate the performances, giving both conclusions and recommendations. The results obtained for the real data sets are presented in Table 2. For the synthetic data, the results are presented as graphs for settings **n1**, **n4**, **n5**, **n6**, **e6** and **l6**. The results for setting **n2** are similar to those of **n1**. All classifiers performed well for **n3**. Settings **n7** and **n8** produced results which were very similar to those of **n6**. The uniform settings **u4**, **u5** and **u6** produced results which were very similar to the multivariate normal equivalents **n4**, **n5** and **n6**. Full results are given in reference (13).

Linear Discriminant Analysis (LDA)

The performance of LDA is mainly determined by the degree to which the data satisfy the assumption of equal class covariance matrices. If the class covariance matrices are the same, LDA is the only method capable

Table 2. Results obtained by the eight classifiers for the nine real data sets. The number in brackets in the column for KNN is the number of nearest neighbours that gave the best result

Data set	QDA	LDA	RDA	1NN	FDP	FF	FR	FV
UNIX commands	11.5%	21.3%	70.5%	62.3% (1)	57.4%	9.4%	14.7%	50.8%
Wine data	99.4%	98.9%	100%	91.0% (7)	88.2%	99.4%	93.8%	74.1%
Tert. institutions	37.5%	31.2%	84.4%	75.0% (1)	84.4%	62.5%	34.4%	71.9%
Cancer types	31.2%	18.7%	62.5%	50.0% (4)	59.4%	37.5%	46.9%	62.5%
NMR data	46.1%	65.4%	80.8%	50.0% (3)	61.5%	69.2%	53.8%	73.2%
Glass types	62.6%	71.2%	74.2%	81.0% (1)	62.0%	58.9%	62.0%	68.1%
LRS data	27.9%	27.9%	91.4%	91.3% (1)	80.7%	—	—	78.1%
Soya diseases	13.8%	13.8%	94.1%	90.3% (1)	38.3%	90.3%	—	59.7%
Fisher's Iris data	97.3%	98.0%	98.0%	96.7% (3)	97.3%	96.0%	70.7%	98.0%

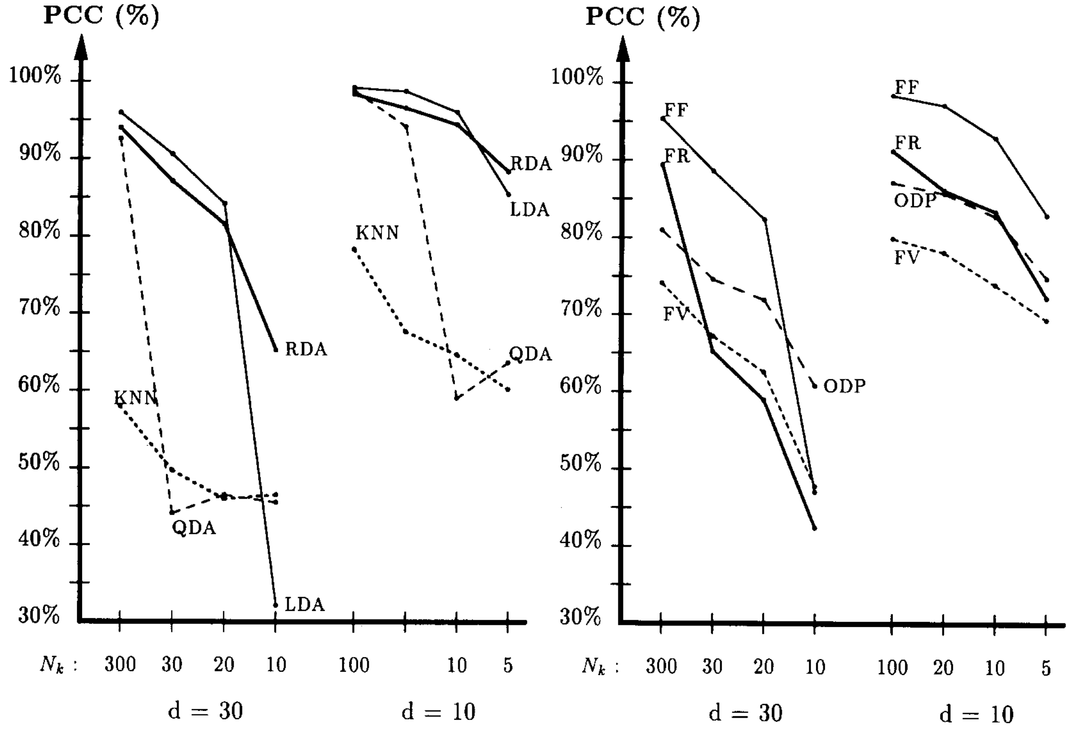

Fig. 2. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings **n5**.

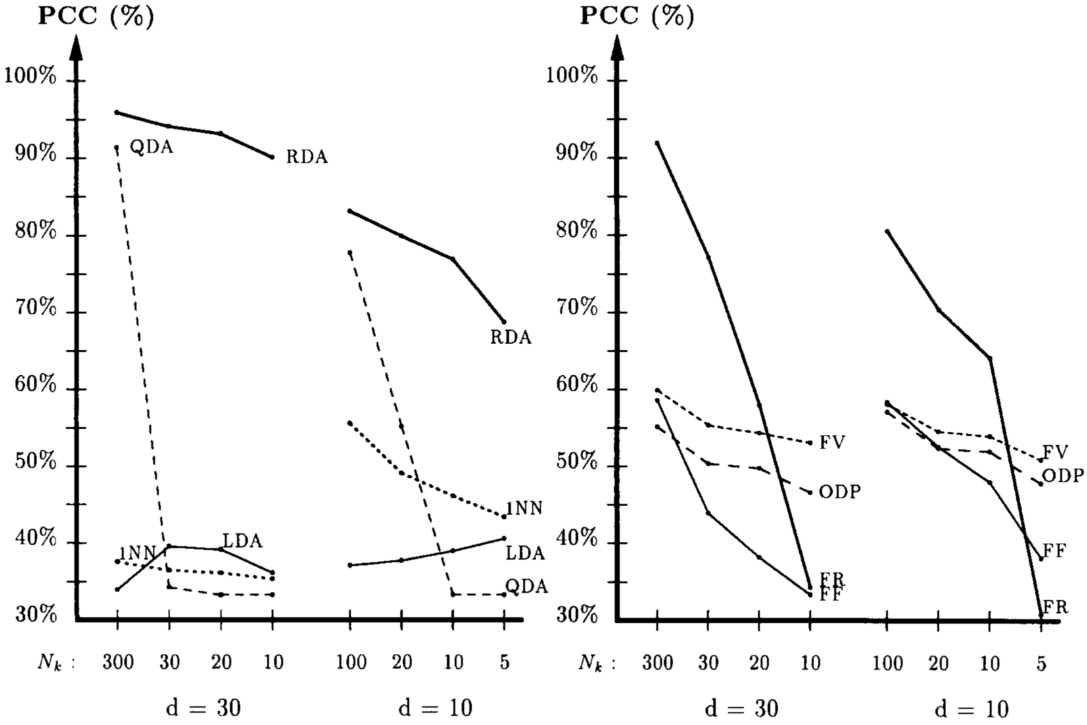

Fig. 3. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings **n1**.

of performing better than RDA (Fig. 2), and only so in the cases of large training sets. If the covariance matrices differ, then the larger the difference, the worse the performance. This was illustrated with settings **n6**, **n7**, and **n8**, where LDA performed best for **n7** and worst for **n8**. LDA has also performed relatively well with the exponential data of settings **e6**. The worst performance was shown with settings **n1** and **n2** (equal class means), where the separability is purely based on the differences in the class shapes (Fig. 3), and generally for worst-posed case of all settings, i.e. for $d = 30$ and $N_k = 10$.

The performance was somewhat mixed with the real data. LDA did well with the three well-posed problems (wine, glass and Iris data), but performed poorly for problems with many classes or high dimensionalities.

Quadratic Discriminant Analysis (QDA)

As with LDA, the results QDA confirm those from previous publications.[2,3,6] QDA performs very well in the cases where many training samples are available, but collapses for poorly- and ill-posed problems, due to poor parameter estimates. For the asymptotic settings ($N_k = 300/100$), QDA performed equally well for any class shapes except perhaps **e6**, which is non-Gaussian. The same observations were made with the real data, where QDA performed well in cases of many training samples, i.e. the wine and the Iris data. QDA displayed unusual behaviour with the lognormal data of setting **l6**, suspected to be due to the perturbation applied to the covariance matrices to enable their inversion.

Regularized Discriminant Analysis (RDA)

RDA has performed best in nearly all settings. RDA was outperformed by LDA for six of the eight settings of **n5** (Fig. 2) and by FDP and FF for the case of $d = 30$ and $N_k = 300$ with the non-Gaussian data of settings **e6** (Fig. 4). However note that RDA was only slightly inferior in these cases.

While RDA has also performed best overall for the real data sets, its superiority was not as clear as with the synthetic data. Most specifically, the 1NN method performed significantly better for the glass data. This is in contrast to the results obtained from the synthetic data. A possible explanation is given in the next section.

Nearest Neighbour Method (1NN)

As 1NN is a nonparametric method, it was worst affected by the high dimensionality of the settings. For $d = 30$, not even 300 training samples were enough to achieve good results. Consequently, 1NN does usually much better for $d = 10$ than for $d = 30$. A second characteristic observed for 1NN was that it performed better in cases where the class covariance matrices differ. However, most noticeable was that 1NN has excelled with some of the real data sets, especially with the glass and the LRS data. This is somewhat in contradiction to the results obtained using the artificial data, where 1NN never performed better than RDA. An explanation is as follows.

The different glass samples were categorized into the two classes corresponding to float processed glass and such which was not float processed. However, it may be possible to further partition the two groups into classes which might be better separable than the two

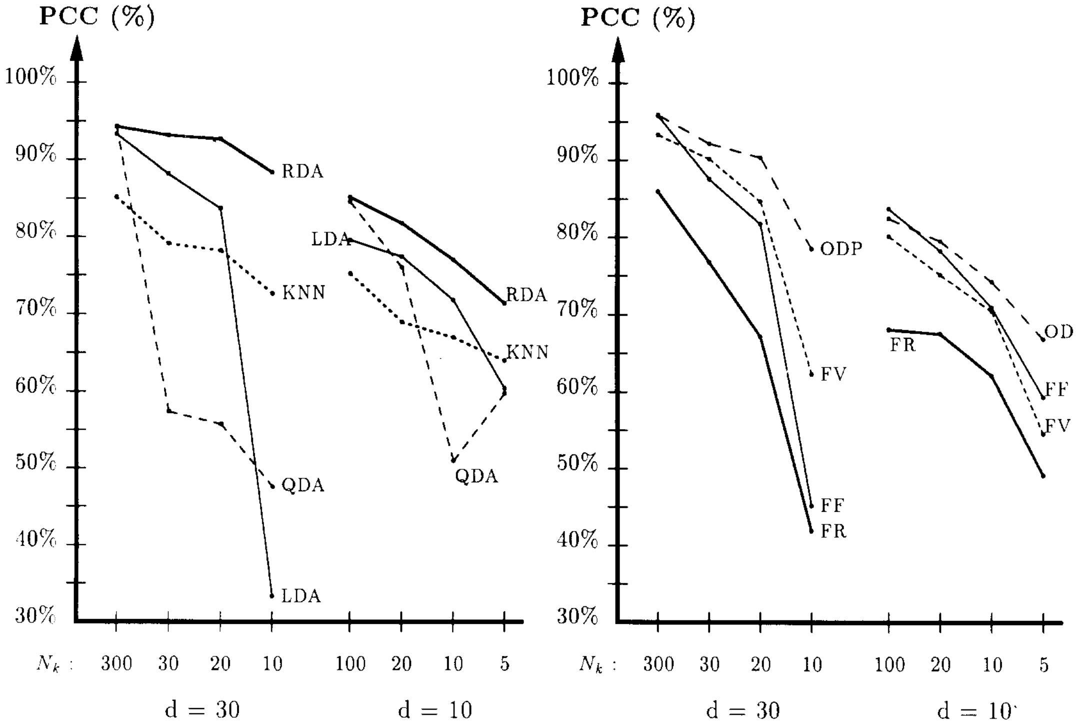

Fig. 4. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings **e6**, which are exponential in nature.

considered here. Say for example that there are only M manufacturers of glass, naturally partitioning the samples into M classes. Now assume that some of the M manufacturers only produce float processed glass, while the remaining only produce non-float processed glass. If the M classes are better separable than the two classes chosen, the 1NN method will essentially classify a test object into one of the M classes, and then "look up" whether the particular class corresponds to a float processing manufacturer or not.

This explanation is further supported by the LRS data, where 1NN again performed surprisingly well. In the original data, subclasses were indicated together with each of the four main classes. Because of the relatively small size of the training sample sets, only the four main classes were used to label the data. Quite likely, these subclasses are better separable than the super-classes, a characteristic which only the non-parametric 1NN method can make use of. Hence the surprisingly good performance of the 1NN method is likely to be an indication that these data violate some of the assumptions made by the other methods.

Fisher's Discriminant Plane (FDP)

FDP is the most traditional of the dimensionality reduction classifiers considered here. If the aims are solely in the context of classification, FDP does not offer any advantage, as RDA generally performs better. However, for data visualization, FDP performed best of all visualization procedures for small N_k. This is because finding the two first Fisher vectors reliably is still possible in these cases. FDP performs better if the class covariance matrices are not too different, and, because it relies on the Fisher criteria, performs very badly in the case where the class means coincide. FDP performed very well for the exponential data.

The results of the real data support the observations made from the simulations. FDP does not perform very well on well-defined data sets (wine data, Iris data), especially when compared to FF. It however compares somewhat better in the other cases, most noticeably in the case of the tertiary institutions data, where it equals the classification performance of RDA. Note that FDP performs very badly with the soya bean data, a result which is likely due to the high number of classes (15).

Fisher–Fukunaga–Koonz Transform (FF)

This new method has been said to perform asymptotically better than FDP for the two class case.[9] The results have shown that this is also true for the modified version proposed here, enabling the application of this method to more than two classes. FF has overall performed second best for the two largest N_k for each dimension. Only RDA performed better.

However, compared to FDP, FF deteriorates at a faster rate, making FDP the better method for small sizes of the training sets. This is because FF makes use of all the eigenvalues and eigenvectors. As the number of training samples is decreased, it becomes more and more difficult to obtain good estimates of the eigenstructure, with the smallest eigenvalues worst affected.

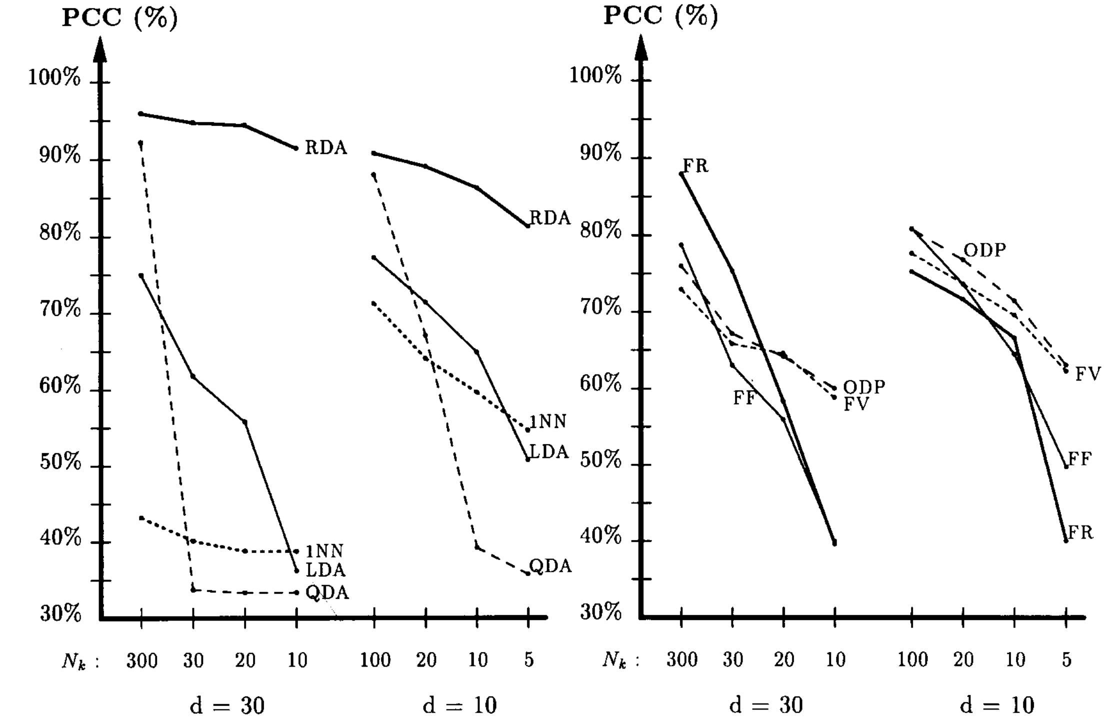

Fig. 5. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings n4.

In the case of an ill-posed problem, singular value perturbation replaces the zero eigenvalues with small constants. However, good estimates of the smallest eigenvalues are essential for FF to perform well, as they are used to determine the orientation of the Fisher–Fukunaga–Koonz plane. Hence FF performs badly in cases of small training samples when these estimates are unreliable.

The above observations are supported by the results obtained on the real data sets. FF performs well in well-posed problems, while it performs very poorly for some of the poorly- and ill-posed problems. There is one exception though. The soya data is, though quite well separable, ill-posed and consists of 15 classes. The performance of FF is excellent, only inferior to that of RDA. This indicates that the extension to FF proposed here also works well in cases where the number of classes is very large.

Fisher Radius transform (*FR*)

Asymptotically, FR has often achieved the best results of all dimensionality reduction techniques, especially for $d = 30$ (Figs 3, 5 and 6). However, even more than FF, FR proved to be extremely sensitive to a decrease in the number of training samples. As with FF, the reason for this lies in the dependence of the method on good estimates of the smallest eigenvalues, which cannot be obtained in poorly and ill-posed settings. Because of its nonlinear mapping, FR performed very well in cases where the class means coincide (Fig. 3) and large N_k.

One interesting point is that asymptotically (N_k = 300 or 100), FR performs better for $d = 30$ rather than for $d = 10$. The following is an intuitive reason why this might be the case. The class associated with the large variance may still have test objects which are relatively close to the Fisher axis, since only the variances are larger, but the mean is the same. These objects are then wrongly classified into the other class, associated with the small variance. The difference in high dimensionalities is that a much larger proportion of the objects are at the tail of the distribution. As a result, fewer objects of the class with the large variance are close to the Fisher axis. Hence fewer objects of the large variance class will be misclassified into the class associated with the small variance (and vice versa).

Another interesting finding is that FR does not perform much better for a particular type of distribution than for another. The first reason for this is that FR involves a nonlinear mapping. Having a particular distribution (e.g. normal) does not provide any advantage. The second reason is that because the extended FR method pools two classes at a time to reduce a three class problem to a two class problem, any distribution gets distorted in a similar manner.

FR has not performed well for any of the real data sets.

Fisher-Variance transform (*FV*)

The FV classifier performed very poorly asymptotically, because maximizing the variance does not necessarily lead to good class separation. However, as

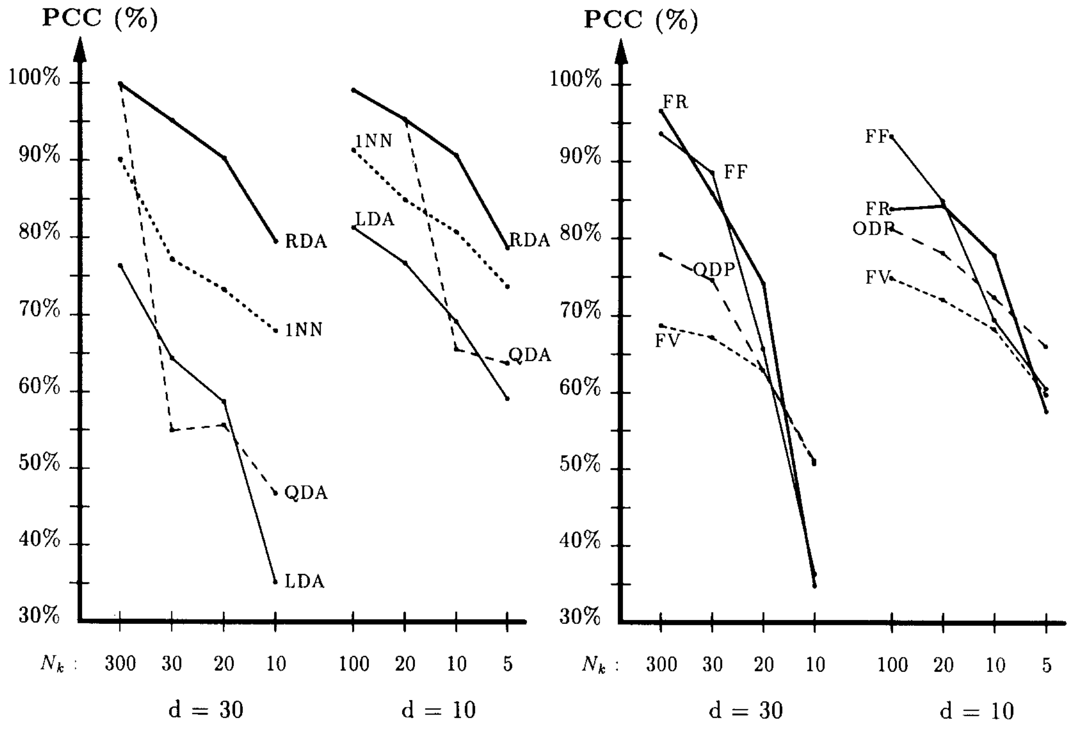

Fig. 6. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings n6.

the first Fisher vector and the direction of largest variance can be found reliably even in the case of small N_k, FV proves relatively robust in cases of small sample sizes.

FV did perform relatively well with some of the real data sets. Most noticeably among these are the cancer data, where FV achieves the same probability of correct classification as RDA.

DISCUSSION

The results show that if RDA is used, high dimensionalities are likely to be beneficial to how well the classes can be separated. It was found that in the vast majority of cases, reducing the dimensionality of a particular problem by the feature extraction methods investigated here lead to inferior classification results compared to those achieved by RDA in the full feature space.

The above observations extend to the design stage of classification. That is, when deciding on which features to measure on the sampled data, the conclusion is that the more features measured, the better, as long as each of these features is likely to contribute to the class separation. This is not to say that careful thought is not needed any longer in the design phase, but is to say that the number of features should not be restricted purely on the basis of avoiding a high dimensional classification problem.

Among the full feature space classifiers, LDA was found to be the only method capable of outperforming RDA, and only so in the case of identical class covariance matrices with many training objects. RDA has otherwise outperformed all other methods except for some of the exponential data. A method which did not perform well with the artificial data was 1NN. However, the results obtained from the real data sets have shown that 1NN is capable of outperforming RDA for reasons given above.

Generally, if classification is the sole goal, projecting the data into two dimensions is likely to result in a decreased PCC. We suggest that, in this case, RDA and perhaps KNN be applied (among the methods investigated by us). If the aim is to classify visualized data, then the use of either FDP or FF was found to be most effective. Which of the two performs better depends on the number of training samples and the shapes of the class populations.

Asymptotically, FF performs somewhat better than FDP, and hence may be used in these cases. As the number of training samples decreases, it was observed that FF deteriorates at a much faster rate than FDP, resulting in FDP performing significantly better than FF for small sample sizes.

These observations lead to a more general conclusion for the case where a two-stage classifier is to be applied to an ill- or poorly-posed problem. The first stage of the classifier tries to map the data in such a way that the class separation is maximized. It was found that in ill- and poorly-posed settings, simple but robust rules performed better than the more sophisticated ones. This is because more sophisticated rules have more

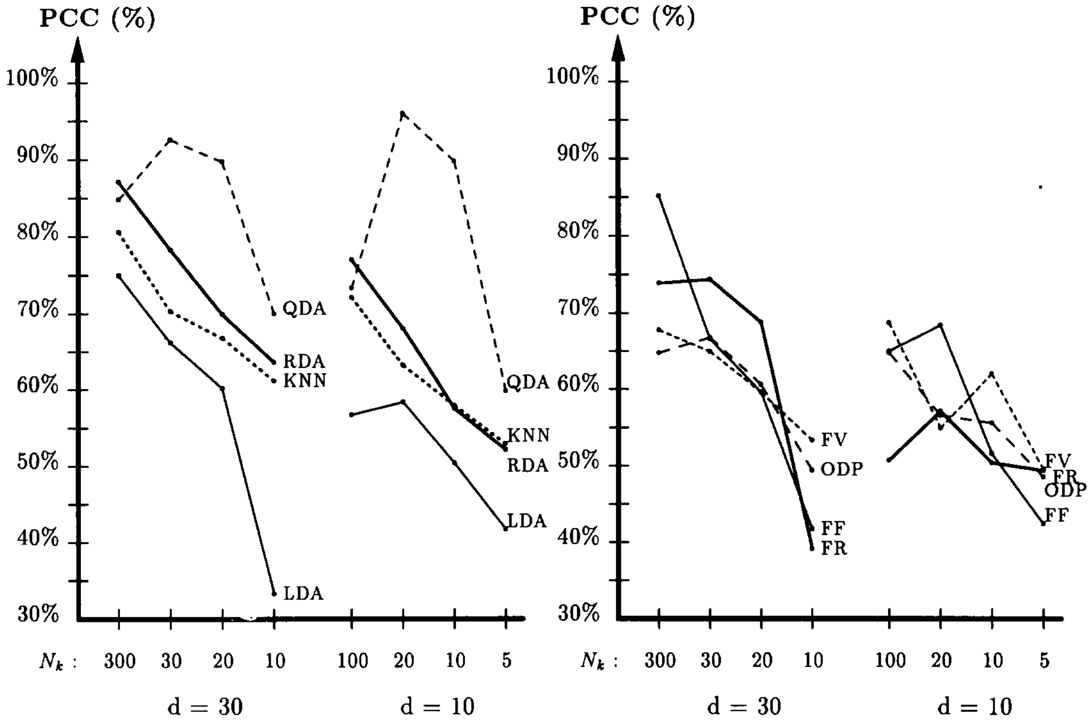

Fig. 7. Probabilities of correct classification (PCC) obtained by the eight classifiers for the eight settings I6, which are lognormal in nature.

parameters to estimate, estimates which become unreliable in ill- and poorly-posed settings. Hence the resulting projections are likely to yield bad class separations.

The FF method as initially proposed was only applicable for two class cases, a limitation which was removed by an extension proposed here. While increasing the computational cost of the FF method, the extension has proven effective with respect to three class classification problems. Results from the real data sets indicate that the method may still perform well even if the number of classes is significantly larger, as illustrated with the soya bean data.

FV and FR showed good performance in particular settings, but were overall not as capable in separating the classes as the other two-stage classifiers, FDP and FF. FR proved a powerful method in the special case where the class means coincide and sufficient training samples were available.

REFERENCES

1. D. Coomans and I. Broeckaert, *Potential Pattern Recognition*. Research Studies Press, Letchworth (1986).
2. A. G. Wacker and T. S. El-Sheikh, Average classification accuracy over collections of Gaussian problems—common covariance matrix case, *Pattern Recognition* **17**, 259–273 (1984).
3. K. Fukunaga and R. R. Hayes, Effects of sample size in classifier design, *IEEE Trans. Pattern Analysis Mach. Intell.* **11**, 873–885 (1989).
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, New York (1990).
5. Z. Q. Hong and J. Y. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, *Pattern Recognition* **24**, 317–324 (1991).
6. J. H. Friedman, Regularized discriminant analysis, *J. Am. Statist. Ass.* **84**, 165–175 (1989).
7. S. Aeberhard, D. Coomans and O. de Vel, Improvements in the performance of regularized discriminant analysis, *J. of Chemometrics* **7**, 99–115 (1993).
8. G. A. F. Seber, *Multivariate Observations*. Wiley, New York (1984).
9. I. D. Longstaff, On extensions to Fisher's linear discriminant function, *IEEE Trans. Pattern Analysis Mach. Intell.* **9**, 321–325 (1987).
10. J. Duchene and S. Leclercq, An optimal transformation for Discriminant and Principal Components Analysis, *IEEE Trans. Pattern Analysis Mach. Intell.* **10**(6), 979–983 (1988).
11. W. Rayens and T. Greene, Covariance pooling and stabilization for classification, *Computat. Statist. Data Analysis* **11**(17), 17–42 (1991).
12. M. Forina, R. Leard, C. Armanino and S. Lauter, PARVUS—an extendible package for data exploration, classification and correlation, Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy (1988).
13. S. Aeberhard, D. Coomans and O. de Vel, The performance of statistical pattern recognition methods in high dimensional settings, Tech. Rep. TR93-4, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland (1993).
14. P. M. Murphy and D. W. Aha, UCI repository of machine learning databases [Machine-readable data repository], University of California, Department of Information and Computer Science, Irvine, California (1991).
15. R. A. Fisher, The use of multiple measures in taxonomic problems, *Ann. Eugenica* **7**(II), 179–188 (1936).

About the Author—STEFAN AEBERHARD grew up in Switzerland and moved to Australia at the age of 20. He received his B.Sc.(Hons) from James Cook University in 1992, and is presently pursuing his Ph.D. on high dimensional pattern recognition.

About the Author—DANNY COOMANS was born in Belgium and received his Ph.D. in pharmaceutical sciences from the Free University of Brussels, Before he came to Australia in 1990, he was Research Director of the Dental Institute of the Free University of Brussels. He is Associate Professor in the Department of Mathematics and Statistics, James Cook University of North Queensland, Townsville. He has written over 80 international journal articles, and has authored two chemometrics books. His present research interest lies in the areas of pattern recognition and dynamic models. Applied fields of interest are chemometrics, industrial statistics and environment. He consults for the chemical industry in Australia and overseas.

About the Author—OLIVIER DE VEL obtained a M.Sc.(Hons) from the University of Waikato (New Zealand) in 1974 and a Docteur 3ieme Cycle from the Institut National Polytechnique of Grenoble (France) in 1978. Prior to his current position as Senior Lecturer in the Department of Computer Science at James Cook University, he has worked in various organizations in developing and developed nations. His research interests include computational learning, computer vision, and parallel computation.