Biometrika (2006), **93,** 4, *pp.* 1018–1024

Printed in Great Britain

Discriminant analysis with common principal components

By MU ZHU

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

m3zhu@uwaterloo.ca

Summary

Zhu & Hastie (2003) presented a general criterion for finding discriminant directions. To optimise their criterion, iterative methods are needed unless each class has a Gaussian distribution with a common covariance matrix. In this short paper, we present a slightly more general case where iterative methods can also be avoided.

Some key words: Lagrange multiplier; Likelihood ratio; Linear discriminant analysis; Proportional covariance model; Quadratic discriminant analysis; Spectral decomposition; Swiss banknotes data.

1. Introduction

Suppose we have data $\{(x_i, y_i); i = 1, \ldots, n\}$, where $x_i \in \mathscr{R}^d$ is a vector of d predictors and $y_i \in \{1, \ldots, K\}$ is a class label. When d is relatively large, information useful for distinguishing the classes is often contained in a few directions $\alpha_1, \alpha_2, \ldots, \alpha_M \in \mathscr{R}^d$, where $M < d$; these directions are sometimes referred to as discriminant directions.

To find these directions, Zhu & Hastie (2003) proposed a general likelihood-ratio criterion for measuring the discriminatory power for a given direction α, with $\|\alpha\| = 1$:

$$\textsc{lr}(\alpha) = \log\left\{\frac{\max\limits_{p_k} \prod_{k=1}^{K} \prod_{x_j \in C_k} p_k^{(\alpha)}(\alpha^{\mathrm{T}} x_j)}{\max\limits_{p} \prod_{k=1}^{K} \prod_{x_j \in C_k} p^{(\alpha)}(\alpha^{\mathrm{T}} x_j)}\right\}, \tag{1}$$

where $p_k^{(\alpha)}(.)$ is the marginal density along the projection defined by α for class k, and $p^{(\alpha)}(.)$ is the corresponding marginal density under the null hypothesis that the classes share the same density function.

Zhu & Hastie (2003) showed how important discriminant directions can be derived by recursively maximising (1) even when there is no specific parametric assumption about the class density function $p_k (k = 1, \ldots, K)$. Of course, there are practical issues when we try to maximise $\textsc{lr}(\alpha)$ using iterative methods, most notably the problem of multiple local solutions.

Such problems are inevitable, except in a few special cases. If, for each k, p_k is the $N(\mu_k, \Sigma)$ density, then criterion (1) is equivalent to the well-known criterion used in Fisher's linear discriminant analysis:

$$\textsc{lda}(\alpha) = \frac{\alpha^{\mathrm{T}} B \alpha}{\alpha^{\mathrm{T}} W \alpha}, \tag{2}$$

where B and W are between- and within-class sample covariance matrices (Zhu & Hastie, 2003). The maximising solution is then the leading eigenvector of $W^{-1}B$ (Mardia et al., 1979, Ch. 11); iterative methods are not necessary.

It is natural to consider next the case when p_k corresponds to $N(\mu_k, \Sigma_k)$. This leads to quadratic discriminant analysis in that the Bayes decision boundaries are quadratic functions of x (Hastie et al., 2001, Ch. 4). Under such a model, apart from a constant not depending on α,

$$\text{LR}(\alpha) \propto \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)(\log \alpha^{\mathrm{T}} S\alpha - \log \alpha^{\mathrm{T}} S_k \alpha) \tag{3}$$

(Zhu & Hastie, 2003), where S is the total sample covariance matrix and S_k is the sample covariance matrix for class k. Unfortunately, iterative methods are still needed to maximise (3).

In this short paper, we look at another special case, one that is more general than Fisher's linear discriminant analysis but less general than quadratic discriminant analysis. For simplicity, we also follow a common practice and assume that the data are preprocessed such that $S = I$. Equation (3) then simplifies to

$$\text{LR}(\alpha) \propto \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)(-\log \alpha^{\mathrm{T}} S_k \alpha), \tag{4}$$

since $\alpha^{\mathrm{T}} S\alpha = \|\alpha\|^2 = 1$.

2. Main result

Consider the following reparameterisation of Σ_k based on its spectral decomposition:

$$\Sigma_k = U_k \Lambda_k U_k^{\mathrm{T}},$$

where $\Lambda_k = \text{diag}\{\lambda_{1k}, \lambda_{2k}, \ldots, \lambda_{dk}\}$. Let a_k be the largest eigenvalue of Σ_k. If we take out a_k as a factor and write $\lambda_{jk} = a_k q_{jk}$, then the parameters a_k, $Q_k \equiv \text{diag}\{q_{jk}\}$, and U_k can be seen to describe the size, shape and orientation of Σ_k (Banfield & Raftery, 1993; Bensmail & Celeux, 1996). A hierarchy of models can then be constructed by restricting some of these parameters to be identical across the K classes. Table 1 lists all the possible combinations.

Table 1. *A hierarchy of models, where* 's' *means* '*same*' *and* 'd' *means* '*different*'

Case	Size (a_k)	Shape (Q_k)	Orientation (U_k)	Also known as
1	s	s	s	LDA
2	s	d	s	–
3	d	s	s	–
4	d	d	s	–
5	s	s	d	–
6	s	d	d	–
7	d	s	d	–
8	d	d	d	QDA

LDA, linear discriminant analysis; QDA, quadratic discriminant analysis.

Clearly, the two extreme cases in which a_k, Q_k and U_k are either all identical, Case 1 in Table 1, or all different, Case 8 in Table 1, across the K classes correspond to linear discriminant analysis and quadratic discriminant analysis. In this paper, we focus on the four cases in which the class covariance matrices Σ_k have the same orientation, that is $U_k = U$ for all $k = 1, \ldots, K$. These correspond to the cases numbered one to four in Table 1 and are also known collectively as the common principal component model (Flury, 1988).

In what follows, we will use $S_k = \hat{U}\hat{\Lambda}_k\hat{U}^{\mathrm{T}} = \hat{a}_k\hat{U}\hat{Q}_k\hat{U}^{\mathrm{T}}$ to denote the empirical estimate of Σ_k under the common principal component model. Flury (1988) gives details on how to obtain maximum likelihood estimates of these quantities.

DEFINITION 1. *A vector $w \in \mathscr{R}^d$ is called a weighting if $\sum_{j=1}^d w_j = 1$ and $w_j \geqslant 0$, for all j.*

DEFINITION 2. *Let $S_k = \hat{U}\hat{\Lambda}_k\hat{U}^{\mathrm{T}}$ be the empirical estimate of Σ_k under the common principal component model. We define the average eigenvalue of S_k with respect to the weighting w as*

$$\phi_k(w) \equiv \sum_{j=1}^d w_j \hat{\lambda}_{jk}.$$

DEFINITION 3. *Let $S_k = \hat{U}\hat{\Lambda}_k\hat{U}^{\mathrm{T}}$ be the empirical estimate of Σ_k under the common principal component model. Then the estimated common eigenvectors $\hat{u}_i$ and $\hat{u}_j$ are dissimilar with respect to the weighting w if*

$$\sum_k \left(\frac{n_k}{N}\right)\left\{\frac{\hat{\lambda}_{ik}}{\phi_k(w)}\right\} \neq \sum_k \left(\frac{n_k}{N}\right)\left\{\frac{\hat{\lambda}_{jk}}{\phi_k(w)}\right\}. \tag{5}$$

They are uniformly dissimilar if they are dissimilar with respect to all weightings $w \in \mathscr{R}^d$.

THEOREM 1. *Let $S_k = \hat{U}\hat{\Lambda}_k\hat{U}^{\mathrm{T}}$ be the empirical estimate of Σ_k under the common principal component model. If the estimated common eigenvectors $\hat{u}_i$ and $\hat{u}_j$ are uniformly dissimilar for all $i \neq j$, then* LR(α) *as in equation* (4) *is maximised by the common eigenvector $\hat{u}_j$ for which*

$$\sum_{k=1}^K \left(\frac{n_k}{N}\right)(-\log \hat{\lambda}_{jk}) \tag{6}$$

is the largest.

Proof. Since $S_k = \hat{U}\hat{\Lambda}_k\hat{U}^{\mathrm{T}}$, we can simply choose to work in the basis of $\{\hat{u}_1, \ldots, \hat{u}_d\}$. If we write out LR($\alpha$) explicitly in this case, our problem becomes

$$\max_{\alpha_j;\, j=1,\ldots,d} \sum_{k=1}^K \left(\frac{n_k}{N}\right)\left\{-\log\left(\sum_{j=1}^d \alpha_j^2 \hat{\lambda}_{jk}\right)\right\}$$

subject to $\sum_{j=1}^d \alpha_j^2 = 1$. The Lagrangian function for this constrained optimisation problem is

$$\sum_{k=1}^K \left(\frac{n_k}{N}\right)\left\{-\log\left(\sum_{j=1}^d \alpha_j^2 \hat{\lambda}_{jk}\right)\right\} + \theta\left(\sum_{j=1}^d \alpha_j^2 - 1\right)$$

and the first-order conditions are

$$\sum_k 2\left(\frac{n_k}{N}\right)\left(\frac{\alpha_j \hat{\lambda}_{jk}}{\sum_j \alpha_j^2 \hat{\lambda}_{jk}}\right) - 2\theta\alpha_j = 0$$

for all j, or

$$\alpha_j\left\{\sum_k \left(\frac{n_k}{N}\right)\left(\frac{\hat{\lambda}_{jk}}{\phi_k(\alpha^2)}\right) - \theta\right\} = 0$$

for all j, where $\phi_k(\alpha^2) \equiv \sum_j \alpha_j^2 \hat{\lambda}_{jk}$. Therefore, for every j, either $\alpha_j = 0$ or

$$\sum_k \left(\frac{n_k}{N}\right)\left\{\frac{\hat{\lambda}_{jk}}{\phi_k(\alpha^2)}\right\} = \theta. \tag{7}$$

Since the eigenvectors are uniformly dissimilar for every $j \neq i$, this means that there exists at the most one $j = j^*$ for which (7) can be true. The fact that $\sum_{j=1}^d \alpha_j^2 = 1$ means that the optimal α must be 1 at the j^*th position and 0 everywhere else. Plugging this back into the objective function, we obtain

$$\mathrm{LR}(\alpha) \propto \sum_{k=1}^K \left(\frac{n_k}{N}\right)(-\log \hat{\lambda}_{j^*k}).$$

Obviously, the correct j^* that maximises this is

$$j^* = \arg\max_j \left\{ \sum_{k=1}^{K} \left(\frac{n_k}{N} \right) (-\log \hat{\lambda}_{jk}) \right\}. \qquad \square$$

The theorem tells us that if, under the common principal component model, the common eigenvectors are uniformly dissimilar then the function LR(α) can be maximised by searching over a finite number of candidates, namely the common eigenvectors $\hat{u}_1, \ldots, \hat{u}_d$; no iterative method is necessary, and we now have a closed-form solution for a more general case than just for the case corresponding to linear discriminant analysis. For the remaining Cases 5–8 in Table 1, of course, iterative methods are required.

3. Discussion

3·1. *Special cases*: $Q_k = Q$

If we further restrict the class covariance matrices to have the same shape, that is $Q_k = Q$ for all k, Cases 1 and 3 in Table 1, then $\Sigma_k = a_k U Q U^{\mathrm{T}} \equiv a_k Q_0$. This is sometimes called the proportional covariance model (Flury, 1986).

Corollary 1. *Let* $S_k = \hat{a}_k \hat{U} \hat{Q} \hat{U}^{\mathrm{T}} = \hat{a}_k \hat{Q}_0$ *be the empirical estimate of* Σ_k *under the proportional covariance model. Then* LR(α) *as in equation* (4) *is maximised by the common eigenvector* $\hat{u}_j$ *for which* $\hat{q}_j$ *is the smallest.*

Proof. Under the extra condition, $\hat{\Lambda}_k = \hat{a}_k \hat{Q}$, as specified by the proportional covariance model, (6) becomes

$$\begin{aligned} \sum_{k=1}^{K} \left(\frac{n_k}{N} \right) (-\log \hat{\lambda}_{jk}) &= - \sum_{k=1}^{K} \left(\frac{n_k}{N} \right) \{ \log(\hat{a}_k) + \log(\hat{q}_j) \} \\ &= -\log(\hat{q}_j) + \text{a constant not depending on } j, \end{aligned}$$

which clearly is maximised by the smallest $\hat{q}_j$. $\square$

Small eigen-directions are often regarded as containing little or no information. The reason for the above seemingly counter-intuitive result is that we have assumed that the data are pre-standardised to have the total sample covariance matrix S equal to I. Apart from scaling factors, the total sample covariance matrix S is just the 'sum' of the between- and within-class covariance matrices, B and W. If $S_k = \hat{a}_k \hat{Q}_0$, then $W = \bar{a} \hat{Q}_0$, where $\bar{a}$ is the average size of the S_k's. It is then easy to see that the smallest eigen-direction of $\hat{Q}_0$, and hence of W, must contain the largest amount of between-class separation if B and W must 'sum' up to a fixed constant.

3·2. *Uniform dissimilarity*

We now provide some intuition for how the notion of uniform dissimilarity can be understood. Note that $\phi_k(w)$ is a weighted average of the eigenvalues within class k, averaged across all directions; hence $\hat{\lambda}_{jk}/\phi_k(w)$ can be regarded as a standardised eigenvalue. Then

$$\sum_k \left(\frac{n_k}{N} \right) \left\{ \frac{\hat{\lambda}_{jk}}{\phi_k(w)} \right\}$$

can be regarded as a weighted average of the standardised eigenvalues in the jth eigen-direction, averaged between classes. The notion of uniform dissimilarity, therefore, can be taken to mean the following: no matter how you standardise the eigenvalues within each class, the average eigenvalues are still different from one dimension to another when averaged between classes.

4. APPLICATIONS

4·1. *A simulated example*

The results presented above can be used to answer a common question that has been raised by a number of readers regarding an example used by Zhu & Hastie (2003) to compare the criterion LR(α) and the sliced average variance estimator, an alternative criterion proposed by Cook & Yin (2001) for finding discriminant directions.

Zhu & Hastie (2003) constructed a simple two-dimensional problem in which two classes, with equal sample sizes, $n_1 = n_2 = N/2$, differ in the mean but have the same marginal variance in the direction x_1 whereas, in the direction x_2, they have the same mean but different marginal variances. The two classes were constructed to have sample means equal to

$$\mu_1 = \begin{pmatrix} -\sqrt{0{\cdot}5} \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} \sqrt{0{\cdot}5} \\ 0 \end{pmatrix};$$

and sample covariance matrices equal to

$$S_1 = \begin{pmatrix} 0{\cdot}5 & 0 \\ 0 & 0{\cdot}3 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 0{\cdot}5 & 0 \\ 0 & 1{\cdot}7 \end{pmatrix}.$$

Zhu & Hastie (2003) then evaluated the different criterion functions at x_1 and x_2 and found that the Cook & Yin method would pick x_2 whereas their method would pick x_1 as the best discriminant direction.

A number of readers have claimed that it was not fair just to compare the different criterion functions at x_1 and x_2 alone because it was not obvious that maximising LR(α) would not result in a third direction, neither x_1 nor x_2, as the best discriminant direction. However, it is easy to see that the common principal component model applies to this example; moreover x_1 and x_2 are the common eigen-directions. Our result above then implies that it suffices to consider just these two directions provided that the two directions are uniformly dissimilar. It is easy to check that this is the case, because there is no w between 0 and 1 such that

$$\frac{0{\cdot}5}{0{\cdot}5w + 0{\cdot}3(1-w)} + \frac{0{\cdot}5}{0{\cdot}5w + 1{\cdot}7(1-w)} = \frac{0{\cdot}3}{0{\cdot}5w + 0{\cdot}3(1-w)} + \frac{1{\cdot}7}{0{\cdot}5w + 1{\cdot}7(1-w)}.$$

Therefore, instead of maximising LR(α) iteratively, we only need to compare the two eigenvectors, in this case x_1 and x_2, according to

$$e_1 = \tfrac{1}{2}(-\log 0{\cdot}5 - \log 0{\cdot}5) \simeq 0{\cdot}69, \quad e_2 = \tfrac{1}{2}(-\log 0{\cdot}3 - \log 1{\cdot}7) \simeq 0{\cdot}34.$$

Since $e_1 > e_2$, the conclusion is that x_1 is the best discriminant direction.

4·2. *Swiss banknotes data*

Flury (1988, Ch. 4) studied an interesting dataset known now as the Swiss banknotes dataset, which can be obtained from a library called `ncomplete` as part of the R package (R Development Core Team, 2004). The dataset consists of six measurements made on 100 genuine and 100 forged Swiss banknotes. The measured variables are as follows: X_1, width of the banknote; X_2, height of the left-hand side of the banknote; X_3, height of the right-hand side of the banknote; X_4, distance between the top of the inner box to the upper border; X_5, distance between the bottom of the inner box to the lower border; X_6, diagonal of the inner box.

Flury (1988) tells us that this dataset satisfies the common principal component model and provides us with an algorithm, called the Flury–Gautschi algorithm, for finding the estimates of the common eigenvectors $\hat{u}_1, \ldots, \hat{u}_6$ as well as the two sets of eigenvalues $\hat{\lambda}_{jk}$ for $j = 1, \ldots, 6$ and $k = 1, 2$. Once the $\hat{\lambda}_{jk}$'s are computed, it is easy to check whether or not the uniform dissimilarity condition is satisfied for any two common eigenvectors $\hat{u}_i$ and $\hat{u}_j$ by verifying numerically that

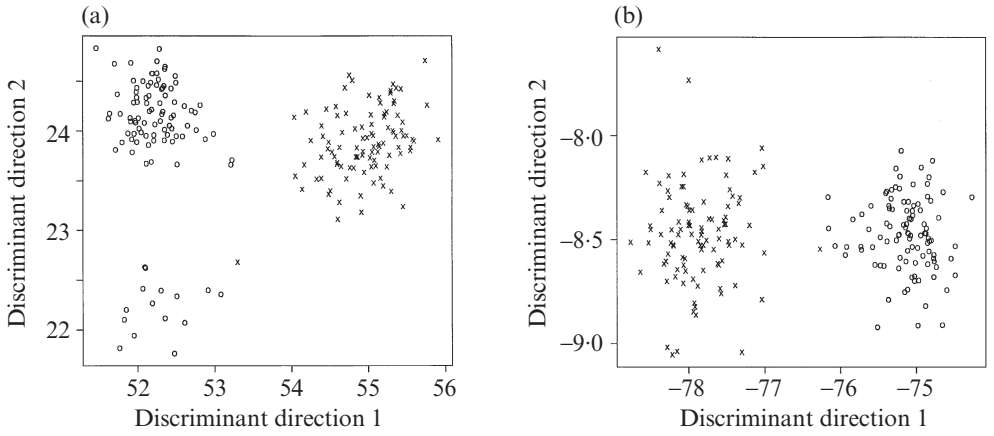

Fig. 1: Swiss banknotes data. Data projected on to the first two discriminant directions: (a) common principal components, (b) linear discriminant analysis. Genuine notes are coded with a circle whereas forged notes are coded with a cross. The second direction is meaningless for linear discriminant analysis.

there is no feasible solution for w to the equation

$$\sum_k \left(\frac{n_k}{N}\right)\left\{\frac{\hat{\lambda}_{ik}}{\phi_k(w)}\right\} - \sum_k \left(\frac{n_k}{N}\right)\left\{\frac{\hat{\lambda}_{jk}}{\phi_k(w)}\right\} = 0.$$

Figure 1 shows the data projected on to the first two discriminant directions. Genuine notes are coded with a circle whereas forged notes are coded with a cross. In Fig. 1(a), the discriminant directions are the common principal components, reordered according to criterion (6). In Fig. 1(b), the first direction is the maximiser of (2) whereas the second direction is meaningless because the between-class covariance matrix has rank 1 for a two-class problem; it is used here just so that we can draw a scatterplot.

Table 2 shows that, aside from an inconsequential scaling factor, the first direction found by maximising LR(α) under the common principal component model is very close to the linear discriminant direction. Using just the first discriminant direction or simply linear discriminant analysis, we see that most forged notes can be detected relatively easily, except for one particular case, which seems to be a particularly well-made forgery.

Here the second direction displayed in Fig. 1(a) is especially useful. On this scale, it appears that the genuine notes can be further divided into two different types, those appearing in the upper left corner of the plot, which we call type I, and those appearing in the lower left corner, which we

Table 2: *Swiss banknotes data. Normalised loadings of the first discriminant directions.* 'LR-CPC' *refers to maximising* LR(α) *under the common principal component model*; 'LDA' *refers to linear discriminant analysis*

Dimension	LR-CPC	LDA
1	−0·04	0·00
2	0·28	−0·33
3	−0·39	0·33
4	−0·41	0·44
5	−0·47	0·46
6	0·62	−0·61

call type II. Clearly, the well-made forgery conforms to type II, which may suggest a line of criminal investigation!

One could also ignore the fact that the common principal component model is satisfied for these data and try to find disciminant directions by recursively maximising criterion (4) directly. As mentioned earlier in § 1, numerous local solutions often exist. In this case, the underlying Newton-type algorithm actually fails to converge for some starting values, giving us completely useless 'solutions'. Therefore, if the common principal component model holds, it is much better to take full advantage of it.

ACKNOWLEDGEMENT

This research is partially supported by the Natural Science and Engineering Research Council of Canada. The author would like to thank Professor Trevor Hastie for introducing him to Flury's work on common principal component models, a referee for his encouragement and thought-provoking comments, and the editor for his editorial advice.

REFERENCES

BANFIELD, J. & RAFTERY, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–21.

BENSMAIL, H. & CELEUX, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J. Am. Statist. Assoc.* **91**, 1743–8.

COOK, R. D. & YIN, X. (2001). Dimension reduction and visualization in discriminant analysis (with Discussion). *Aust. New Zeal. J. Statist.* **43**, 147–99.

FLURY, B. (1986). Proportionality of k covariance matrices. *Statist. Prob. Lett.* **4**, 29–33.

FLURY, B. (1988). *Common Principal Components and Related Multivariate Models*. New York: Wiley.

HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning: Data-Mining, Inference and Prediction*. New York: Springer-Verlag.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.

R DEVELOPMENT CORE TEAM (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; URL http://www.R-project.org.

ZHU, M. & HASTIE, T. J. (2003). Feature extraction for nonparametric discriminant analysis. *J. Comp. Graph. Statist.* **12**, 101–20.

[*Received October* 2005. *Revised January* 2006]