"Small Sample Size": A Methodological Problem in Bayes Plug-in Classifier for Image Recognition

Technical Report 6/2001

Carlos E. Thomaz and Duncan F. Gillies

Department of Computing, Imperial College of Science Technology and Medicine,
180 Queen's Gate, London SW7 2BZ, United Kingdom
{cet,dfg}@doc.ic.ac.uk

Abstract. New technologies in the form of improved instrumentation have made it possible to take detailed measurements over recognition patterns. This increase in the number of features or parameters for each pattern of interest not necessarily generates better classification performance. In fact, in problems where the number of training samples is less than the number of parameters, i.e. "small sample size" problems, not all parameters can be estimated and traditional classifiers often used to analyse lower dimensional data deteriorate. The Bayes plug-in classifier has been successfully applied to discriminate high dimensional data. This classifier is based on similarity measures that involve the inverse of the sample group covariance matrices. However, these matrices are singular in "small sample size" problems. Thus, several other methods of covariance estimation have been proposed where the sample group covariance estimate is replaced by covariance matrices of various forms. In this report, some of these approaches are reviewed and a new covariance estimator is proposed. The new estimator does not require an optimisation procedure, but an eigenvector-eigenvalue ordering process to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. The effectiveness of the method is shown by some experimental results.

1 Introduction

New technologies in the form of improved instrumentation have made it possible to take detailed measurements over recognition patterns in a relatively cheap and efficient way. This increase in the number of features or parameters for each pattern of interest provides a wealth of detailed information, but not necessarily better classification accuracy. In fact, when the dimension of the feature space is larger than the number of training examples per class, not all parameters can be estimated and consequently pattern recognition techniques often used to analyse lower dimensional data deteriorate. This problem, which is called a "small sample size" problem [Fuk90], is indeed quite common nowadays, espe-

cially in image recognition applications where patterns are usually composed of a huge number of pixels.

Statistical pattern recognition techniques have been successfully applied to discriminate high dimensional data. In particular, the Bayes plug-in classifier is known as one of the most often used statistical parametric methods for high dimensional problems. This classifier is based on similarity measures that involve the inverse of the true covariance matrix of each class. Since in practical cases these matrices are not known, estimates must be computed based on the patterns available in a training set. The usual choice for estimating the true covariance matrices is the maximum likelihood estimator defined by their corresponding sample group covariance matrices. However, in "small sample size" applications the sample group covariance matrices are singular.

One way to overcome this problem is to assume that all groups have equal covariance matrices and to use as their estimates the weighting average of each sample group covariance matrix, given by the pooled covariance matrix calculated from the whole training set. This pooled covariance matrix will potentially have a higher rank than the sample group covariance matrices (and would be eventually full rank) and the variances of their elements are smaller than the variances of the corresponding sample group elements. Yet the choice between the sample group covariance matrices and the pooled covariance matrix represents a limited set of estimates for the true covariance matrices.

Thus, in the last 25 years, several other methods of covariance estimation have been proposed where the sample group covariance estimate is replaced by covariance matrices of various forms. As far as known, all these covariance estimators have been verified to improve the classification accuracy for computer-generated data, small training set recognition problems and moderate number of groups. In fact, these ideas have showed to be true in cases where no more than 20 groups are required, but have not been verified for a large number of groups.

In this way, one of the objectives of this report is to investigate the performance of proposed covariance estimators in image recognition problems that consider small training sets, large number of features and large number of groups. Biometric image recognition

problems, such as face recognition, are examples of promising applications. Another objective of this work is to propose a new covariance estimator for the sample group covariance matrices given by a convenient combination of the sample group covariance matrix and the pooled covariance one. This new estimator should have the property of having the same rank as the pooled estimate, while allowing a different estimate for each group.

The organisation of this report is as follows. Section 2 contains a brief and theoretical description of statistical pattern recognition problems, including a discussion between parametric and non-parametric methods. Section 3 describes the Bayes plug-in classifier and formally states the "small sample size" problems related to this type of classifier. A number of Bayes plug-in covariance estimators available in statistical pattern recognition regarding the difficulties inherent to small sample size settings is reviewed in section 4. Section 5 presents the basic concepts of the new covariance estimator proposed in this work. An account of experiments and results carried out to evaluate and compare the new covariance approach in two pattern recognition applications is presented in section 6. Finally, section 7 summarises the main ideas presented and concludes the work.

2 Statistical Pattern Recognition

Statistical pattern recognition has been used successfully to design several recognition systems. In statistical pattern recognition, a pattern is represented by a set of *p* features or measurements and is viewed as a point in a *p*-dimensional space.

The decision making process in statistical pattern recognition consists of assigning a given pattern with p feature values $x = [x_1, x_2, ..., x_p]^T$ to one of g groups or classes $\pi_1, \pi_2, ..., \pi_g$ on the basis of a set of measurements obtained for each pattern. The measurements associated with the population of patterns belonging to the π_i class are assumed to have a distribution of values with probability conditioned on the pattern class (or probability density function). That is, a pattern vector *x* belonging to class π_i is viewed as an observation drawn randomly from the class-conditional probability function $p(x \mid \pi_i)$.

There are a number of decision rules available to define appropriate decision-making boundaries, but the Bayes rule that assigns a pattern to the group with the highest conditional probability is the one that achieves minimal misclassification risk among all possible rules [And84].

The idea behind the Bayes rule is that all of the information available about group membership is contained in the set of conditional (or posterior) probabilities. The Bayes decision rule for minimizing the risk in the case of a 0/1 loss function can be formally stated as follows [Jam85]: Assign input pattern x to class π_i if

$$P(\pi_i \mid x) > P(\pi_j \mid x) \text{ , for all } j \neq i . \qquad (\mathbf{1})$$

If there is more than one group with the largest conditional probability then the tie may be broken by allocating the object randomly to one of the tied groups. Yet quantities such as $P(\pi_i \mid x)$ are difficult to find by standard methods of estimation, this is not the case, however, for quantities such as $p(x \mid \pi_i)$. The probability of getting a particular set of measurements x given that the object comes from class π_i, that is the class-conditional probability $p(x \mid \pi_i)$, can be estimated simply by taking a sample of patterns from class π_i (likelihood information). Fortunately there is a connection between $P(\pi_i \mid x)$ and $p(x \mid \pi_i)$ known as the Bayes theorem [Jam85]:

$$P(\pi_i \mid x) = \frac{p(x \mid \pi_i)\, p(\pi_i)}{\sum_{\text{all i}} p(x \mid \pi_i)\, p(\pi_i)} . \qquad (\mathbf{2})$$

It is important to note that all the items on the right-hand side of the equation (2) are measurable quantities and so can be found by sampling. The item $p(\pi_i)$ is simply the probability that the pattern comes from class π_i in the absence of any information (prior probability), i.e. it is the proportion of class π_i in the population. Using the Bayes theorem described in (2) with the previous Bayes rule (1) gives the following decision rule: Assign input pattern x to class π_i if

$$\frac{p(x \mid \pi_i)\, p(\pi_i)}{\sum_{\text{all k}} p(x \mid \pi_k)\, p(\pi_k)} > \frac{p(x \mid \pi_j)\, p(\pi_j)}{\sum_{\text{all k}} p(x \mid \pi_k)\, p(\pi_k)} \text{, for all } j \neq i \text{ and } 1 \leq k \leq g . \qquad (\mathbf{3})$$

As on both sides of the inequality the denominators are equal, the Bayes rule can be conveniently written as follows: Assign pattern x to class π_i if

$$p(x \mid \pi_i)\, p(\pi_i) = \max_{1 \le j \le g} p(x \mid \pi_j)\, p(\pi_j)\,. \qquad \textbf{(4)}$$

The classification rule defined in (4) is the final practical form of the optimal Bayes decision rule. This practical form of the Bayes decision rule is also called the maximum a posteriori (MAP) rule.

2.1 Parametric and Non-Parametric Methods

Several methods have been utilized to design a statistical pattern recognition classifier. Strategies for choosing the most appropriate method basically depend on the type and the amount of information available about the class-conditional probability densities.

The optimal Bayes rule discussed in the previous section can be used to design a classifier when all of the class-conditional densities are specified. In practice, however, the true class-conditional densities are typically not known and must be estimated from the available samples or training sets. If at least the form of the class-conditional densities is known (e.g. multivariate Gaussian distributions) but some of the parameters of these densities (e.g. mean vectors and covariance matrices) are unknown, then this problem is defined as a parametric decision problem. A common strategy to tackle this problem is to replace the unknown parameters in the density functions by their respective estimated values calculated from the training sets. This strategy is often referred as the Bayes plug-in classifier and will be fully described in the next section.

When the form of the class-conditional densities is either not known or assumed, non-parametric models have to be considered. In non-parametric problems, either the density functions must be estimated by using kernel functions or the class decision boundary has to be directly constructed based on the available training samples. These ideas are respectively the bases of the two most common non-parametric models: the Parzen Classifier and the *k*-nearest neighbour (k-NN) classifier – see [Fuk90] for more details.

Another subtle point in choosing a convenient statistical pattern method is related to the amount of information available. When a classifier is designed using a finite number of training samples, the expected probability of error is greater than if an infinite number of training samples were available. It is reasonable to expect that the probability of error decreases as more training samples are added and this behaviour obviously depends on the type of classifier used [Hof95]. Raudys and Jain [RJ91] found that the additional error due to finite training sample size for parametric classifiers was inversely proportional to *N*, where *N* is the total number of training samples. On the other hand, they showed that for the non-parametric Parzen classifier this additional error was inversely proportional to $\sqrt{N}$. This result indicates that the additional error due to finite training sample size decreases more quickly for parametric classifiers than for non-parametric ones [Hof95]. Therefore, since in image recognition applications the number of training samples is usually limited and significantly less than the number of parameters considered, parametric techniques seem to be more convenient than non-parametric ones.

3 The Bayes Plug-in Classifier

One of the most common parametric methods applied to statistical pattern recognition systems is the Bayes plug-in classifier.

The Bayes plug-in classifier, also called the Gaussian maximum likelihood classifier, is based on the *p*-multivariate normal or Gaussian class-conditional probability densities

$$p(x \mid \pi_i) \equiv f_i(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \left|\Sigma_i\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right], \qquad (5)$$

where μ_i and Σ_i are the true class *i* population mean vector and covariance matrix. The notation | |, i.e. a pair of vertical lines, denotes the determinant of a matrix.

The class-conditional probability densities $f_i(x \mid \mu_i, \Sigma_i)$ defined in (5) are also known as the likelihood density functions. Substituting equation (5) into equation (4) and assuming that all of the *g* groups or classes have the same prior probabilities, that is

$$p(\pi_i) = p(\pi_j), \text{ for all } j \neq i \text{ and } 1 \leq i, j \leq g, \qquad (6)$$

lead to the following Bayes classification rule form: Assign pattern x to class i if

$$f_i(x \mid \mu_i, \Sigma_i) = \max_{1 \le j \le g} f_j(x \mid \mu_j, \Sigma_j) \qquad (7)$$

Another way to specify equation (7) is to take the natural logarithms of the quantities involved, such as

$$d_i(x) = \ln[f_i(x \mid \mu_i, \Sigma_i)] = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) \qquad (8)$$

where $d_i(x)$ is often called the quadratic discriminant score for the ith class. Since the constant $(p/2)\ln(2\pi)$ is the same for all classes, it can be ignored. Therefore, the optimal Bayes classification rule defined in equation (7) may be simplified to: Assign pattern x to class i if

$$\begin{aligned} d_i(x) &= \max_{1 \le j \le g}[-\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)] \\ &= \min_{1 \le j \le g}[\ln|\Sigma_j| + (x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)] \\ &= \min_{1 \le j \le g} d_j(x) \end{aligned} \qquad (9)$$

The Bayes classification rule specified in (9) is known as the quadratic discriminant rule (QD). In addition, the measure $d_j(x)$ is sometimes referred to as the generalized distance between x and μ_j. The first term is related to the generalized variance of the jth group and the second term is the familiar Mahalanobis distance between x and the mean vector for the jth group.

In practice, however, the true values of the mean and covariance matrix, i.e. μ_i and Σ_i, are seldom known and must be replaced by their respective estimates calculated from the training samples available- this is when the term "plug-in" of the Bayes plug-in classifier takes place. The mean is estimated by the usual sample mean $\bar{x}_i$ which is the maximum likelihood estimator of μ_i, that is

$$\begin{aligned} &\frac{\partial}{\partial \mu}\left[\ln(f_i(x \mid \mu_i, \Sigma_i))\right] = 0 \quad \text{when} \\ &\mu_i \equiv \bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \ , \end{aligned} \qquad (10)$$

where $x_{i,j}$ is observation j from class i, and n_i is the number of training observations from class i. The covariance matrix is commonly estimated by the sample group covariance matrix S_i which is the unbiased maximum likelihood estimator of Σ_i, that is

$$\frac{\partial}{\partial \Sigma}\left[\ln(f_i(x \mid \mu_i = \bar{x}_i, \Sigma_i))\right] = 0 \quad \text{when}$$
$$\Sigma_i \equiv S_i = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \ . \tag{11}$$

Then, assuming normal and statistically independent data, the sample mean and the sample group covariance matrix estimates have the property that they maximise the joint likelihood of the training observations [JW98], that is, they maximise the product of the marginal normal density functions:

$$(\bar{x}_i, S_i) = \arg\max_{\mu_i, \Sigma_i} \prod_{j=1}^{n_i} f_i(x_{i,j} \mid \mu_i, \Sigma_i) \, . \tag{12}$$

From replacing ("plug-in") the true values of the mean and covariance matrix in (9) by their respective estimates, the QD rule can be rewritten as: Assign pattern x to class i that minimizes the generalized distance between x and $\bar{x}_i$

$$d_i(x) = \ln|S_i| + (x - \bar{x}_i)^T S_i^{-1} (x - \bar{x}_i) \, . \tag{13}$$

3.1 "Small Sample Size" Problems

The similarity measures used for Bayes plug-in classifiers involve the inverse of the true covariance matrices. Since in practice these matrices are not known, estimates must be computed based on the patterns available in a training set.

Although $\bar{x}_i$ and S_i are maximum likelihood estimators of μ_i and Σ_i, the misclassification rate defined in (13) approaches the optimal rate obtained by equation (9) only when the sample sizes in the training set approach infinity. Furthermore, the performance of (13) can be seriously degraded in small samples due to the instability of $\bar{x}_i$ and S_i estimators. In fact, for p-dimensional patterns the use of S_i is especially problematic if less

than $p + 1$ training observations from each class i are available, that is, the sample group covariance matrix is singular if n_i is less than the dimension of the feature space.

One method routinely applied to overcome the "small sample size" problem and consequently deal with the singularity and instability of the S_i is to employ the so-called linear discriminant rule (LD) which is obtained by replacing the S_i in (13) with the pooled sample covariance matrix

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{N - g}, \qquad (\mathbf{14})$$

where $N = n_1 + n_2 + \cdots + n_g$. Since more observations are taken to calculate the pooled covariance matrix, S_p is indeed a weighted average of the S_i, S_p will potentially have a higher rank than S_i (and would be eventually full rank) and the variances of their elements are smaller than the variances of the corresponding S_i elements. Thus, although theoretically S_p is a consistent estimator of the true covariance matrices Σ_i only when $\Sigma_1=\Sigma_2=\ldots=\Sigma_g$, various simulation studies including [MD74, WK77] have showed that LD can outperform QD for small sample sizes even when the individual Σ_i differ. In contrast, these simulation studies have also showed that QD can significantly outperform LD when the samples sizes are large and Σ_i differ.

In view of these studies, the choice between the sample group covariance matrices S_i (or QD classifier) and the pooled covariance matrix S_p (or LD classifier) represents a limited set of estimates for the true covariance matrices Σ_i. Several other methods of covariance estimation have been proposed where the sample group covariance estimate is replaced by covariance matrices of various forms. Some of these approaches are reviewed in the next section.

4 Covariance Estimators

As discussed in the previous section, a critical issue for the Bayes plug-in classifier is the instability and singularity of the sample group covariance estimate. Several approaches have been applied to overcome these problems and provide higher classification accuracy

with small training set size. Some of these covariance estimators can also be viewed as choosing an intermediate classifier between the LD and QD classifiers.

4.1 Eigenvalues Shrinkage Methods

When the class sample sizes n_i are small compared with the dimension of the measurement space p, the sample group covariance estimates defined in equation (11) become highly variable or even not invertible when $n_i < p+1$.

The effect of that instability has on the QD classifier can be seen by representing the sample group covariance matrices by their spectral decompositions [Fuk90]

$$S_i = \Phi_i \Lambda_i \Phi_i^T = \sum_{k=1}^{p} \lambda_{ik} \phi_{ik} \phi_{ik}^T , \qquad (\mathbf{15})$$

where λ_{ik} is the kth eigenvalue of S_i ordered in decreasing value and ϕ_{ik} is the corresponding eigenvector. According to this representation, the inverse of the sample group covariance matrix is

$$S_i^{-1} = \sum_{k=1}^{p} \frac{\phi_{ik} \phi_{ik}^T}{\lambda_{ik}} . \qquad (\mathbf{16})$$

Substituting equation (15) into equation (13) and using the well-known theorem that the determinant of a matrix Q is equal to the product of all its eigenvalues [Fuk90], the discriminant score of the QD rule becomes

$$d_i(x) = \sum_{k=1}^{p} \ln \lambda_{ik} + \sum_{k=1}^{p} \frac{[\phi_{ik}^T (x - \bar{x}_i)]^2}{\lambda_{ik}} . \qquad (\mathbf{17})$$

As can be observed, the discriminant score described in equation (17) is heavily weighted by the smallest eigenvalues and the directions associated with their eigenvectors. Therefore, the low-variance subspace spanned by the eigenvectors corresponding to the smallest sample eigenvalues has strong effect on the Bayes plug-in discriminant rule [Fri89].

Another problem related to equation (17) is the well-known upward bias of the large eigenvalues and downward bias of the smaller eigenvalues of the sample group covariance matrix. When the sample size decreases the estimates based on equation (11) produce

biased estimates of the eigenvalues, that is, the largest eigenvalues are larger than the eigenvalues of the true covariance and the smallest ones are biased toward lower values. In fact, when the sample covariance matrix is singular the smallest ($p - n_i + 1$) eigenvalues are estimated to be 0 and the corresponding eigenvectors are arbitrary, constraint to the orthogonality assumption [Fri89].

Hence, the effect of the instability and biasing incurred in estimating the Bayes plug-in parameters tend to exaggerate the importance associated with the low-variance information and consequently distort the corresponding discriminant analysis. Several investigators have demonstrated that Stein-like biased estimators, which basically shrink or expand the sample eigenvalues depending on their magnitude, dominate the sample covariance matrix under a variety of loss functions [Haf79, DS85]. Moreover, in [EM76] an estimate for the inverse of the sample covariance matrix that shrinks the eigenvalues of the sample group covariance matrix toward a common value has been developed. Also in [PN82], shrinkage estimates of this type have been substituted for S_p and the resulting rules outperform LD considering classification accuracy. In these works, the problem of estimating the covariance matrix S_i is based on its distribution, often called the Wishart distribution [And84].

In nearly all of the eigenvalues shrinkage methods quoted, the sample covariance matrix S_i must be non-singular, since the probability density function of the Wishart distribution does not exist unless the sample size n_i is greater than the number of parameters p [JW98]. As discussed earlier, this constraint is quite restrictive and, therefore, alternative covariance estimators have been provided by biasing the S_i towards non-singular matrices. The most often used approaches are described in the following sub-sections.

4.2 Discriminant Reguralization Method

Several regularization methods have been successfully used in solving poorly and ill-posed inverse problems [OSu86]. An estimation problem can be briefly stated as a poorly posed problem when the number of observations available (training set) is not considerably larger than the number of parameters to be estimated and ill-posed if this number of

parameters exceeds the training sample size. As a result, such parameter estimates become highly variable due to limited training set size.

Regularization methods attempt to reduce the variability of poorly and ill-posed estimates by biasing them toward values that are considered to be more "physically plausible" [Fri89]. The idea behind the term "regularization" is to decrease the variance associated with the limited sample-based estimate at the expense of potentially increased bias. The extent of this variance-bias trade-off is controlled by one or more regularization parameters [Fri89].

Friedman [Fri89] has proposed one of the most important regularization procedures called "reguralized discriminant analysis" (RDA). RDA is an alternative to the usual Bayes plug-in estimates for the covariance matrices and can be viewed as an intermediate classifier between the LD and QD classifiers.

The Friedman's approach is basically a two-dimensional optimisation method that shrinks both the S_i towards S_p and also the eigenvalues of the S_i towards equality by blending the first shrinkage with multiples of the identity matrix. In this context, the sample covariance matrices S_i of the discriminant rule defined in (13) are replaced by the following $S_i^{rda}(\lambda,\gamma)$

$$\begin{aligned} S_i^{rda}(\lambda,\gamma) &= (1-\gamma)S_i^{rda}(\lambda) + \gamma\left(\frac{tr(S_i^{rda}(\lambda))}{p}\right)I, \\ S_i^{rda}(\lambda) &= \frac{(1-\lambda)(n_i-1)S_i + \lambda(N-g)S_p}{(1-\lambda)n_i + \lambda N} \end{aligned} \quad , \qquad (\mathbf{18})$$

where the notation "tr" denotes the trace of a matrix, that is, the sum of all eigenvalues. Thus the regularization parameter λ controls the degree of shrinkage of the sample group covariance matrix estimates toward the pooled covariance one, while the parameter γ controls the shrinkage toward a multiple of the identity matrix. Since the multiplier $tr(S_i^{rda}(\lambda))/p$ is just the average eigenvalue of $S_i^{rda}(\lambda)$, the shrinkage parameter γ has the effect of decreasing the larger eigenvalues and increasing the smaller ones [Fri89]. This effect counteracts the aforementioned upward and downward biases of the sample group covariance estimates and favours true covariance matrices that are some multiples

of the identity matrix. In fact, the RDA method provides a number of regularization alternatives. Holding the mixing parameter γ at 0 and varying λ yields classification models between LD and QD classifiers. Holding λ at 0 and increasing γ attempts to unbias the sample-based eigenvalue estimates while holding λ at 1 and varying γ gives rise to ridge-like estimates of the pooled covariance matrix [DPi77, Cam80].

The mixing parameters λ and γ are restricted to the range 0 to 1 (optimisation grid) and are selected to maximise the leave-one-out classification accuracy based on the corresponding rule defined in (13). That is, the following classification rule is developed on the $N-1$ training observations exclusive of a particular observation $x_{i,v}$ and then used to classify $x_{i,v}$: Choose class *k* such that

$$d_k(x_{i,v}) = \min_{1 \le j \le g} d_j(x_{i,v}) \text{, with}$$
$$d_j(x_{i,v}) = \ln\left|S_{j/v}^{rda}(\lambda,\gamma)\right| + (x_{i,v} - \bar{x}_{j/v})^T \left(S_{j/v}^{rda}(\lambda,\gamma)\right)^{-1} (x_{i,v} - \bar{x}_{j/v}) \,. \qquad \textbf{(19)}$$

where the notation */v* represents the corresponding quantity with observation $x_{i,v}$ removed. Each of the training observations is in turn held out and then classified in this manner. The resulting misclassification loss, i.e. the number of cases in which the observation left out is allocated to the wrong class, averaged over all the training observations is then used to choose the best grid-pair (λ, γ).

Although Friedman's RDA method is theoretically a well-established approach and has the benefit of being directly related to classification accuracy, it has some practical drawbacks. First of all, RDA is a computationally intensive method. For each point on the two-dimensional optimisation grid, RDA requires the evaluation of the proposed estimates of every class. In situations where the optimisation has to be done over a fine grid and a large number of *g* groups is considered, for instance *g* is a number of 10^2 order, the RDA seems to be unfeasible. Also, despite the substantial amount of computation saved by taking advantage of matrix updating formulas based on the Sherman-Morrison-Woodbury formula [GL89], RDA requires the computation of the eigenvalues and eigenvectors for a (*p* by *p*) matrix for each value of the mixing parameter λ. In addition to the computational limitation, Greene and Rayens [GR91] have observed that RDA has the disadvan-

tage of partially ignoring information from a considerable portion of the data in the selection of the mixing parameters λ and γ - the same error rates could take place over a wide range of parameter values - and the optimal values of the grid-pair (λ, γ) are not unique. Therefore, a tie-breaking method needs to be applied. Finally, as RDA maximises the classification accuracy calculating all covariance estimates simultaneously, it is restricted to using the same value of the mixing parameters for all the classes. These same values may not be optimal for all classes.

4.3 Leave-One-Out Likelihood Method

The RDA [Fri89] method described in the previous sub-section uses the leave-one-out procedure to optimise its respective mixing parameters under a classification loss function. Since this function depends on calculating all covariance estimates simultaneously, Friedman's approach must employ the same mixing parameters for all classes. In practice, however, it is common to have classes with different forms and, consequently, different covariance matrices. In such situations, it seems appropriate to allow these covariance matrices to be estimated by distinct mixing parameters.

Hoffbeck [Hof95] has proposed a leave-one-out covariance estimator (LOOC) that depends only on covariance estimates of single classes. In LOOC each covariance estimate is optimised independently and a separate mixing parameter is computed for each class based on the corresponding likelihood information. The idea is to examine pair-wise mixtures of the sample group covariance estimates S_i (defined in (11)) and the unweighted common covariance estimate S, defined as

$$S = \frac{1}{g}\sum_{i=1}^{g} S_i \,, \qquad (\mathbf{20})$$

together with their diagonal forms. The LOOC estimator has the following form:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\mathrm{diag}(S) & 2 < \alpha_i \le 3 \end{cases}. \qquad (\mathbf{21})$$

The mixing or shrinkage parameter α_i determines which covariance estimate or mixture of covariance estimates is selected, that is: if $\alpha_i = 0$ then the diagonal of sample covariance is used; if $\alpha_i = 1$ the sample covariance is returned; if $\alpha_i = 2$ the common covariance is selected; and if $\alpha_i = 3$ the diagonal form of the common covariance is considered. Other values of α_i lead to mixtures of two of the aforementioned estimates [Hof95].

In order to select the appropriate mixing parameter α_i, the leave-one-out likelihood (LOOL) parameter estimation has been considered. In the LOOL technique, one training observation of the *i*th class training set is removed and the sample mean (defined in (10)) and the covariance estimates (defined in (21)) from the remaining $n_i - 1$ samples are estimated. Then the likelihood of the excluded sample is calculated given the previous mean and covariance estimates. This operation is repeated $n_i - 1$ times and the average log likelihood is computed over all the n_i observations. The Hoffbeck strategy is to evaluate several values of α_i over the optimisation grid $0 \le \alpha_i \le 3$, and then choose α_i that maximizes the average log likelihood of the corresponding *p*-variate normal density function, computed as follows:

$$\begin{aligned} LOOL_i(\alpha_i) &= \frac{1}{n_i}\sum_{v=1}^{n_i}\left[f\left(x_{i,v} \mid \bar{x}_{i/v}, S_{i/v}^{looc}(\alpha_i)\right)\right] \\ &= \frac{1}{n_i}\sum_{v=1}^{n_i}\left[-\ln\left|S_{i/v}^{looc}(\alpha_i)\right| - \frac{1}{2}(x_{i,v}-\bar{x}_{i/v})^T\left(S_{i/v}^{looc}(\alpha_i)\right)^{-1}(x_{i,v}-\bar{x}_{i/v})\right], \end{aligned} \quad \textbf{(22)}$$

where the notation */v* represents the corresponding quantity with observation $x_{i,v}$ left out. Once the mixture parameter α_i is selected, the corresponding leave-one-out covariance estimate $S_i^{looc}(\alpha_i)$ is calculated using all the n_i training observations and substituted for S_i into the Bayes discriminant rule defined in (13) [Hof95].

As can be seen, the computation of the LOOC estimate requires only one density function be evaluated for each point on the α_i optimisation grid, but also involves calculating the inverse and determinant of the (*p* by *p*) matrix $S_i^{looc}(\alpha_i)$ for each training observation belonging to the *i*th class. Although this is a one-dimensional optimisation procedure and consequently requires less computation than the two-dimensional RDA estimator previously discussed, LOOC is still computationally expensive. Hoffbeck has reduced signifi-

cantly the LOOC required computation by considering valid approximations of the covariance estimates and using the Sherman-Morrison-Woodbury formula [GL89] to write the estimates in a form that allows the determinant and inverse of each corresponding class to be computed only once, followed by a relatively simple computation for each left out observation. Therefore, the final form of LOOC requires much less computation than RDA estimator.

LOOC differs from the previous covariance method described basically in the mixtures it considers and the optimisation index utilised to select the best estimator. Although RDA estimator investigates the sample covariance matrix, pooled covariance matrix and the identity matrix multiplied by a scalar, LOOC employs the sample covariance matrix, unweighted common covariance matrix and the diagonal forms of these matrices. In LOOC the optimisation search is one-dimensional and limited to pair-wise mixtures, while in RDA estimator more general two-dimensional mixtures are considered. Moreover, the optimisation index maximised in LOOC is the leave-one-out likelihood that allows a separate mixing parameter to be computed for each class. On the other hand, RDA estimator uses leave-one-out optimisation procedures based on all the training observations of all classes and is restricted to using the same mixing parameters for all classes [Hof95].

Hoffbeck and Landgrebe have carried out several experiments with computer generated and remote sensing data to compare LOOC and RDA performances [Hof95, HL96]. In about half of these experiments, LOOC has led to higher classification accuracy than RDA, but required much less computation.

5 A New Covariance Estimator

As discussed previously, when the group sample sizes are small compared with the number of parameters the covariance estimates become highly variable and, consequently, the performance of the Bayes plug-in classifier deteriorates.

According to Friedman's RDA and Hoffbeck's LOOC approaches described in the previous section, and several other similar methods [GR89, GR91, Tad98, TL99] not de-

scribed in this report, optimised linear combinations of the sample group covariance matrices S_i and, for instance, the pooled covariance matrix S_p not only overcome the "small sample size" problem but also achieve better classification accuracy than LD and standard QD classifiers. However, in situations where S_i are singular, such approaches may lead to inconvenient biasing mixtures. This statement, which is better explained in the following section, forms the basis of the new covariance estimator idea, called Covariance Projecting Ordering method.

5.1 Covariance Projection Ordering Method

The Covariance Projection Ordering (COPO) estimator examines the combination of the sample group covariance matrices S_i and the pooled covariance matrix S_p in the QD classifiers using their spectral decomposition representations. This new estimator has the property of having the same rank as the pooled estimate, while allowing a different estimate for each group.

First, in order to understand the aforementioned inconvenient biasing mixtures, let a matrix S_i^{mix} be given by the following linear combination:

$$S_i^{mix} = aS_i + bS_p\,, \qquad (\mathbf{23})$$

where the mixing parameters *a* and *b* are positive constants, and the pooled covariance matrix S_p is a non-singular matrix. The S_i^{mix} eigenvectors and eigenvalues are given by the matrices Φ_i^{mix} and Λ_i^{mix}, respectively. From the covariance spectral decomposition formula described in equation (15), it is possible to write

$$(\Phi_i^{mix})^T S_i^{mix} \Phi_i^{mix} = \Lambda_i^{mix} = \begin{bmatrix} \lambda_1^{mix} & & & 0 \\ & \lambda_2^{mix} & & \\ & & \ddots & \\ 0 & & & \lambda_s^{mix} \end{bmatrix} = diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_s^{mix}]\,, \qquad (\mathbf{24})$$

where $\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_s^{mix}$ are the S_i^{mix} eigenvalues and *s* is the dimension of the measurement space considered[1]. Using the information provided by equation (23), equation (24) can be rewritten as:

$$\begin{aligned}(\Phi_i^{mix})^T S_i^{mix} \Phi_i^{mix} &= diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_s^{mix}] \\ &= (\Phi_i^{mix})^T \left[aS_i + bS_p\right] \Phi_i^{mix} \\ &= a(\Phi_i^{mix})^T S_i \Phi_i^{mix} + b(\Phi_i^{mix})^T S_p \Phi_i^{mix} \\ &= a\Lambda^{i*} + b\Lambda^{p*} \\ &= diag[a\lambda_1^{i*} + b\lambda_1^{p*}, a\lambda_2^{i*} + b\lambda_2^{p*}, ..., a\lambda_s^{i*} + b\lambda_s^{p*}]\end{aligned} \quad (\mathbf{25})$$

where $\lambda_1^{i*}, \lambda_2^{i*}, ..., \lambda_s^{i*}$ and $\lambda_1^{p*}, \lambda_2^{p*}, ..., \lambda_s^{p*}$ are the corresponding spread values of sample group covariance and pooled covariance matrices spanned by the S_i^{mix} eigenvectors matrix Φ_i^{mix}. Then, from equation (17), the discriminant score of the QD rule becomes

$$\begin{aligned}d_i(x) &= \sum_{k=1}^{s} \ln \lambda_k^{mix} + \sum_{k=1}^{s} \frac{[(\phi_{ik}^{mix})^T (x - \bar{x}_i)]^2}{\lambda_k^{mix}} \\ &= \sum_{k=1}^{s} \ln\left(a\lambda_k^{i*} + b\lambda_k^{p*}\right) + \sum_{k=1}^{s} \frac{[(\phi_{ik}^{mix})^T (x - \bar{x}_i)]^2}{a\lambda_k^{i*} + b\lambda_k^{p*}}\end{aligned} \quad (\mathbf{26})$$

As can be observed, the discriminant score described in equation (26) considers the dispersions of sample group covariance matrices spanned by all the S_i^{mix} eigenvectors. Therefore, in problems where the group sample sizes n_i are small compared with the dimension of the feature space *s*, the corresponding $(s - n_i + 1)$ lower dispersion values are estimated to be 0 or approximately 0. In this way, a linear combination as defined in equation (23) of the sample group covariance matrix and the pooled covariance in a subspace where the former is poorly represented seems to be not convenient. Other covariance estimators have not addressed this problem and have used the same parameters *a* and *b* defined in equation (23) for the whole feature space.

The COPO estimator is a simple approach to overcome this problem. Basically, the idea is to use all the sample group covariance information available whenever possible and the pooled covariance information otherwise. Regarding equations (23) and (25), this idea can be derived as follows:

[1] All over the text the dimension of the measurement space or, analogously, the number of parameters have been represented by the variable *p*. In this section, this variable representation was changed in order to avoid misunderstandings.

$$S_i^{copo} = \sum_{k=1}^{p} \lambda_k^{copo} \phi_{ik}^{copo} (\phi_{ik}^{copo})^T \text{ , where}$$
$$\lambda_k^{copo} = \begin{cases} \lambda_k^{i*} & \text{if } 1 \le k \le rank(S_i), \\ \lambda_k^{p*} & \text{otherwise,} \end{cases} \qquad (27)$$

and ϕ_{ik}^{copo} is the corresponding *k*-th eigenvector of the matrix given by $S_i + S_p$ ordered in λ_k^{i*} decreasing values. Then the discriminant scored described in equation (26) becomes:

$$d_i(x) = \sum_{k=1}^{r} \ln \lambda_k^{i*} + \sum_{k=r+1}^{s} \ln \lambda_k^{p*} + \sum_{k=1}^{r} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\lambda_k^{i*}} + \sum_{k=r+1}^{s} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\lambda_k^{p*}}, \qquad (28)$$

where $r = rank(S_i)$.

The COPO estimator provides a new combination of the sample group covariance matrices S_i and the pooled covariance matrix S_p in such a way that this combination is strongly related to the rank of S_i or, equivalently, to the number of training samples n_i. It can be viewed as an *s*-dimensional non-singular approximation of an *r*-dimensional singular matrix. Although several other covariance methods of combining or biasing S_i towards S_p have been developed, their optimisation procedures have not explicitly considered the sample group singularity effects. The COPO method does not require an optimisation procedure, but an eigenvector-eigenvalue ordering process to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. Therefore, the computational issues regarding the Covariance Projection Ordering approach is less severe than the Friedman's RDA and Hoffbeck's LOOC approaches. In addition, the COPO method is not restricted to use the same covariance combination for all classes, allowing covariance matrices to be distinctly estimated.

6 Experiments and Results

In order to evaluate the Covariance Projection Ordering (COPO) approach, two image recognition applications were considered: face recognition and facial expression recognition. The evaluation used two different image databases.

6.1 Databases

In the face recognition experiments the ORL Face Database[2] was used. This database contains a set of face images taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, U.K, with ten images for each of 40 individuals, a total of 400 images. All images were taken against a dark homogeneous background with the person in an upright frontal position, with tolerance for some tilting and rotation of up to about 20 degrees. Scale varies about 10%. The original size of each image is 92x112 pixels, with 256 grey levels per pixel.

The Tohoku University has provided the database for the facial expression experiment. This database [LBA99] is composed of 193 images of expressions posed by nine Japanese females. Each person posed three or four examples of each six fundamental facial expression: anger, disgust, fear, happiness, sadness and surprise. The database has at least 29 images for each fundamental facial expression.

For implementation convenience all images were first resized to 64x64 pixels.

6.2 Experiments

The experiments were carried out as follows. First the well-known dimensionality reduction technique called Principal Component Analysis (PCA) reduced the dimensionality of the original images and secondly the Bayes plug-in classifier using one of the five covariance estimators was applied:

1. Sample group covariance estimate *Sgroup* or S_i – equation (11);

2. Pooled covariance estimate *Spooled* or S_p – equation (14);

3. Friedman's reguralized discriminant estimate *Srda* or S_i^{rda} – equation (18);

4. Hoffbeck's leave-one-out estimate *Slooc* or S_i^{looc} – equation (21);

5. Covariance projection ordering estimate *Scopo* or S_i^{copo} – equation (27).

Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean and the standard deviation of the

[2] The ORL database is available free of charge, see http://www.cam-orl.co.uk/facedatabase.html

recognition rate were calculated. The face recognition classification was computed using for each individual in the ORL database 5 images to train and 5 images to test. In the facial expression recognition, the training and test sets were respectively composed of 20 and 9 images.

In order to compare the covariance methods based solely on the optimisation indexes utilised to select the best estimator, only RDA and LOOC mixtures that shrink the sample group covariance matrices towards the pooled covariance matrix were considered. In other words, the RDA γ parameter was held at 0 and the λ optimisation range was taken to be 20, that is, $\lambda = [0.05, 0.10, \ldots, 1]$. Analogously, the size of the LOOC mixture parameter was $\alpha_i = [1.05, 1.10, \ldots, 2]$. The stars over both label results indicate these special parameter selections.

6.3 Results

Tables 1 and 2 present the training and test average recognition rates (with standard deviations) of the ORL and Tohoku face and facial expression databases, respectively, over the different PCA dimensions.

Since only 5 images of each individual were used to form the ORL face recognition training sets, the results relative to the sample group covariance estimate (*Sgroup*) were limited to 4 PCA components. Table 1 shows that all the quadratic discriminant covariance estimators (*Srda*, *Slooc* and *Scopo*) performed better than the linear covariance estimator (*Spooled*), leading to higher training and test recognition rates. For the training samples, the *Scopo* estimator led to higher classification accuracy than the other two quadratic estimators considered. The *Srda* and *Slooc* estimators outperformed the *Scopo* in lower testing dimensional space, but these performances deteriorated when the dimensionality increased. The *Scopo* estimator achieved the best recognition rate – 96.6% – for all PCA components considered. In terms of how sensitive the covariance results were to the choice of training and test sets, the covariance estimators similarly had the same performances, particularly in high dimensional space.

PCA	Sgroup		Spooled		Srda*		Slooc*		Scopo	
	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
4	99.5(0.4)	51.6(4.4)	73.3(3.1)	59.5(3.0)	94.7(2.9)	75.9(3.4)	90.1(2.1)	70.8(3.2)	97.0(1.1)	69.8(3.4)
10			96.6(1.2)	88.4(1.4)	99.8(0.3)	93.8(1.7)	99.4(0.5)	92.0(1.5)	99.9(0.2)	90.2(2.5)
30			99.9(0.2)	94.7(1.7)	100.0	96.0(1.4)	100.0	95.9(1.5)	100.0	95.6(1.8)
50			100.0	95.7(1.2)	100.0	96.4(1.5)	100.0	96.4(1.5)	100.0	96.6(1.7)
60			100.0	95.0(1.6)	100.0	95.4(1.6)	100.0	95.8(1.6)	100.0	95.9(1.5)

Table 1. ORL face database results.

The results of the Tohoku facial expression recognition are presented in table 2. For more than 20 PCA components, when the sample group covariance estimates became singular, all the quadratic discriminant covariance estimators performed better than the linear covariance one for training and test samples. In lower dimension space, *Srda* led to higher classification accuracy, followed by *Scopo* and *Slooc*. However, analogously to the ORL face results, when the dimensionality increased, *Scopo* estimator apparently outperformed the other estimators, achieving the highest recognition rate – 84.7% – for all PCA components considered. In this recognition application, all the computed covariance estimators were quite sensitive to the choice of the training and test sets.

PCA	Sgroup		Spooled		Srda*		Slooc*		Scopo	
	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
10	76.3(3.6)	38.8(5.6)	49.6(3.9)	26.5(6.8)	69.8(4.6)	33.8(6.4)	58.5(3.7)	27.8(5.6)	66.5(3.2)	31.5(5.8)
15	99.7(0.5)	64.3(6.4)	69.1(3.6)	44.4(5.3)	92.7(5.3)	59.2(7.1)	82.9(2.9)	49.7(7.7)	90.4(3.0)	60.0(7.4)
20			81.2(2.6)	55.9(7.7)	98.1(1.5)	71.2(7.4)	91.4(2.8)	61.3(7.1)	95.6(1.9)	66.5(7.4)
40			95.9(1.4)	75.6(7.0)	98.5(1.5)	78.9(6.2)	98.3(1.1)	77.2(5.7)	98.2(0.8)	74.7(5.7)
65			99.5(0.6)	83.3(5.5)	99.7(0.5)	84.4(6.0)	99.8(0.4)	84.5(6.2)	99.9(0.2)	84.7(6.0)

Table 2. Tohoku facial expression database results.

7 Conclusion

A number of Bayes plug-in covariance estimators available in statistical pattern recognition have been described regarding the difficulties caused by small sample sizes. Some experiments carried out have confirmed that choosing an intermediate estimator between

the linear and quadratic classifiers improve the classification accuracy in settings for which samples sizes are small and number of parameters or features is large.

The new covariance estimator, called Covariance Projection Ordering method (COPO), has showed to be a powerful technique in small sample size image recognition problems, especially when concerns about computational costs exist. However, this statement is not conclusive. Comparisons between estimators like RDA and LOOC have to be analysed considering synthetic data and standard biometric databases, such as the FERET database (US Army Face Recognition Technology facial database).

A high-dimensional image recognition model involves not only the decision-making or classification process but also the data representation or feature extraction. In practice, a feature extraction algorithm is usually used to reduce the dimensionality of the data and, consequently, the number of parameters and computation time. The new covariance estimator might be helpful when using a feature selection algorithm, such as Linear Discriminant Analysis, that requires the estimation of covariance matrices in the high-dimensional space.

The aforementioned comparisons and feature analysis will provide a strong assessment of the COPO method.

References

[And84] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, second edition. New York: John Wiley & Sons, 1984.

[Cam80] N.A. Campbell, "Shrunken estimator in discriminant and canonical variate analysis", *Applied Statistics*, vol. 29, pp. 5-14, 1980.

[DPi77] P.J. Di Pillo, "Biased Discriminant Analysis: Evaluation of the optimum probability of misclassification", *Communications in Statistics-Theory and Methods*, vol. A8, no. 14, pp. 1447-1457, 1977.

[DS85] D.K. Dey and C. Srinivasan, "Estimation of a covariance matrix under Stein's loss", *Annals of Statistics*, vol. 13, pp. 1581-1591, 1985.

[EM76] B. Efron and C. Morris, "Multivariate empirical Bayes and estimation of covariance matrices", *Annals of Statistics*, vol. 4, pp. 22-32, 1976.

[Fri89] J.H. Friedman, “Reguralized Discriminant Analysis”, *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, March 1989.

[Fuk90] K. Fukunaga, Introduction to Statistical Pattern Recognition, second edition. Boston: Academic Press, 1990.

[GL89] G.H. Golub and C.F. Van Loan, *Matrix Computations*, second edition. Baltimore: Johns Hopkins, 1989.

[GR89] T. Greene and W.S. Rayens, “Partially pooled covariance matrix estimation in discriminant analysis”, *Communications in Statistics-Theory and Methods*, vol. 18, no. 10, pp. 3679-3702, 1989.

[GR91] T. Greene and W.S. Rayens, “Covariance pooling and stabilization for classification”, *Computational Statistics & Data Analysis*, vol. 11, pp. 17-42, 1991.

[Haf79] L.R. Haff, “Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity”, *Annals of Statistics*, vol. 7, pp. 1264-1276, 1979.

[HL96] J.P. Hoffbeck and D.A. Landgrebe, “Covariance Matrix Estimation and Classification With Limited Training Data”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763-767, July 1996.

[Hof95] J.P. Hoffbeck, “Classification of High Dimensional Multispectral Data”, PhD thesis, Purdue University, West Lafayette, Indiana, 1995.

[Jam85] M. James, *Classification Algorithms.* London: William Collins Sons & Co. Ltd, 1985.

[JW98] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, fourth edition. New Jersey: Prentice Hall, 1998.

[LBA99] M. J. Lyons, J. Budynek, and S. Akamatsu, “Automatic Classification of Single Facial Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, December 1999.

[MD74] S. Marks and O.J. Dunn, “Discriminant functions when the covariance matrices are unequal”, *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 555-559, June 1974.

[OSu86] F. O’Sullivan, “A Statistical Perspective on Ill-Posed Inverse Problems”, *Statistical Science*, vol. 1, pp. 502-527, 1986.

[PN82] R. Peck and J. Van Ness, “The use of shrinkage estimators in linear discriminant analysis”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 531-537, September 1982.

[RJ91] S.J. Raudys and A.K. Jain, “Small sample size effects in statistical pattern recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252-264, 1991.

[Tad98] S. Tadjudin, “Classification of High Dimensional Data With Limited Training Samples”, PhD thesis, Purdue University, West Lafayette, Indiana, 1998.

[TL99] S. Tadjudin and D.A. Landgrebe, “Covariance Estimation With Limited Training Samples”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, July 1999.

[WK77] P.W. Wahl and R.A. Kronmall, “Discriminant functions when the covariance are equal and sample sizes are moderate”, *Biometrics*, vol. 33, pp. 479-484, September 1977.