Penalized Discriminant Analysis

Trevor Hastie
Statistics and Data Analysis Research Department
AT&T Bell Laboratories
Murray Hill, New Jersey

Andreas Buja
Statistics and Data Analysis Research Group
Bellcore
Morristown, New Jersey

Robert Tibshirani
Department of Preventive Medicine and Biostatistics
and Department of Statistics
University of Toronto
Toronto, Ontario

May 25, 1994
©AT&T Bell Laboratories

Abstract

Fisher's linear discriminant analysis (LDA) is a popular data-analytic tool for studying the relationship between a set of predictors and a categorical response. In this paper we describe a penalized version of LDA. It is designed for situations in which there are many highly correlated predictors, such as those obtained by discretizing a function, or the greyscale values of the pixels in a series of images. In cases such as these it is natural, efficient, and sometimes essential to impose a spatial smoothness constraint on the coefficients, both for improved prediction performance and interpretability. We cast the classification problem into a regression framework via optimal scoring. Using this, our proposal facilitates the use of any penalized regression technique in the classification setting. The technique is illustrated with examples in speech recognition and handwritten character recognition.

AMS 1991 Classifications: Primary 62H30, Secondary 62G07

1 Introduction

Linear discriminant analysis (LDA) is a standard tool both for classification and dimension reduction. Here is roughly how LDA works:

- Classification is based on the Mahalanobis distances $(x - m^j)\Sigma_W^{-1}(x - m^j)$ to class centroids m_j, where Σ_W is the pooled within-class covariance of the predictors x.

- Data reduction entails a sequence of unit-variance, linear discriminant variables $\beta_k^T x$, chosen to successively maximize $\beta_k^T \Sigma_{Bet} \beta_k$, with Σ_{Bet} the between-class covariance matrix. These discriminant variables represent a subspace for which the class centroids are spread out as much as possible.

LDA enjoys a number of favorable properties, such as reasonable robustness to non-normality and even to mildly different class covariances (see the many references in Seber 1984, chap. 6).On the down side, there are two deficiencies which apply under opposite circumstances:

- LDA is too flexible in situations with large numbers (e.g., hundreds) of highly correlated predictor variables.
- It is too rigid in situations where the class boundaries in predictor space are complex and nonlinear.

In the first case, LDA overfits, in the second case, LDA underfits the data.

We show that both problems can be overcome by modifications of LDA that effectively regularize a large, nearly or fully degenerate within-class covariance matrix Σ_W. This paper focuses on the first situation where LDA has problems — large numbers of correlated predictor variables. In a companion paper (Hastie, Tibshirani & Buja 1992), we address the other problem, interestingly by using results of this paper.

We call the technique discussed here *penalized discriminant analysis* or PDA for short. We rely on the well-known relationship between linear discriminant analysis and canonical correlation analysis, or more precisely its asymmetric cousin, here called optimal scoring. If viewed appropriately, optimal scoring contains *linear regression* as a building block. This building block lends itself to generalization by replacing linear least squares regression with other types of regression. In the present paper, we use *penalized least squares regression* to overcome the problems of high-dimensional (e.g. > 200) correlated predictors, while in the companion paper we use nonparametric adaptive regression methods to solve the problem of complex class boundaries when the predictors are relatively low-dimensional (probably no more than 30).

Along with the transfer of regression methods to discriminant analysis goes the transfer of related notions. For example, we can now use degrees of freedom in classification problems; these arise naturally when selecting or comparing smoothing parameters in fixed-bandwidth non-adaptive regression methods for optimal scoring.

In this paper we study two examples with large numbers of correlated predictor variables:

1. In speech recognition, one is interested in classifying short speech frames into one of several phoneme classes, based on the log-periodogram of the frame (for example); a typical log-periodogram estimate forms a predictor vector of dimension 256.
2. In handwritten character recognition, one wishes to classify small images of characters and digits into the obvious classes; a typical size-normalized image might be represented by 16 by 16 grayscale pixels, and again form a predictor of dimension 256.

The data in both these examples arise from the discretization of analog signals. It is obvious that an empirical (within-class) covariance matrix Σ_W of size 256 by 256 will have unfavorable statistical properties for almost any realistic size of the training sample, and straightforward Mahalanobis distances calculated in LDA will suffer or become useless as a result. Although we used a discretization resolution of 256 in both cases, one could use a higher resolution and the problems would be worse. Some form of "borrowing strength" or regularization appears to be needed, and can be formally motivated in the context of a population model (see section 4 as well as Leurgans, Moyeed & Silverman (1993)).

Even if we have sufficient data the estimates are likely to be spatially rough. To see this consider the usual linear discriminant functions with coefficient vectors $\Sigma_W^{-1} m^j$, where m^j is the mean vector for the jth class. If adjacent predictor (log-spectral or grayscale) values have strong positive correlations due to locally smooth behavior, then the spectral decomposition of Σ_W will favor low-frequency functions (large eigenvalues for smooth eigenvectors, small eigenvalues for rough eigenvectors). Σ_W^{-1} will have the inverse structure, and hence $\Sigma_W^{-1} m^j$ will emphasize the rough components of m^j. This phenomenon has two undesirable consequences: 1) rough coefficient contours that lack smoothness on the index domain are not interpretable, and 2) misclassification rates deteriorate since jagged coefficient contours indicate reliance on irrelevant local contrasts that are easily thrown off if the behavior of a test predictor vector deviates slightly from those found in the training sample.

Similar problems have been recognized by several authors (Di Pillo 1976, 1979; Campbell 1980; Friedman 1989) who introduce ridge-type regularizations of LDA. The obvious idea is to stabilize the (within-class) covariance matrix by adding a diagonal matrix, often a small multiple of the identity. The relation of this approach to ridge regression goes beyond a formal analogy due to the above mentioned relation between LDA and optimal scoring: subjecting the regression building block of optimal scoring to a ridge-type modification gives essentially the regularized version of LDA proposed by these authors. Also equivalent is Vinod's (1976) ridge-regularization of canonical correlation analysis when applied to a discriminant context (see also Vinod and Ullah, 1981). Vinod's aim was to counter the detrimental effects of collinear sets of variables on canonical correlation analysis.

It appears, though, that the examples of speech and character recognition call for a different type of regularization. We often wish to interpret the coefficient arrays of discriminant variables, in particular in the two cases above. These coefficient arrays can be interpreted as discretized curves and images since they, too, are indexed by frequency and pixel location, respectively. In order to achieve spatially smooth behavior, the usual ridge method does not seem entirely appropriate since its bias towards an overall mean ignores the spatial structure of the index domain. It may therefore be more plausible to bias the discriminant coefficients toward smoothness as a function of the frequency or the pixel location, e.g., through regularization along the lines of smoothing spline models. Smoothing splines enforce smoothness by penalizing local contrasts such as second order differences. That is, the coefficient array of a discriminant variable pays a price for locally rough behavior as a function of frequency or pixel location.

To summarize, two distinct motivations for regularization have emerged:

- When the number-of-variables to sample-size ratio is too high, we cannot reliably estimate a covariance matrix. Since the large number of variables arise from discretizing an analog signal, natural methods of regularization can be found.

- Even if the sample size were sufficient to estimate the structured covariance matrix, coefficients of spatially smooth variables tend to be spatially rough. Since we hope to interpret these coefficients, we would prefer smoother versions , especially if they do not compromise the fit.

Place figure 1 around here

Our proposal replaces Σ_W by a regularized version $\Sigma_W + \lambda\Omega$, where Ω is a "roughness" type penalty matrix; the LDA analysis then proceeds as usual. Figures 1 and 2 show the results for the two examples. In each we see the raw LDA coefficients displayed as curves and images respectively, as well as the corresponding regularized versions. In both cases the regularized versions improved classification performance on test data significantly. Both these examples are treated in detail in sections 5 and 6.

The possibility of using other than ridge-type penalties was mentioned in passing by Friedman (1989, end of Sec. 8). He proposes to penalize local averages to enforce smoothness, which is distinct from our proposal since the splines we use penalize local differences. We do not know whether Friedman's proposal achieves what it sets out to do. The thrust of his paper is to combine ridge-shrinkage, LDA and quadratic discriminant analysis (QDA) in a single framework. That is, he considers problems of lower dimensionality than we do since QDA estimates within-class covariances individually for each class, thus further compounding the problem caused by high dimensionality.

Place figure 2 around here

A different and quite complex approach is taken by Kiiveri (1992) whose motivation also arose from the analysis of high-dimensional spectral data. He assumes a factor analysis/growth curve model for the within-class covariance: $\Sigma_W = \Lambda\Phi\Lambda^T + \Psi$, where the first term is structural and of low-rank, and Ψ is typically diagonal stemming from uncorrelated errors. Based on normal assumptions, the components are estimated with an EM algorithm. As an aside, Kiiveri (1992, end of Sec. 2.2) mentions the possibility of using a stationary rather than diagonal form for Ψ. His Ψ is reminiscent of a penalty, but its actual role is that of a factor analytic "uniqueness". He seems exclusively interested in interpretability of canonical variates since he does not mention misclassification rates.

Related to the present work is Leurgans et al. (1993) who study canonical-correlation analysis of pairs of discretized analog signals. They also use a penalty approach and they provide some asymptotic theory. More details are given in Section 4.

We should also mention an approach that seems obvious but does not work: motivated by the fact that log-spectra are often smoothed before further processing, one might argue that LDA should be applied to smoothed log-spectra, and, by generalization, to smoothed images. The problem with this approach is that smoothing adds to already existing correlations among neighboring log-spectral and gray-scale values, thus causing the canonical variate arrays to be even more jagged. The within-class covariance of smoothed data will be even more degenerate than that of the raw data, thus calling for even more regularization. Thus we see that regularization at the level of the within-class covariance is indispensable. Pre-smoothing then turns out to be superfluous, adding to the computational cost of PDA with zero benefit in classification performance.

Another approach to regularization is *filtering*, popular in the signal and speech processing literature. Here the spatial structure of the predictors is acknowledged by approximating the data by its projection onto a low-dimensional, spatially-smooth basis. The predictors are thus replaced by their basis coefficients. A drawback of this approach is that one has to supply a suitable basis, which tends to involve more subjectivity than choosing a smoothing penalty. In our examples the filtering approach produced results similar but slightly inferior to those achieved by regularization.

2 Notations and Conventions

Consider a discrimination problem with J classes and N training samples. The training samples consist of observed predictor vectors and their known class memberships. These memberships will be represented by a categorical response variable G with J levels; the individual responses by g_i. It is sometimes convenient to code the N responses in terms of an indicator matrix Y of size $N \times J$, with columns that correspond to the dummy-variable codings of the J classes. We denote by h_i the m-vector of predictor values (input features) for the ith training sample. In the digit-recognition example, h_i is a 256-vector of grayscale pixel measurements for the ith training image. In the context of flexible discriminant analysis (Hastie et al. 1992), the h_i will be expanded bases (splines, polynomials,...) of observed predictor variables x_i, that is $h_i = h(x_i)$. The vectors h_i for all the training observations will be stored as rows of the $N \times m$ matrix H, which we call the predictor matrix. We will assume that the columns of H are centered (i.e., orthogonal to the constant one vector). If this is assumption is not satisfied, the constant one vector should be an element of the column space of H, in which case the eigenproblems encountered below will exhibit a trivial eigenvalue which is easily weeded out.

We assume that some penalty matrix Ω of size $m \times m$ is given, for penalizing the coefficients of h. It should be symmetric and non-negative definite. If a smoothing parameter λ is part of the penalty, we assume that it is absorbed in Ω. These penalty matrices arise naturally in the context of the examples. If the data are log-spectra or images, we define Ω in such a way so as to force nearby components of β_k to be similar. If $h_i = h(x_i)$ represent a basis expansion of the actual input vectors x_i, a penalty matrix may be chosen so that the compositions $h(x_i)^T \beta_k$ are smooth as functions of x. See Sections 5, 6 and 7.

3 The Equivalence of Penalized Linear Discriminant Analysis, Canonical Correlation Analysis and Optimal Scoring

It is known that discriminant variates are up to scale factors the same as the so-called "canonical variates" which result from an associated canonical correlation analysis (CCA), and often the latter term is used interchangeably with discriminant variates. Somewhat lesser known is that an asymmetric version of canonical correlation analysis, here called "optimal scoring" and abbreviated "OS", also yields a set of dimensions which coincide up to scalars with those of LDA and CCA.

In the following subsections we introduce OS, CCA and LDA with penalization built in. We then show that the three are equivalent just as they are without penalization. This has the important consequence that we can simply use our tools for penalized and nonparametric regression to perform penalized and nonparametric discriminant analysis.

We also discuss the dimension reduction aspect of linear discriminant analysis. Dimension reduction means re-expressing the data in fewer variables while minimizing the loss of essential information for the problem at hand. Such reduction can actually be beneficial when the "lost dimensions" show only spurious or weak structure. The reduced dimensions resulting from LDA are variously called "discriminant variates", "discriminant coordinates", or sometimes "crimcoords" for short.

Each of the three problems (OS, CCA, LDA) to be defined below has an associated criterion and constraints under which the criterion is to be optimized or made stationary.

3.1 Penalized Optimal Scoring

The point of optimal scoring is to turn categorical variables into quantitative ones by assigning scores to classes (groups, categories). In our notation above $\theta(G)$ assigns a real number, say θ_j, to the jth level of G. Given a J-vector of such scores θ for the J classes, the N-vector $Y\theta$ represents a vector of scored training data which one may try to regress onto the predictor matrix H.[1] The simultaneous estimation of scores and regressions constitutes the optimal scoring problem:

Definition 1 *The penalized optimal scoring problem is defined by the criterion*

$$\begin{aligned} ASR(\theta, \beta) &= N^{-1}\left(\sum_{i=1}^{N}[\theta(g_i) - h(x_i)^T\beta]^2 + \beta^T\Omega\beta\right) && (1) \\ &= N^{-1}\left(\|Y\theta - H\beta\|^2 + \beta^T\Omega\beta\right) && (2) \end{aligned}$$

which is to be minimized (made stationary) under the constraint $N^{-1}\|Y\theta\|^2 = 1$.

Although definition 1 is stated in terms of a single solution (θ, β), implicit is a sequence of solutions (θ_k, β_k) with orthogonality defined by the implied inner-product $N^{-1}\langle Y\theta_k, Y\theta_\ell\rangle = \delta_{k\ell}$. We use this compact notation in the subsequent definitions as well, and thus avoid a more cumbersome notation involving traces of matrices.

Optimal scoring (or scaling) is common in correspondence analysis (e.g., Lebart et al. 1984; chap. III for a comparison with discriminant analysis) and the psychometric literature (e.g., Gifi 1981, 1990, in particular chap. 7.2 for discriminant analysis; and the series of papers by de Leeuw, Takane, Young, in various permutations: 1976, 1978, 1979).

It is useful to interpret the criterion (2) as a quadratic form in the combined θ-β vector:

$$ASR(\theta, \beta) = \theta^T\Sigma_{11}\theta - 2\theta^T\Sigma_{12}\beta + \beta^T\Sigma_{22}\beta \tag{3}$$

where the obvious definitions are:

- $\Sigma_{11} = N^{-1}Y^TY$, a diagonal matrix with the class proportions $p_i = N_i/N$ in the diagonal,
- $\Sigma_{22} = N^{-1}(H^TH + \Omega)$, the penalized covariance matrix of the predictors,
- $\Sigma_{12} = N^{-1}H^TY, \quad \Sigma_{21} = \Sigma_{12}^T$.

If we assume that all classes are non-empty ($N_j > 0$), Σ_{11} is invertible. As far as Σ_{22} is concerned, we must assume that the penalty Ω is chosen intelligently so as to prevent exact or near collinearities among the predictors, i.e., Σ_{22} should be invertible, too.

For a given score vector θ, the minimizing β for the OS problem is the penalized least squares estimate:

$$\beta_{\mathrm{OS}} = (H^TH + \Omega)^{-1}H^TY\theta = \Sigma_{22}^{-1}\Sigma_{21}\theta, \tag{4}$$

and the partially minimized criterion becomes:

$$\min_{\beta} ASR(\theta, \beta) = 1 - N^{-1}\theta^TY^TS(\Omega)Y\theta = 1 - \theta^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\theta \tag{5}$$

[1]With a slight abuse of notation, we use θ to represent both the function or the vector of J real numbers that represent the function.

where $S(\Omega) = H(H^T H + \Omega)^{-1} H^T$ denotes the "hat" or "smoother" matrix of H regularized by Ω. This form of the OS problem is computationally useful because it shows that one can run a penalized multi-response regression of Y onto H once: $\hat{Y} = S(\Omega)Y$, and then eigen-analyze $Y^T\hat{Y}$ to obtain the stationary θ vectors. In section 7 we discuss various forms for $S(\Omega)$. The intent of the following sections is to translate the results of OS analysis into a LDA form.

3.2 Penalized Canonical Correlation Analysis

Definition 2 *The penalized canonical correlation problem is defined by the criterion*

$$COR(\theta, \beta) = \theta^T \Sigma_{12} \beta$$

which is to be maximized (made stationary) under the constraints

$$\theta^T \Sigma_{11} \theta = 1 \qquad and \qquad \beta^T \Sigma_{22} \beta = 1$$

Formally, this looks like the usual canonical correlation problem as applied to linear discrimination, except for the penalization built into Σ_{22}.

The criteria of optimal scoring and canonical correlation analysis are related to each other by (3): under the CCA constraints, $ASR = 2 - COR$, which shows that the two problems differ only in the additional constraint on β which is missing in OS. For a given set of scores θ, the maximizing β for the CCA problem is, up to a scalar, the same as the minimizing β (4) for the OS problem:

$$\beta_{\mathrm{CCA}} = \beta_{\mathrm{OS}} / \sqrt{\beta_{\mathrm{OS}}^T \Sigma_{22} \beta_{\mathrm{OS}}}, \tag{6}$$

which entails:

$$\max_{\beta^T \Sigma_{22} \beta = 1} COR(\theta, \beta) = (\theta^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \theta)^{\frac{1}{2}}. \tag{7}$$

A comparison with (5) shows that the OS and CCA problems produce identical stationary score vectors θ.

3.3 Penalized Linear Discriminant Analysis

LDA finds linear combinations of the predictors that maximize the between class variance (of the class means), relative to the within class variance. To introduce the *penalized* linear discriminant problem, we need the customary decomposition of the *total covariance* Σ_{22} into *between-class covariance* and *within-class covariance* (or respective cross-products, if the columns of the expanded predictor matrix H are not centered). The between-class covariance is the covariance of H regressed onto Y, or, equivalently, the class-weighted covariance of the class means. The within-class covariance is what's left, and is a pooled estimate of the common covariance matrix for the J classes. Let $P_Y = Y(Y^T Y)^{-1} Y^T$ be the projector onto Y-column space:

- $M = \Sigma_{11}^{-1} \Sigma_{12}$, a $J \times m$ matrix whose rows are the class means $m^j = ave\{h_i \; ; \; i \in Class\ j\}$: $M = (m^1, ..., m^J)^T$.
- $\Sigma_{Bet} = N^{-1}(P_Y H)^T (P_Y H) = \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} = M^T \Sigma_{11} M$
- $\Sigma_W = N^{-1}[((I - P_Y)H)^T (I - P_Y)H + \Omega] = \Sigma_{22} - \Sigma_{Bet}$

Accordingly, penalization affects only the within-class covariance.

Definition 3 *The criterion of the penalized linear discriminant problem is the (unpenalized) between-class variance:*

$$BVAR(\beta) = \beta^T \Sigma_{Bet} \beta,$$

which is to be maximized (made stationary) under a constraint on the penalized within-class variance:

$$WVAR(\beta) = \beta^T \Sigma_W \beta = 1.$$

The equivalence of the CCA and the LDA problem is standard (e.g., 11.5.4 of Mardia et al., 1979), but it will follow from the following section as well. The interpretation of this particular form of penalization is more easily made in the context of a particular example. Highly correlated predictors, such as in the case of discretized spectra, can meet the (unpenalized) normalization constraint by using negatively-correlated coefficients, since the constraint is in terms of the variance of the derived variable. Viewed as a function, the negatively correlated coefficients will appear wiggly; the appropriate penalty in this case would limit the spatial roughness of these coefficients.

3.4 Translation of Optimal Scoring Dimensions into Discriminant Coordinates

It is convenient to use CCA as a link between OS and LDA. CCA is a generalized singular value problem for Σ_{12} with regard to the metrics given by Σ_{11} and Σ_{22}. The associated singular value decomposition, essentially a collection of stationary solutions of the CCA problem, takes on this form:

$$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} = \Theta D_\alpha \mathrm{B}^T, \tag{8}$$

$$\Theta^T \Sigma_{11} \Theta = I_L, \tag{9}$$

$$\mathrm{B}^T \Sigma_{22} \mathrm{B} = I_L, \tag{10}$$

where $L = \min(J, m)$, Θ is a $J \times L$ matrix whose columns θ_k are left-stationary vectors, B is a $m \times L$ matrix whose columns β_k are right-stationary vectors, and D_α is a diagonal matrix of size $L \times L$ with non-negative diagonal elements α_k sorted in descending order.

The usual proof of (8) is by coordinate transformations $\theta = \Sigma_{11}^{-\frac{1}{2}} \theta'$ and $\beta = \Sigma_{22}^{-\frac{1}{2}} \beta'$ which bring (9) and (10) to a euclidean form, so the simple (non-generalized) SVD can be applied to $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$ (i.e., Σ_{12} expressed in the new coordinates). A simple SVD of the form $A = UDV^T$ entails the trivial consequences $AV = UD$, $A^T U = VD$, $U^T AV = D$, $V^T A^T AV = D^2$, $U^T AA^T U = D^2$; these are translated to the generalized SVD as follows:

$$\Sigma_{11}^{-1} \Sigma_{12} \mathrm{B} = \Theta D_\alpha, \tag{11}$$

$$\Sigma_{22}^{-1} \Sigma_{21} \Theta = \mathrm{B} D_\alpha, \tag{12}$$

$$\Theta^T \Sigma_{12} \mathrm{B} = D_\alpha, \tag{13}$$

$$\Theta^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Theta = D_{\alpha^2}, \tag{14}$$

$$\mathrm{B}^T \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \mathrm{B} = D_{\alpha^2}. \tag{15}$$

In particular, (13) implies $COR(\theta_k, \beta_k) = \alpha_k$.

According to (5) and (7) the stationary θ vectors of OS and CCA are the same, while the β vectors of OS and CCA are related according to (4) and (12) by

$$\mathrm{B}_{\mathrm{OS}} = \mathrm{B} D_{\alpha}^{T}, \tag{16}$$

B_{OS} being a matrix of OS-stationary column vectors $\beta_{OS,k}$. From (5) and (7) it follows that $ASR(\theta_k, \beta_k) = 1 - \alpha_k^2$.

To link CCA and LDA, we translate (15) directly into:

$$\mathrm{B}^T \Sigma_{Bet} \mathrm{B} = D_{\alpha^2}, \tag{17}$$

$$\mathrm{B}^T \Sigma_W \mathrm{B} = I_L - D_{\alpha^2} = D_{1-\alpha^2}. \tag{18}$$

This shows that B diagonalizes both Σ_{Bet} and Σ_W. If we define,

$$\mathrm{B}_{\mathrm{LDA}} = \mathrm{B} D_{(1-\alpha^2)^{-\frac{1}{2}}}, \tag{19}$$

we get a matrix whose columns $\beta_{LDA,k}$ are stationary solutions of the LDA problem:

$$\mathrm{B}_{\mathrm{LDA}}^T \Sigma_W \mathrm{B}_{\mathrm{LDA}} = I_L, \qquad \mathrm{B}_{\mathrm{LDA}}^T \Sigma_{Bet} \mathrm{B}_{\mathrm{LDA}} = D_{\alpha^2/(1-\alpha^2)}. \tag{20}$$

Finally, the relation between the LDA and OS solutions is

$$\mathrm{B}_{\mathrm{LDA}} = \mathrm{B}_{\mathrm{OS}} D_{[\alpha^2(1-\alpha^2)]^{-\frac{1}{2}}}. \tag{21}$$

3.5 Graphical Projections

The strongest discriminant coordinates are useful for graphical examination of both training and test data, the former to visually assess overlap of groups in predictor space, the latter to judge the strength of evidence for assigning test data to specific classes. The elements that need to be plotted are: projections of

1. the class centers m^j,
2. the predictor vectors h_i of the training data with an indication of their class memberships, and
3. the predictor vectors h of test data, if any.

We assume that the analysis has been performed by regression and eigenanalysis, i.e., we are given sets of scores Θ and sets of fits

$$\eta = \mathrm{B}_{\mathrm{OS}}^T h \tag{22}$$

for a given predictor vector h. It is convenient to translate these directly into the required projections. For example, η might have been produced by a complicated nonparametric regression model (for which we have a software algorithm) with many hidden basis functions; we prefer to avoid these and the innards of the software by simply asking for the fits, or predictions of η at new points.

Starting with the class means, we rewrite (11) using the definition of M and apply (19):

$$\Theta D_{\alpha} = M\mathrm{B} = M\mathrm{B}_{\mathrm{LDA}} D_{(1-\alpha^2)^{\frac{1}{2}}}, \tag{23}$$

and, denoting with θ^j the scores of class j: $\Theta = (\theta^1, ...\theta^J)^T$, we get

$$\mathrm{B}^T_{\mathrm{LDA}} m^j = \alpha_j/(1-\alpha_j^2)^{\frac{1}{2}}\theta^j. \tag{24}$$

For actual plotting, one peels off the first two or three dimensions, i.e., one uses the first two or three scores for each class; see figures 4 and 9 for an example.

To map sets of fits η to projected predictors $\mathrm{B}^T_{\mathrm{LDA}} h$, we only need to apply (21):

$$\mathrm{B}^T_{\mathrm{LDA}} h = D_{[\alpha^2(1-\alpha^2)]^{-\frac{1}{2}}} \mathrm{B}^T_{\mathrm{OS}} h = D_{[\alpha^2(1-\alpha^2)]^{-\frac{1}{2}}} \eta \tag{25}$$

Again, for plotting, one picks the first two or three dimensions, i.e., the first two or three elements from the set of fits.

In figure 6 we plotted the penalized class discriminant functions $\Sigma_W^{-1} m^j$. From (19) and (36) in the appendix, we see that $\Sigma_W^{-1} m^j = \mathrm{B}_{\mathrm{LDA}} \mathrm{B}^T_{\mathrm{LDA}} m^j$. Here we need $\mathrm{B}_{\mathrm{LDA}}$ or equivalently B_{OS}.

3.6 Distance Calculation and Classification

Like many other statistical procedures, LDA can be derived from suitable normal assumptions. In the present paper, these assumptions are not made, but the derivations from them are used as heuristic guides. The usual assumptions for LDA are that the predictor vectors follow multivariate normal distributions with different mean vectors but common covariance matrices among the classes. Assuming (unrealistically) that the parameters in this model are known, classification according to maximum posterior class probability results in a *nearest class mean* rule, where *"nearest"* is measured in terms of the shared within-class Mahalanobis distance. In practice, the rule is applied with parameters estimated from training data (m^j and the penalized Σ_W being our choices) and adjustments for unequal class sizes: classify a test predictor h as class j if

$$(h - m^j)^T \Sigma_W^{-1} (h - m^j) - 2\log p_j \tag{26}$$

is minimized over $j = 1, ..., J$. The last term is the class size adjustment ($p_j = N_j/N$), while the crucial first term is the (penalized) Mahalanobis distance $D(h, m^j)$. Sometimes the sample priors p_j do not reflect the population priors π_j (e.g. a stratified sample); in this case some external estimate of the π_j should be used.

In order to appreciate the effect of penalization on the metric in $D(h, m^j)$, write Σ_W^{-1} as $(W + \Omega)^{-1}$ where W is the unpenalized within covariance. Consider the deviations of two different observations from a centroid: O_1 being close in a "smooth" coordinate, far in a "rough", and O_2 being the opposite. We would prefer to say that O_1 is closer. However, W^{-1} will have large eigenvalues for rough directions, and will favor O_2; the penalty Ω will typically have the reverse eigenstructure to W and so cancels out this effect.

In the appendix, we first reconstruct the fact that classification can be done based on euclidean distance in the full space of discriminant variates $\mathrm{B}^T_{\mathrm{LDA}} h$ of h. Second, we show that the results of optimal scoring: the scores Θ and the fits $\mathrm{B}^T_{\mathrm{OS}} h$ are all that is needed for classification, a fact which we learned from Breiman & Ihaka (1984).This is very useful when we want the regression procedure to select the amount of smoothing, for it shows that the residual sum-of-squares in the regression problem is intimately related to the classification distance (Hastie et al. 1992). Our proofs are shorter and cover the case of penalized within-class covariances. Third, one can perform classification using distances in the $K < J-1$ reduced dimensional discriminant subspace by using the corresponding reduced set of regression fits and scores.

These results establish that the original class-adjusted Mahalanobis distance (26) is equivalent to each of the following distance measures (when used in a relative sense for classification):

1. $D(h, m^j) - 2\log p_j$
2. $\left\|\mathrm{B}_{\mathrm{LDA}}^T(h - m^j)\right\|^2 - 2\log p_j$
3. $\left\|D^{-1}_{\alpha(1-\alpha^2)^{\frac{1}{2}}}(\eta - \bar{\eta}^j)\right\|^2 - 2\log p_j$
4. $\left\|D_{(1-\alpha^2)^{-\frac{1}{2}}}(\theta^j - \eta)\right\|^2 - \left\|\theta^j\right\|^2 - 2\log p_j$
5. $\left\|D_{(1-\alpha^2)^{-\frac{1}{2}}}(\theta^j - \eta)\right\|^2 - 1/p_j - 2\log p_j$

Among these, only the last can not be used in a dimension-reduction mode since it relies on the presence of $J-1$ discriminant coordinates. In the third expression, $\bar{\eta}^j = \mathrm{B}^T m^j = D_\alpha^2 \theta^j$.

4 Model Considerations

Both the examples in this paper arise from discretized analog signals — cases where the data can be viewed as functions, sampled for computer representation. In the first example, the spectrum for an utterance is a function of frequency; in the second, the picture of the handwritten digit is a function of two spatial coordinates. As such the data for the ith sample can be interpreted as a discretized realization $h_{i,k} = h_i(s_k)$ of a stochastic predictor process $h(s)$ evaluated at discrete values s_k of a (continuous) index domain. We assume that the jth class

- is observed with relative frequency $\pi_j = P[G = j]$,
- its predictor process $h(s)$ has a class-specific mean function, $\mu_j(s) = E[h(s)|G = j]$, and
- a covariance function, $\Sigma(s,t) = E[(h(s) - \mu_j(s))(h(t) - \mu_j(t)|G = j]$, that is shared among the classes.

This provides us with a model underlying the methods described in this paper. It is natural to think of a functional version of LDA (Kiiveri 1992, section 1.2) in terms of this model. Ramsay & Dalzell (1991) coined the name *Functional Data Analysis* for this kind of problem.

The functional canonical variate problem in discriminant analysis consists of finding normalized functions $\beta_k(s)$ such that the associated functionals $\eta_k = \int \beta_k(s)h(s)ds$ have means that are optimally separated among the classes, i.e., $\sum_{j=1}^{J} \pi_j E(\eta_k|G = j)^2$ is maximized. The normalization has the form $\int\int \beta_k(s)\Sigma(s,t)\beta_\ell(t)dsdt = \delta_{k\ell}$.

For the functional classification problem in discriminant analysis, one assumes that the predictor processes are Gaussian. The Bayes optimal classification assigns a predictor function $h(s)$ to the class j that minimizes $D(h, \mu_j) - 2\log\pi_j$, where the Mahalanobis distance D is defined as

$$D(h,\mu_j) \;=\; \int\int [h(s) - \mu_j(s)]\Sigma^{-1}(s,t)[h(t) - \mu_j(t)]dsdt \; .$$

For this to be meaningful, one has to assume that an inverse Σ^{-1} of the covariance operator $\beta(.) \mapsto \int \Sigma(.,t)\beta(t)dt$ exists, but even if it exists it generally can not be represented by a kernel $\Sigma^{-1}(s,t)$. The double integral in the definition of $D(h, \mu_j)$ therefore has to be taken with a grain of salt.

For estimation in the functional model, we have to consider two levels of asymptotics:

1. we obtain N realizations h_i of the predictor process (sample size N),

2. we observe the realizations at m discrete locations s_k (resolution m).

Leurgans et al. (1993), in the context of canonical correlation analysis, focus on the first case and essentially view the second as a computational approximation of integrals by sums. For infinite-dimensional predictors $h(s)$ ($m = \infty$), they show that in order to obtain consistent estimates of $\beta(s)$ it is essential to regularize the sample estimate of the covariance function (see their Propositions 1 and 2). This is intuitively obvious, for with infinite resolution m and finite sample size N, we have always more variables than observations, and well-known degeneracies occur. They then derive the rates at which the amount of regularization should decrease to zero as N grows large in order to achieve consistent estimates of canonical variates. These results carry over to discriminant analysis with minor modifications: in their functional canonical variates problem, specialize the second process to a stochastic indicator vector indicating class membership, and remove penalization for this J-dimensional "class membership process".

Estimation of the Bayes optimal classification criterion poses an equally obvious problem: any estimate $\hat{\Sigma}(s,t)$ of the covariance function $\Sigma(s,t)$ will have rank no greater than the sample size N. It is therefore impossible to invert the estimated covariance operator in order to calculate the Mahalanobis distance, unless it is regularized with a suitable penalty. There is no necessity, however, to develop asymptotic theory for this form of the classification criterion since other equivalent forms of the criterion are based on canonical variates (section 3.6) and hence require inversion of the covariance operator only on a finite-dimensional subspace.

As a final remark, one might argue that asymptotic theory of estimation in a functional framework is not realistic since for all practical purposes a finite resolution has to be chosen. The argument could continue along the lines that the resolution should be selected commensurate with the sample size, and a realistic asymptotic theory should determine at what rate the resolution m can be increased as a function of the sample size N in order to achieve consistency. Such an approach would essentially use resolution as a regularization parameter. The problem with this approach is that the choice of resolution is and should be guided by computational feasibility rather than sample size: if we face a very small N, it might be feasible to choose a rather large m without running into problems of compute time and memory exhaustion. Low resolution amounts to throwing away data — a rather crude regularization method. It is more informative to make use of as much data as possible and derive regularization from substantive arguments, such as smoothness considerations on the index domain.

5 Example: Smooth Canonical Functions for Classifying Log-Periodograms

The following analysis is taken from joint work of Andreas Buja, Werner Stuetzle and Martin Maechler. The data in this example were extracted from the TIMIT database,[2] which is a widely used resource for research in speech recognition. We formed a small test problem by selecting five phonemes for classification based on digitized speech from this database. The phonemes are transcribed as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". From continuous speech of 50 male speakers, we selected 1000 speech frames of 32 msec duration, approximately 2 examples of each phoneme from each speaker. Each speech frame is represented by 512 samples at a 16kHz sampling rate, and each frame represents one of the above five phonemes. The breakdown of the 1000 speech frames into phoneme frequencies is as follows: aa ao dcl iy sh 147 226 167 269 191

sh	dcl	iy	aa	ao
191	167	269	147	467

From each speech frame, we computed a log-periodogram, which is one of several widely used methods for casting speech data in a form suitable for speech recognition. Thus the data used in what follows consist of 1000 log-periodograms of length 256, with known class (phoneme) memberships.

PLACE FIGURE 3 AROUND HERE

PLACE FIGURE 4 AROUND HERE

Figure 3 shows a sample of 10 log-periodograms in each phoneme class. It is known that periodograms have rather erratic statistical properties; the figure supports this. In order to turn them into reasonable estimates of an underlying spectral density function, a certain amount of smoothing is required. However, the interest in speech recognition is not faithful estimation of spectral densities but discrimination among speech units such as phonemes. We therefore envision a role for smoothing at the level of linear *functionals* on log-periodograms rather than smoothing the periodograms themselves.

If we think of log-periodogram vectors as functions $h(f)$ evaluated at finitely many frequencies f, we have the task of estimating linear functionals $\eta(h) = \int_f \beta(f)h(f)df$, approximated by $\sum_f h(f_j)\beta_j$, that best discriminate between the phonemes. The representers $\beta(f)$ and their discretized versions (the sets of linear coefficients) should be estimated in such a way that they depend smoothly on the frequency f, for reasons presented in sections 1 and 4. This can be achieved in an obvious way, for example, with a second derivative penalty on the coefficients $\beta(f)$: $\lambda \int[\beta''(f)]^2 df$

[2]TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce

(Wahba 1990). The penalty that we actually used for the present speech problem has the form $\lambda \int [\beta''(f)]^2 w(f) df$, where $w(f)$ was chosen to penalize the higher frequencies more. This is suggested by the experimental observation that the capability of the human ear to discern acoustic detail decreases rapidly for the frequencies above 1 kHz. In speech recognition, it is standard to use techniques with higher resolution in lower frequencies. Our weight function $w(f)$ is constant for values of f up to 1 kHz, and decreases linearly thereafter. We refer to this as an *improper spline* penalty; further details are given in section 7.

In figure 1 in section 1 we show the four canonical LDA coefficients $\beta(f)$ plotted as functions of a frequency scale (the plotting range 0-256 maps to 0-8 kHz). The jagged curves represent the raw LDA coefficients, while the smooth curves show penalized coefficients.

The overwhelming impression from these plots is of course the extremely wild behavior of the unpenalized LDA coefficients as compared to the penalized ones. Except for the very lowest frequencies, there is barely any structure that could be discerned by eye in the unpenalized curves. Although not visually evident in figure 3, there are positive correlations among log-periodogram values over large neighborhoods on the frequency scale. In fact, among more than 32,000 (256 choose 2, to be exact) pairwise within-class correlations of log-periodogram values, fewer than 5 percent were negative. For the reasons given in section 1, this leads to negatively correlated coefficients.

The penalized curves shown in figure 1 are easily recognized by their smoothness. In terms of fitted degrees of freedom, the reduction in the penalized coefficients is dramatic: while the unpenalized coefficients have 256 *df*, the penalized coefficients have only 30 *df*. (We computed *df* as the trace of the smoother matrix in the equivalent optimal scoring problem; further details in section 7.) The smoothness of the penalized curves allows interpretation of the coefficients as contrasts that set various frequency ranges against each other. The action is mostly in the low frequencies, as expected. Figure 4 shows that the 1st penalized crimcoord sets the phonemes *aa* and *ao* apart from the phonemes *dcl*, *iy* and *sh*. According to figures 1 and 3, this crimcoord relies on a peak (so-called formant) between 500 Hz and 1000 Hz (roughly frequencies 16 to 32) in the log-periodograms for *aa* and *ao*. According to figure 4, the second penalized crimcoord sets the phoneme *sh* against the phonemes *dcl* and *iy*. A look at figures 1 and 3 shows that this contrast picks up a peak near zero frequency in the log-periodograms for *dcl* and *iy* which is absent for *sh*.

PLACE FIGURE 5 AROUND HERE

So far, we discussed only qualitative features of penalized discriminant analysis for this data example, but the bottom line numbers, namely misclassification rates, are favorable for penalization as well.

In order to compute relevant numbers, we conducted a simulation study using an additional 387 speakers. From the combined set we randomly selected 50 speakers, fit both the LDA model and a variety of PDA models at different levels of regularization, and used them to classifiy the remaining examples. This was repeated fifty times, and the results are shown in figure 5. In the lower left panel we see boxplots of the test error for the improper spline as a function of *df*, with the rightmost entry corresponding to the unpenalized LDA. The minimum median of 0.073 is achieved around 30 *df*, and all are lower than that for unpenalized LDA (median 0.086). The amount of regularization needed depends on the size of the training set, as well as the purpose of the model (we might favor smoother coefficient functions for interpretability). The figure suggests that the test error rates of PDA are relatively insensitive to the amount of smoothing – any reasonable amount between 20 and 80 *df* does considerably better than no smoothing. By comparison, the top left panel shows the corresponding error rates on the training data. As expected, LDA overfits and does the best, while PDA achieves a median value slightly lower than the 0.073 it achieved on the test data.

Included in figure 5 are the results of some other approaches to PDA. The second column of figures was based on a hand-crafted filtering approach popular in speech research, which uses piecewise linear basis functions. The knot spacings are chosen to be uniform up to 1kHz, and then increasing logarithmically. Our piecewise linear weighting function $w(f)$ used with the improper spline penalty was chosen to correspond empirically with this schoice of knots. These filter bases are used to create linear combinations of the original predictors, and are then used in place of them. The misclassification results are comparable to those for the improper spline.

The third column was obtained by using an unstructured ridge penalty to regularize, which amounts to shrinking the within covariance matrix towards a scalar covariance matrix. Again the misclassification results are only slightly worse than for the first two columns.

All three approaches can be viewd as versions of PDA, enforcing spatial smoothness of the coefficient functions. The filter approach explicitly uses piecewise smooth bases to represent the coefficients. Although ridging appears to treat all coefficients equally, one can show that those contrasts corresponding to small eigenvalues of the total covariance are shrunk more than those corresponding to the large eigenvalues. Since the positive autocovariances lead to longer-trend dominant eigenfunctions, to some extent the ridging shrinks towards similarly behaved coefficient functions (Hastie & Mallows 1993). The improper-spline approach, on the other hand, explicitly biases towards smooth coefficient functions.

Place figure 6 around here

Figure 6 shows the *class discriminant functions* for the five classes, comparing LDA, our improper-spline PDA and ridge PDA, each at their optimal *df* according to figure 5. These are the equivalent versions of $\Sigma_W^{-1} m^j$ for linear discriminant analysis, and we discuss their construction further in section 3.5. These are distinguishing contrast coefficients for each phoneme. As expected the two "a" vowels are very similar. The ridge curve is very wiggly, but still manages to portray the main features. Since ridge shrinks towards the origin, far less shrinking was asked for (70 *df*) in order to allow the main features to appear; the improper spline, on the other hand, shrinks towards smoothness and could afford smoother functions. For space reasons we do not display the curves for the filter method, which look similar to those for the improper spline, albeit piecewise linear.

6 Example: Image Analysis and Character Recognition

A current "hot" topic in the field of pattern recognition is the automatic reading of handwritten addresses and zip-codes from envelopes. Le Cun et al. (1990) implemented a successful procedure which included as a primitive the classification of isolated handwritten digits. We used a subset of their data for training and testing, a sample of which are shown in figure 7.

Place figure 7 around here

Le Cun et al. (1990) normalized the binary images for size and orientation, resulting in 8-bit, 16×16 grayscale images. They used the 256 pixel values as the inputs to a multilayer-neural-network model, which is trained to classify the images. They report overall misclassification rates on test data of under 5%. The goal in this section is not to compete with their highly tuned procedure, but to show that a traditional statistical method, LDA, can be improved by spatial regularization. One could expect that any classification procedure that relies on coefficients at a pixel level (including neural networks) would similarly benefit from regularization.

Our approach here was to use the same normalized data as Le Cun et al. (kindly supplied by J. Bromley), but to use simpler and more classical approaches to discrimination. We used the first 2000 images as training data, and the following 2000 as a preliminary validation set.

As a first step we fit a standard LDA model to the training data. The misclassification rate was 3.1% on the training data, but 11% on the validation set. The training error is less than a third of the test error, and suggests we may be in an overfit situation, despite the simplicity of the technique. Indeed, the nine discriminant functions each have 256 coefficients, and despite the normalization constraints, resulted in over 2000 independent parameters!

Apart from the initial normalization, our procedure did not take advantage of the spatial correlation in the data. Even though we have 256 pixels per image—ostensibly 256 independent pieces of information—this spatial correlation suggests the real number is far less.

Figure 2 in section 1 consists of nine pairs of images; the left member of each pair represents the LDA coefficients as images, with the hope of gaining some insight into the important contrasts found. These salt-and-pepper images reveal no structure, and once again reflect the correlation between neighboring pixels, resulting in a strong negative correlation between coefficients.

We fit a PDA model using a Laplacian penalty $J(\beta)$ to constrain the coefficients to be spatially smooth. Denote the coefficients as functions $\beta(x, y)$, where x and y refer to horizontal and vertical coordinates in the two-dimensional image. Then

$$J(\beta) = \int_{x,y} \left[\frac{\partial^2 \beta}{\partial x^2} + \frac{\partial^2 \beta}{\partial y^2} \right]^2 dx\,dy, \tag{27}$$

where the term within square braces is the Laplacian. Further details are given in section 7.

Place figure 9 around here

The right member of each pair in figure 2 represents the PDA fit, where we chose the smoothing parameter to achieve 40 *df* for each coordinate. The misclassification rates are now 6.1% for the training data, and 8.2% for the test data; a 25% reduction in validation error using 85% fewer parameters!

In order to validate these results, we conducted a simulation study similar to that used in the previous example. We randomly sampled 2000 images from the combined set of 4000 as training images, and used the remainder as test images. This was repeated 50 times, and the results for several methods are summarized in figure 8. We have used about 40 *df* for all the examples. We tried different amounts of smoothing, and found that for less *df*, the test results uniformly got worse quite rapidly, and for more *df* they remained roughly the same or got worse very slowly. From these results it seems our initial division into training and test set yielded slightly pessimistic results.

Place figure 8 around here

The smoothness not only gives us improved misclassification rates, but offers interpretability as well. The first PDA coefficient in figure 2 looks like a white 0 with a vertical center darker than the remainder of the image. This contrast image can be expected to yield positive scores for 1s and negative scores for 0s. Figure 9 confirms this, and as we might expect 7s and 9s are on the same side as 1s, while 6s look more like 0s. The second PDA coefficient in figure 2 gives a positive score to images dark in the bottom left corner, and a negative score to images dark in the bottom middle and top left; 6s fall into the first category and 7s and 9s into the second, and both are confirmed in figure 9. The interpretation of the coefficients becomes more difficult for the lower order coefficients.

For the filtering approach we used a tensor-product basis of polynomials in each of the spatial coordinates, with total degree restricted to 8 and thus $m = 44$ basis functions. The images derived for this and the other methods were not visually informative so we have omitted them.

7 Penalized Regression Methods

In this section we give more details on the different forms of penalized regression outlined in section 1, used in the examples, and appearing in equations (5) and (4). The response is a scored version of G, thus a scalar, numeric variable, which for simplicity we refer to as Y, with realizations y_i.

Ridge regression is the oldest and simplest method for regularizing linear regression problems with nearly collinear predictors. It is trivially a penalized least squares method since the ridge estimator $\hat{\beta} = (H^T H + \lambda I)^{-1} H^T y$ minimizes the criterion

$$\|y - H\beta\|_N^2 + \lambda \|\beta\|_m^2 ,$$

where N is the sample size and m the number of predictor variables. The matrix λI and its associated penalization $\lambda \|\beta\|_m^2$ are adequate when neither prior knowledge nor purpose of investigation dictate a more adapted kind of regularization. More generally the estimator has the form $\hat{\beta} = (H^T H + \lambda\Omega)^{-1} H^T y$, and it minimizes

$$\|y - H\beta\|_N^2 + \lambda\beta^T\Omega\beta. \tag{28}$$

Here Ω is a more structured penalty matrix, and imposes smoothness with regard to an underlying space, time or frequency domain.

We also want to consider smoothness at the level of functions $\beta(s)$ on this domain, as in section 4. The functional discriminant analysis model translates via optimal scoring into the functional regression model $\eta = \int \beta(s)h(s)ds$ for the mean of Y, and an appropriate criterion might be

$$E_Y(Y - \eta)^2 + \lambda J(\beta) \tag{29}$$

for some penalty functional J, a semi-norm of the space of functions being considered. In discretizing the problem, we have a sample of responses y_i and functions h_i measured at a set of m values of s, but we can leave $\beta(s)$ as a function, chosen to minimize

$$\sum_{i=1}^{N}(y_i - \sum_{\ell=1}^{m} h_i(s_\ell)\beta(s_\ell))^2 + \lambda J(\beta) \tag{30}$$

We now focus on three particular applications: the two examples given in this paper, and the non-parametric regression approach used in FDA.

7.1 Speech data: improper splines

Here the penalty has the form

$$J(\beta) = \int [\beta''(f)]^2 w(f) df \tag{31}$$

where $w(f)$ allows the penalty to give more weight to the lower frequencies. If w were missing, the solution to (31) would be a natural cubic-spline (Wahba 1990) with knots at the sampled frequencies f_ℓ. With w present, the solution will not in general be a spline, and the character of the solution, if it exists, will depend on w. We sidetrack these issues by adapting the spline penalty to approximate (31), and refer to our solution as an *improper spline*. With w absent, the m-dimensional spline solution can be parametrized by the values $\beta_\ell = \beta(f_\ell)$ themselves, and the penalty can be written as $J(\beta) = \beta^T \Omega \beta$, with $\Omega = \Delta^T C^{-1} \Delta$(Green & Yandell 1985, for example). Here Δ is a second difference operator for approximating second derivatives, and C^{-1} can be viewed as a kernel for approximating the integral. If D_w is a diagonal matrix with entries $w(f_\ell)$, we use $\Omega_w = \Delta^T D_w^{\frac{1}{2}} C^{-1} D_w^{\frac{1}{2}} \Delta$ as the penalty in (28).

7.2 Image data: tensor product splines

Here our coefficients $\beta(x, y)$ are indexed by x and y — horizontal and vertical coordinates in the two-dimensional image, and $h(x, y)$ are grayscale values. We used the Laplacian penalty

$$J(\beta) = \int_{x,y} \left[\frac{\partial^2 \beta}{\partial x^2} + \frac{\partial^2 \beta}{\partial y^2} \right]^2 dx dy, \tag{32}$$

Since x and y each assume 16 uniformly spaced values on a lattice in R^2, for computational simplicity we use the simpler discretized version of $J(\beta)$ described in O'Sullivan (1991, page 635). This uses the discrete approximation $J(\beta) = \beta^T \Omega \beta$, where the 256×256 penalty matrix $\Omega = \Delta^T \Delta$ is based on the discrete approximation to the Laplacian: $\Delta = D_x \otimes I_y + I_x \otimes D_y$. The 256-vector β refers to the vectorized version of the 16×16 matrix of coefficients (in row order). Here D_x and D_y are each 16×16 matrices that approximate a second derivative by second differences. O'Sullivan (1991) gives an explicit Fourier representation of the spectral decomposition $\Omega = \Gamma D_\psi \Gamma^T$. This simplifies the solution of (28), since we can transform to new coordinates for which the penalty is diagonal.

7.3 Nonparametric Regression

The regularized regressions we have seen so far penalize the coefficients of the predictors for roughness (in some spatial domain). In the form of nonparametric regression considered here, we first expand the predictors into a large set of basis functions, and then restrict their coefficients to be smooth as a function of the predictors themselves.

For the purposes of the discussion here, we focus on nonparametric additive models for the regression function: $\eta(x) = \sum_{k=1}^{p} f_k(x_k)$ (Buja, Hastie & Tibshirani 1989, Hastie et al. 1992), and an appropriate penalized least-squares criterion

$$\sum_{i=1}^{N} (y_i - \eta(x_i))^2 + J(\eta; \lambda), \tag{33}$$

with

$$J(\eta;\lambda)=\sum_{k=1}^{p}\lambda_k[f_k''(u)]^2du.$$

The solution can be represented in terms of a finite-dimensional basis $h(x):\ \eta(x)=h(x)^T\beta$. Here $h(x)$ is a union of spline basis functions for each coordinate function. Suppose h has m components, and let H be the $N\times m$ basis matrix with ith row $h(x_i)$. $J(\eta;\lambda)$ can be expressed as a quadratic form in β: $J(\eta;\lambda)=\beta^T\Omega\beta$ where Ω is a $m\times m$ block-diagonal, non-negative definite penalty matrix. This again has the form (28).

The number m of basis functions can be quite large; for a univariate smoothing spline there are as many basis functions as there are unique values of x_i (potentially N.) The additive-spline model can have up to Np basis functions! Fortunately there are efficient iterative algorithms for computing the solution, especially for additive spline models. More details on the additive spline model, and several examples using them in this context, are given in Hastie et al. (1992).

7.4 Degrees of freedom

In the examples we have referred to the effective degrees of freedom of a smooth fit. This is a useful translation of the smoothing parameter into a more meaningful parameter that can be used to calibrate a variety of different linear methods. The vector of fitted values for η in (28) is given by

$$S(\lambda)y=H(H^TH+\lambda\Omega)^{-1}H^Ty. \tag{34}$$

where $S(\lambda)$ is an $N\times N$ linear operator matrix. The degrees of freedom are defined to be $df(\lambda)=\mathrm{tr}S(\lambda)$(Buja et al. 1989, Wahba 1990). The function df is monotone in λ, and in practice we can fix df and determine the appropriate value of λ with minimal additional computational cost. In the examples in this paper we avoided automatic selection of regularization parameters. Instead we have favored an empirical approach of examining discriminant functions and misclassification performance for a few values of df and selecting the most informative values.

8 Discussion

Penalized discriminant analysis appears to be a useful tool for problems such as speech and image recognition, featuring a large number of correlated inputs. An attractive feature of the method is the potential to extract and interpret a small number of discriminant functions and the resulting benefits for scientific understanding of the data. Potential medical applications include the disease classification of pap smears and mammograms.

Acknowledgements

We thank Michael Leblanc and Werner Stuetzle for helpful discussion. We also thank the referees and associate editor for their insightful suggestions on an earlier draft of this paper, which encouraged us to rethink some of the issues and resulted in the inclusion of section 4. The third author was supported by the Natural Sciences and Engineering Research Council of Canada.

9 Appendix- details of the distance calculation

9.1 Reduction of Distances to Discriminant Coordinates

The decomposition in (8) is in terms of at most $J-1$ eigenvectors of Σ_{Bet}; the Jth is trivial with eigenvalue 0 if H is centered, else equally trivial with eigenvalue 1 if H is not centered but includes the constant column. Assuming that $J < m$, it is useful to pad out the decomposition to the full m dimensions:

- let $\mathrm{B}^* = (\mathrm{B} : \mathrm{B}_\perp)$ be of size $m \times m$, such that $\mathrm{B}^{*T}\Sigma_{22}B^* = I_m$, and
- $D_\lambda = D_{(\alpha^2:\, 0)}$,

where B is now the full $J-1$ dimensional nontrivial solution, and D_{α^2} the corresponding nonzero eigenvalues. The following facts can be easily verified:

$$\begin{aligned}
\Sigma_{22}^{-1} &= \mathrm{B}^*\mathrm{B}^{*T} \\
&= \mathrm{B}\mathrm{B}^T + \mathrm{B}_\perp \mathrm{B}_\perp^T; && (35)\\
\Sigma_W^{-1} &= \mathrm{B}^*(I-D_\lambda)^{-1}\mathrm{B}^{*T} \\
&= \mathrm{B}(I-D_\alpha^2)^{-1}\mathrm{B}^T + \mathrm{B}_\perp \mathrm{B}_\perp^T; && (36)\\
\Sigma_{Bet} &= \mathrm{B}^{*-T} D_\lambda \mathrm{B}^{*-1}. && (37)
\end{aligned}$$

An implication is that, confined to the subspace spanned by M, (penalized) Mahalanobis distances differ using the metrics Σ_{22}^{-1} and Σ_W^{-1}, but are the same in the orthogonal subspace.

The following equation shows that classification can be based entirely on discriminant coordinates:

$$D(h, m^j) = \left\|\mathrm{B}_{\mathrm{LDA}}^T(h - m^j)\right\|^2 + \left\|\mathrm{B}_\perp^T h\right\|^2 \qquad (38)$$

The first term follows from (36) together with the relation between $\mathrm{B}_{\mathrm{LDA}}$ and B from (19). From (37) we see that $\mathrm{B}_\perp^T \Sigma_{Bet} = 0$, but since $\Sigma_{Bet} = M^T\Sigma_{11}M$, it follows that $\mathrm{B}_\perp^T M^T = 0$, and hence $\mathrm{B}_\perp^T m^j = 0\ \forall j$. Thus (38) follows.

If the dimension m of the (expanded) predictors h is much larger than the number of classes, J, the reduction to discriminant coordinates may give considerable savings.

9.2 Reduction of Distances to Optimal Scoring

In section 3.5, we established links between projections based on LDA and OS. In terms of distance calculations, they imply that classification can be based on scores θ^j and fits η rather than class centers m^j and predictors h. This can be convenient in settings where the dimension m of h is very large; for example, when η is modeled by an additive spline, $m \sim Np$, where p is the number of original predictors in x. Algorithms for fitting additive spline models efficiently compute the additive fit $\beta_{\mathrm{OS}}^T h$; it is therefore useful to express the classification criterion in terms of this regression as well. To this end, we reformulate (38) based on (24) and (25):

$$\begin{aligned}
\left\|\mathrm{B}_{\mathrm{LDA}}(h - m^j)\right\|^2 &= h^T\mathrm{B}_{\mathrm{LDA}}\mathrm{B}_{\mathrm{LDA}}^T h - 2h^T\mathrm{B}_{\mathrm{LDA}}\mathrm{B}_{\mathrm{LDA}}^T m^j \\
&\quad + {m^j}^T\mathrm{B}_{\mathrm{LDA}}\mathrm{B}_{\mathrm{LDA}}^T m^j \\
&= \eta^T D_{\alpha^{-2}(1-\alpha^2)^{-1}}\eta - 2\eta^T D_{(1-\alpha^2)^{-1}}\theta^j \\
&\quad + {\theta^j}^T D_{\alpha^2(1-\alpha^2)^{-1}}\theta^j \\
&= \eta^T D_{(\alpha^{-2}-1)(1-\alpha^2)^{-1}}\eta + (\theta^j - \eta)^T D_{(1-\alpha^2)^{-1}}(\theta^j - \eta) \\
&\quad - {\theta^j}^T D_{(1-\alpha^2)(1-\alpha^2)^{-1}}\theta^j \\
&= \|D_{\alpha^{-1}}\eta\|^2 + \left\|D_{(1-\alpha^2)^{-\frac{1}{2}}}(\theta^j - \eta)\right\|^2 - \|\theta^j\|^2 && (39)
\end{aligned}$$

This result in itself proves the assertion. A closer look at the derivation reveals that it is of greater generality than meets the eye. Assume one wishes to perform dimension reduction and base the distance calculations for classification on the first $K < J-1$ discriminant coordinates only. One would then use

- $\mathrm{B}_{\mathrm{LDA}}$ reduced to the first K columns,
- θ^j reduced to the first K scores and
- η reduced to the first K fitted values.

It turns out that the derivations of (39) still hold. Thus, scores and fitted values are a base for distance calculation at any level of dimension reduction.

Assuming that no dimension reduction is desired, one can continue and express (39) in a variety of ways, e.g., by rewriting the equation (38) for the Mahalanobis distance with the help of (35):

$$D(h, m^j) = h^T \Sigma_{22}^{-1} h + \left\| D_{(1-\alpha^2)^{-\frac{1}{2}}} (\theta^j - \eta) \right\|^2 - \left\| \theta^j \right\|^2 \tag{40}$$

A final touch concerns the term $\left\| \theta^j \right\|^2$: Let Θ' denote Θ augmented with the trivial column $\theta_J = 1$. Then Θ' is square and nonsingular, with $\Theta'^T \Sigma_{11} \Theta' = I_J$. Thus, since $\Sigma_{11} = D_p$ is diagonal with the class proportions as diagonal elements, we have $\Theta' \Theta'^T = D_p^{-1}$, or $\left\| \theta^j \right\|^2 + 1 = 1/p_j \ \forall j$:

$$D(h, m^j) = h^T \Sigma_{22}^{-1} h + \left\| D_{(1-\alpha^2)^{-\frac{1}{2}}} (\theta^j - \eta) \right\|^2 - 1/p_j + 1 \tag{41}$$

This is the form given by Breiman & Ihaka (1984).

If we neglect terms which are independent of the classes, such as $h^T \Sigma_{22}^{-1} h$, the original class-adjusted Mahalanobis distance (26) is seen to be equivalent to the distances labelled 1–5 in section 3.6.

References

Breiman, L. & Ihaka, R. (1984), Nonlinear discriminant analysis via scaling and ACE, Technical report, Univ. of California, Berkeley.

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models (with discussion)', *Annals of Statistics* **17**, 453–555.

Campbell, N. (1980), 'Shrunken estimators in discriminant and canonical variate analysis', *Applied Statistics* **29**(1), 5–14.

De Leeuw, J., Young, F. & Takane, Y. (1976), 'Additive structure in qualitative data: an alternating least squares method with optimal scaling features', *Psychometrika* **41**, 471–503.

DiPillo, P. (1976), 'The application of bias to discriminant analysis', *Commun. Statist.-Theor. Meth.* **A5(9)**, 843–854.

DiPillo, P. (1979), 'Biased discriminant analysis: Evaluation of the optimum probability of classification', *Commun. Statist.-Theor. Meth.* **A8(14)**, 1447–1457.

Friedman, J. (1989), 'Regularized discriminant analysis', *Journal of the American Statistical Association* **84**(405), 165–175.

Gifi, A. (1981), Nonlinear multivariate analysis, unpublished manuscript.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, Chichester.

Green, P. & Yandell, B. (1985), Semi-parametric generalized linear models, *in* 'Proceedings of the 2nd International GLIM conference, Lancaster', Vol. 32 of *Lecture notes in Statistics*, Springer-Verlag, New York, pp. 44–55.

Hastie, T. & Mallows, C. (1993), 'Comment on "a statistical view of some chemometric regression tools" by j. friedman and i. frank', *Technometrics* **35**(2), 140–143.

Hastie, T., Tibshirani, R. & Buja, A. (1992), Flexible discriminant analysis by optimal scoring, To be published.

Kiiveri, H. (1992), 'Canonical variate analysis of high-dimensional spectral data', *Technometrics* **34**, 321–331.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. (1990), Handwritten digit recognition with a back-propogation network, *in* D. Touretzky, ed., 'Advances in Neural Information Processing Systems', Vol. 2, Morgan Kaufman, Denver, CO.

Lebart, L., Morineau, A. & Warwick, K. (1984), *Multivariate Descriptive Statistical Analysis*, Wiley, New York.

Leurgans, S., Moyeed, R. & Silverman, B. (1993), 'Canonical correlation analysis when the data are curves', *J. Royal Statist. Soc. (Series B)* **55**, 725–740.

Mardia, K., Kent, J. & Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

O'Sullivan, F. (1991), 'Discretized laplacian smoothing by fourier methods', *J. Amer. Statist. Assoc.* **86**, 634–642.

Ramsay, J. & Dalzell, C. (1991), 'Some tools for functional data analysis (with discussion)', *J. Royal Statist. Soc. (Series B)* **53**, 539–572.

Seber, G. (1984), *Multivariate Observations*, Wiley, New York.

Takane, Y., Young, F. & De Leeuw, J. (1979), 'Nonmetric common factor analysis: an alternating least squares method with optimal scaling features', *Behaviormetrika* **6**, 45–56.

Vinod, H. (1976), 'Canonical ridge and econometrics of joint production', *J. Econometrics* **4**, 147–166.

Vinod, H. & Ullah, A. (1981), *Recent Advances in Regression Methods*, Marcell Dekker, New York.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

Young, F., Takane, Y. & De Leeuw, J. (1978), 'The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features', *Psychometrika* **43**, 505–529.

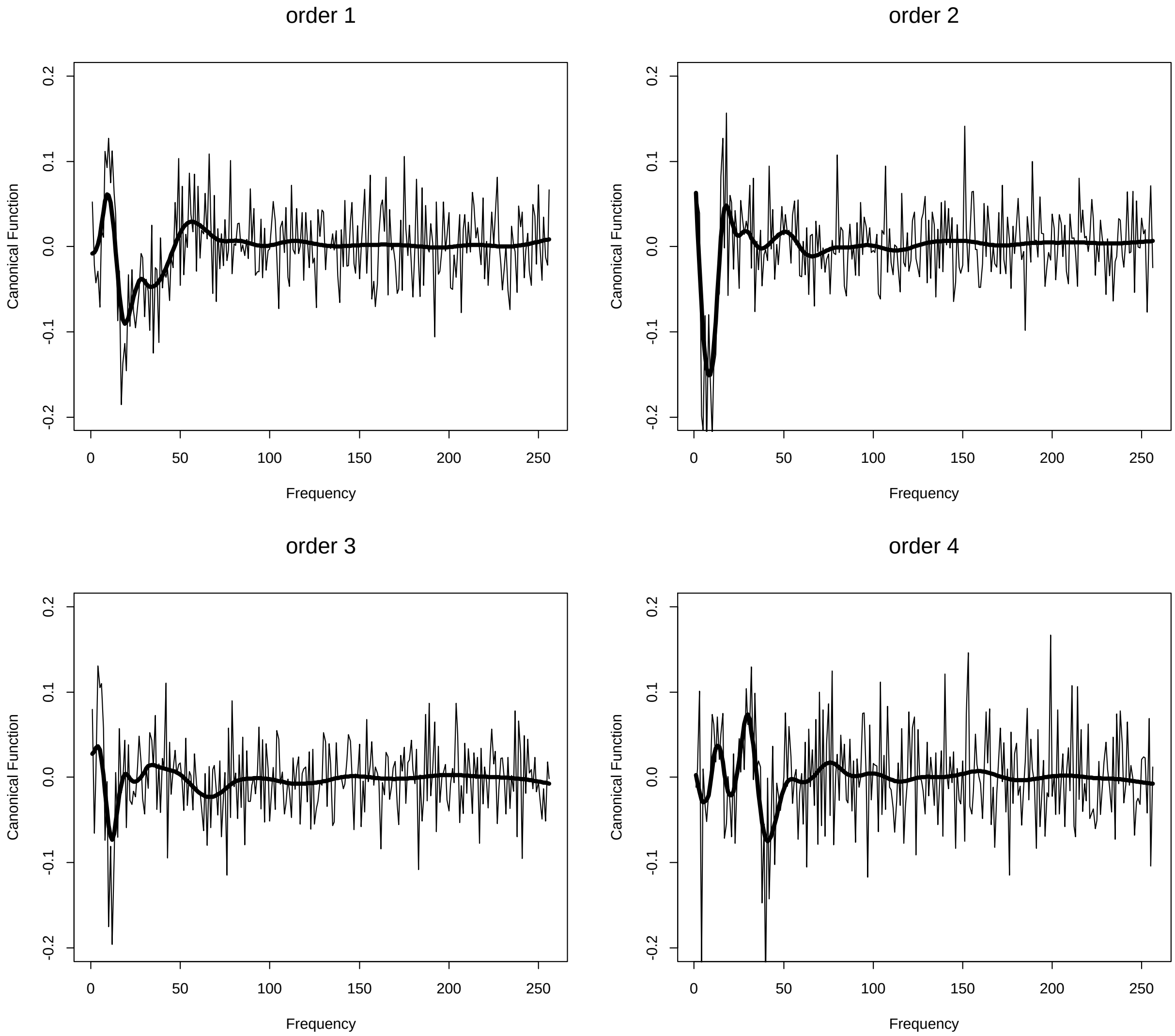

Figure 1: *The jagged functions are the four discriminant coefficient functions for separating the five phonemes in the sampled utterances in example 5. The superimposed smooth functions are the regularized discriminant coefficient functions, using* 30 df *worth of smoothing.*

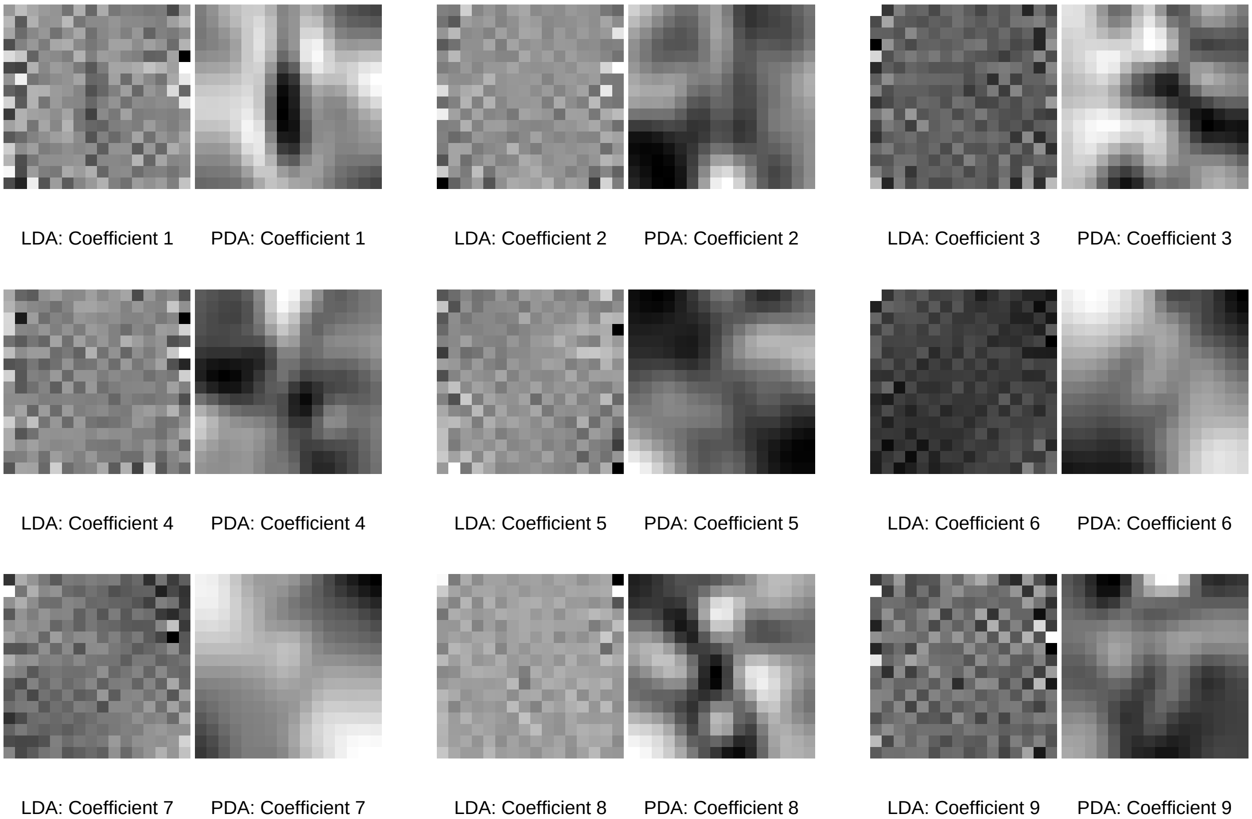

Figure 2: *The bitmaps appear in pairs, and represent as images the nine discriminant coefficient functions for the digit recognition problem. The left member of each pair is the LDA coefficient, while the right member is the PDA coefficient, regularized to enforce spatial smoothness.*

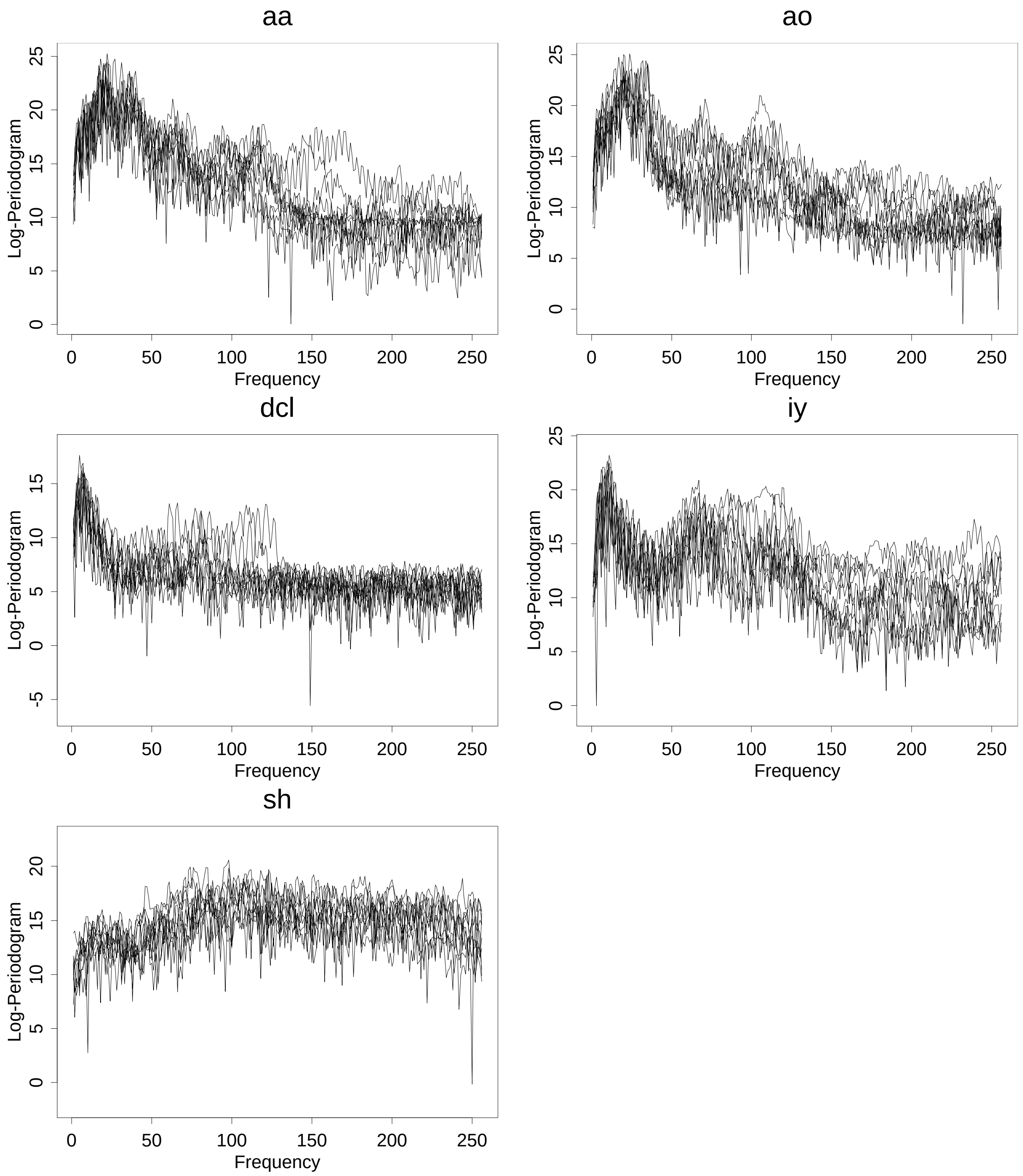

Figure 3: *Each plot shows a sample of 10 log-periodogram's within each phoneme class. The periodograms consist of a smooth trend plus high frequency oscillations.*

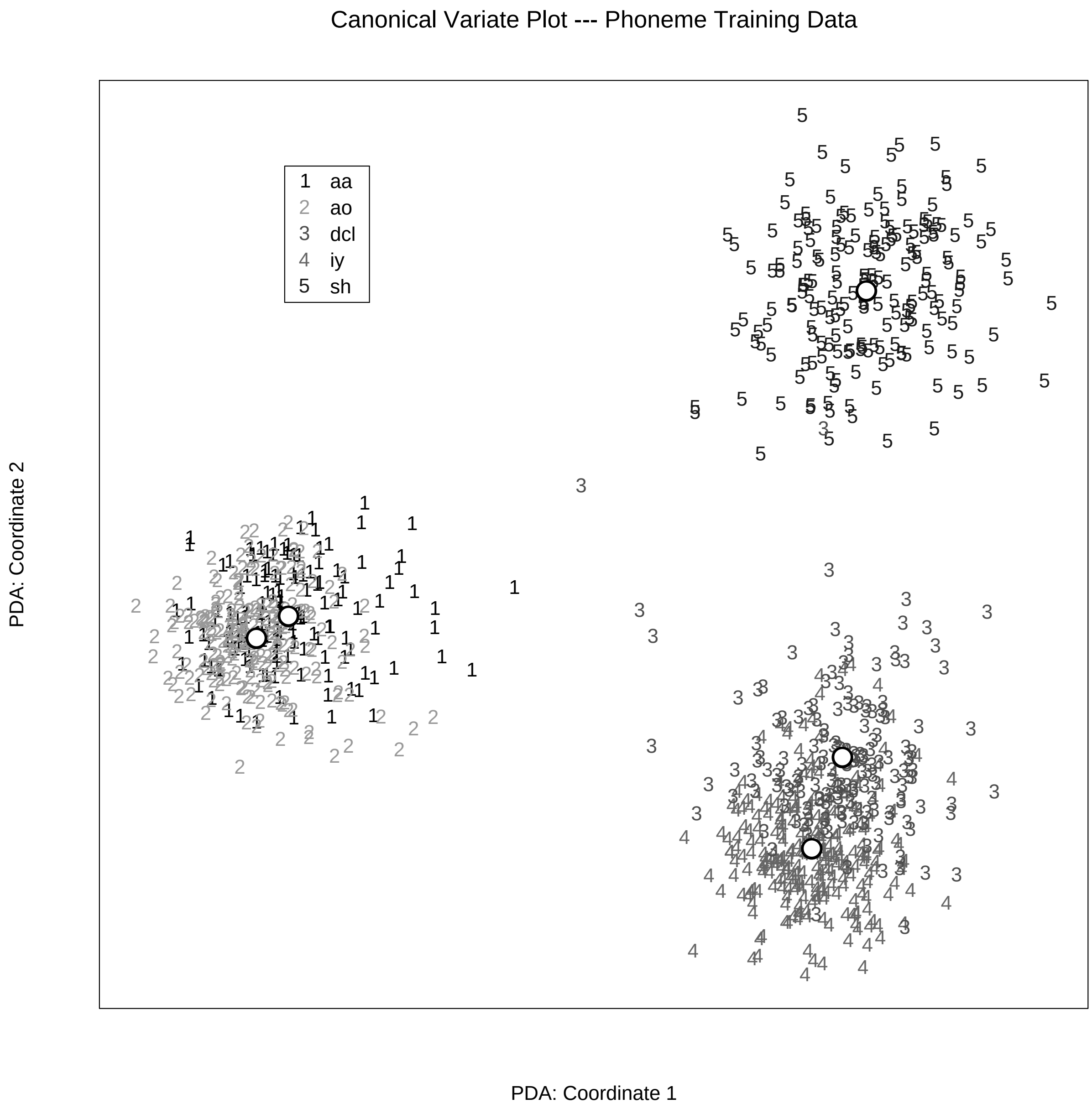

Figure 4: *The first two penalized canonical variates, evaluated for the training data. The true class identities are indicated. As might be expected, the two vowel sounds, "ao" and "aa" are confused. The circles indicate the class centroids.*

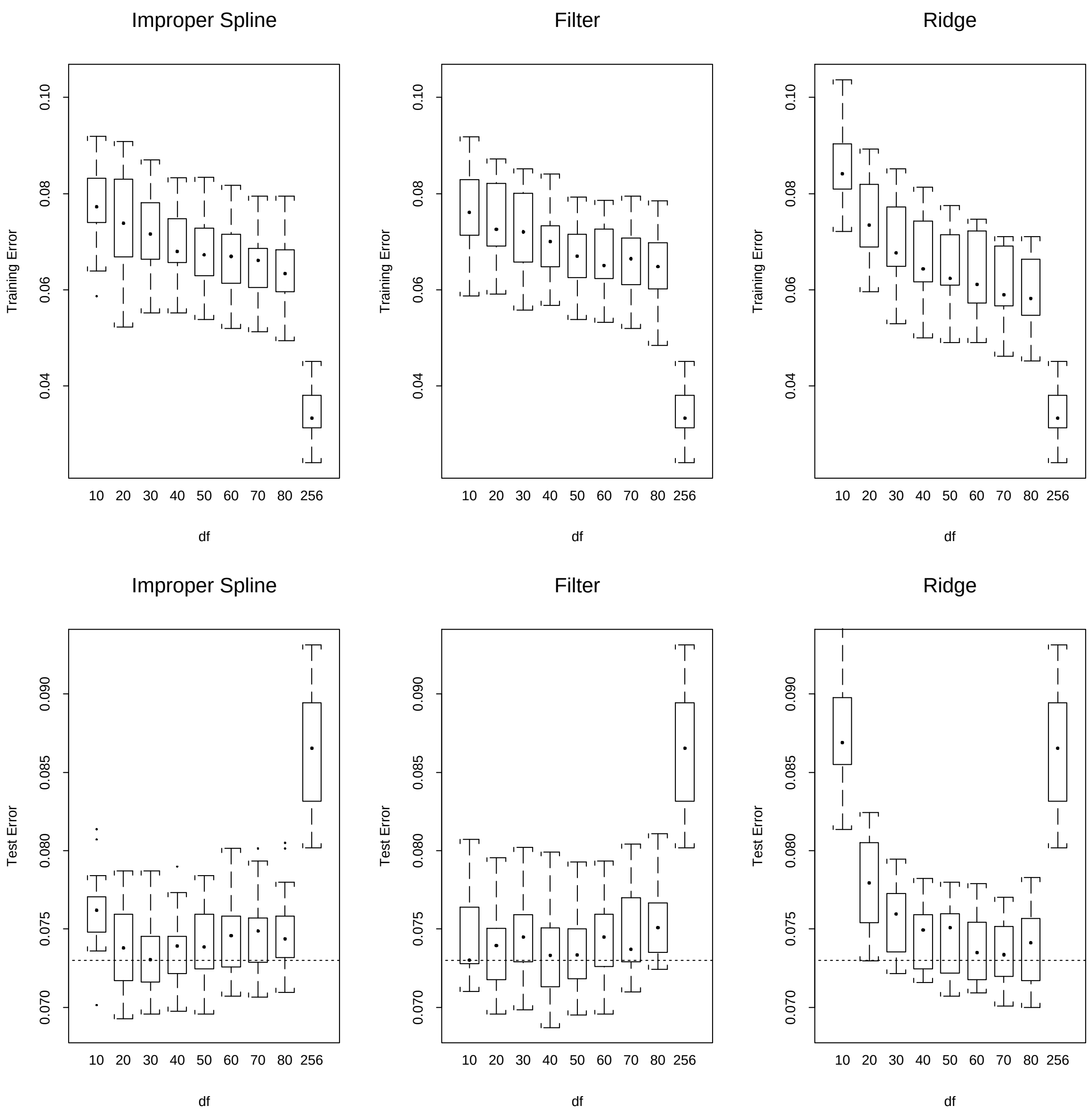

Figure 5: *Training and test errors as a function of* df *for three versions of PDA, compared with the unrestricted LDA (*df *=256). The three version are (left) regularized using a weighted smoothing penalty, (middle) asymmetric filters, and (right) simple ridging. All three show comparable classification performence at their optimum* df.

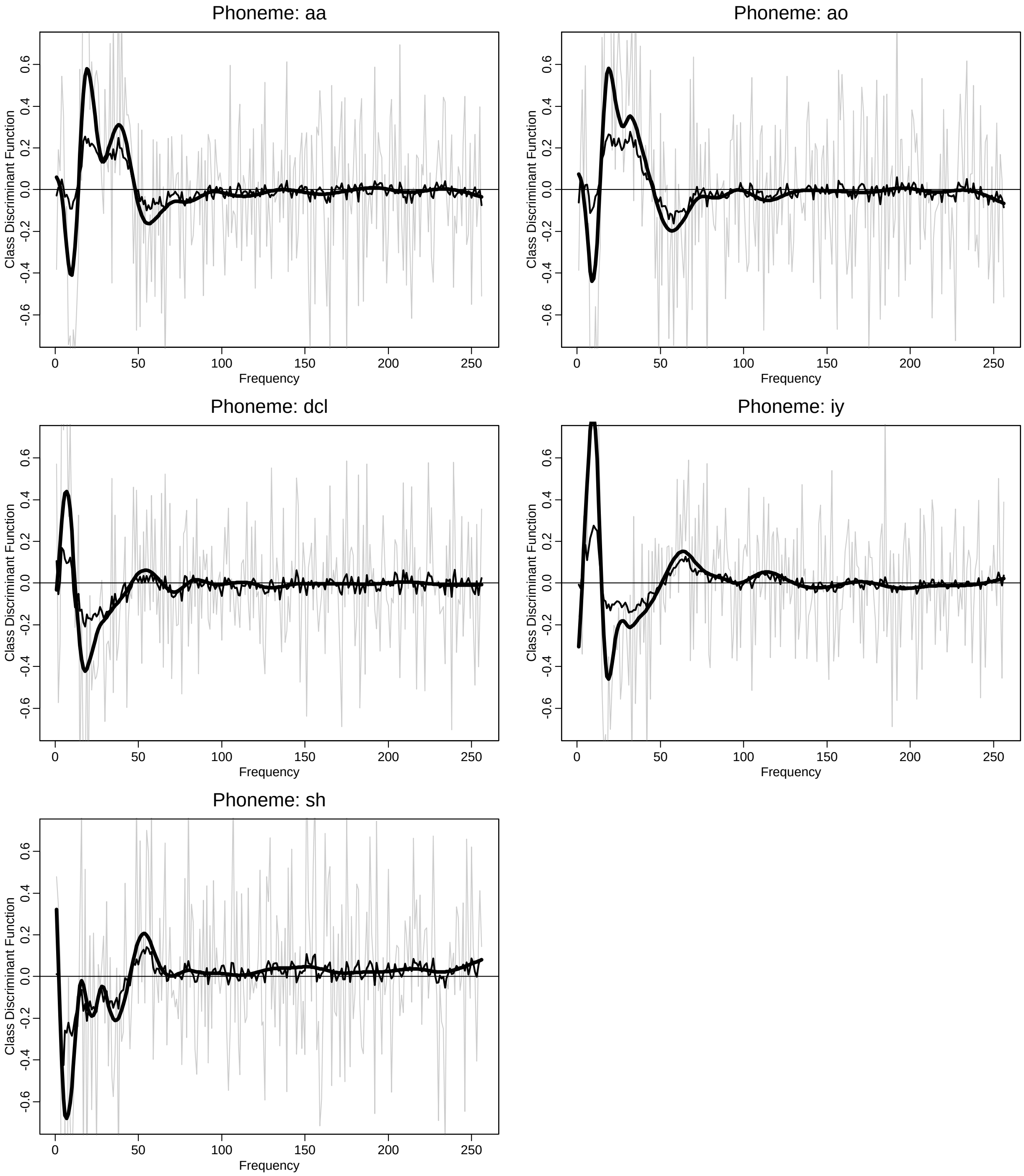

Figure 6: *Class discriminant functions for the 5 phonemes. These are analogs of the functions* $\Sigma_W^{-1} m^j$ *for linear discriminant analysis, except the within covariance here is penalized. The thick solid curve represents PDA using the improper spline penalty and* 30 df. *The raw LDA coefficients are in the background. The wiggly curve is the ridge version of PDA, using* 70 df.

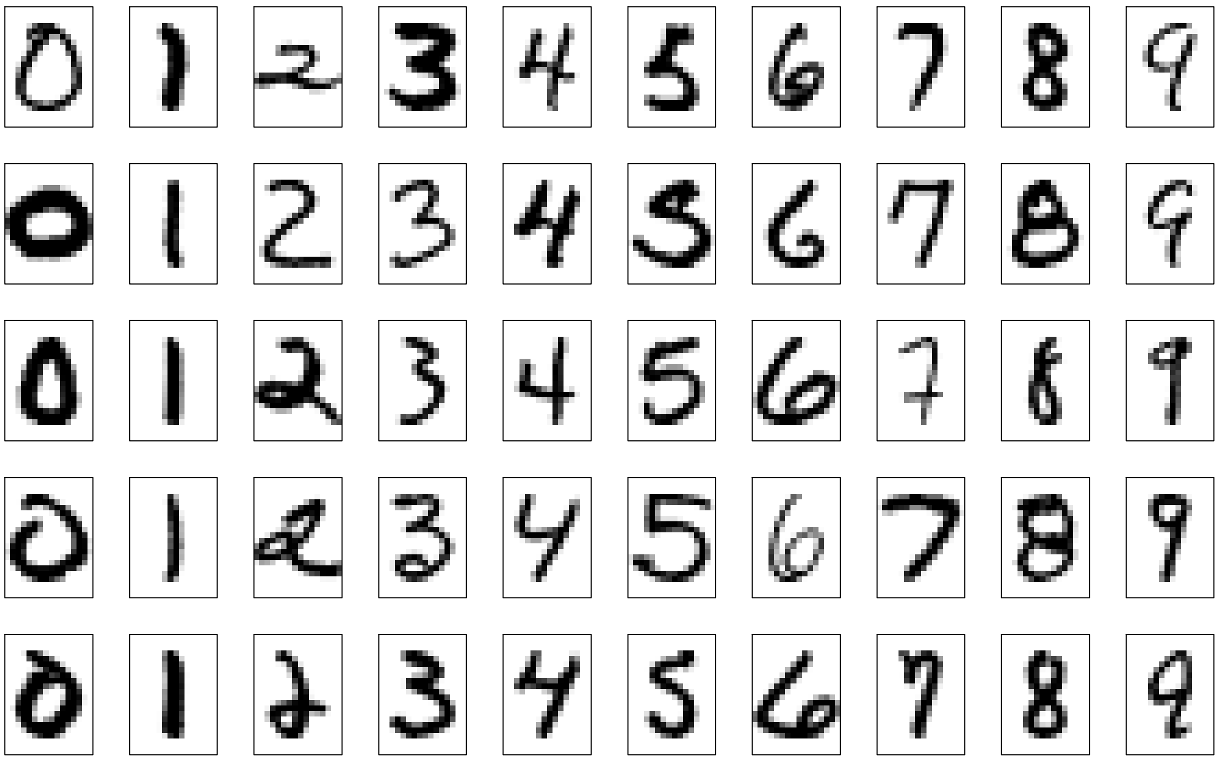

Figure 7: *A random selection of digitized handwritten digits. Each image is a 8 bit, grayscale version of the original binary image, size and orientation normalized to* 16×16 *pixels.*

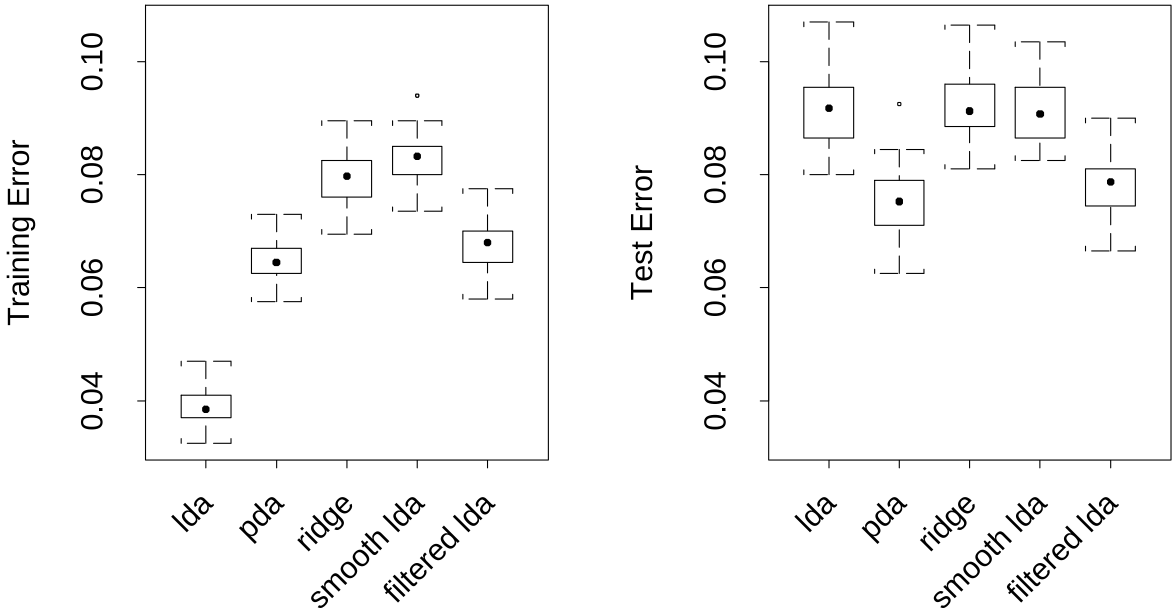

Figure 8: *Each boxplot shows the misclassification results for 50 resampling trials, the left panel training or resubstitution error, the right test error. All methods use about 40 degrees of freedom. The method labelled* smooth LDA *was obtained by smoothing the LDA coefficients as a function of spatial coordinates, using the same Laplacian penalty as the PDA fits. The method labelled* filtered LDA *corresponds to LDA on a 44-dimensional set of filter coefficients, described in the text.*

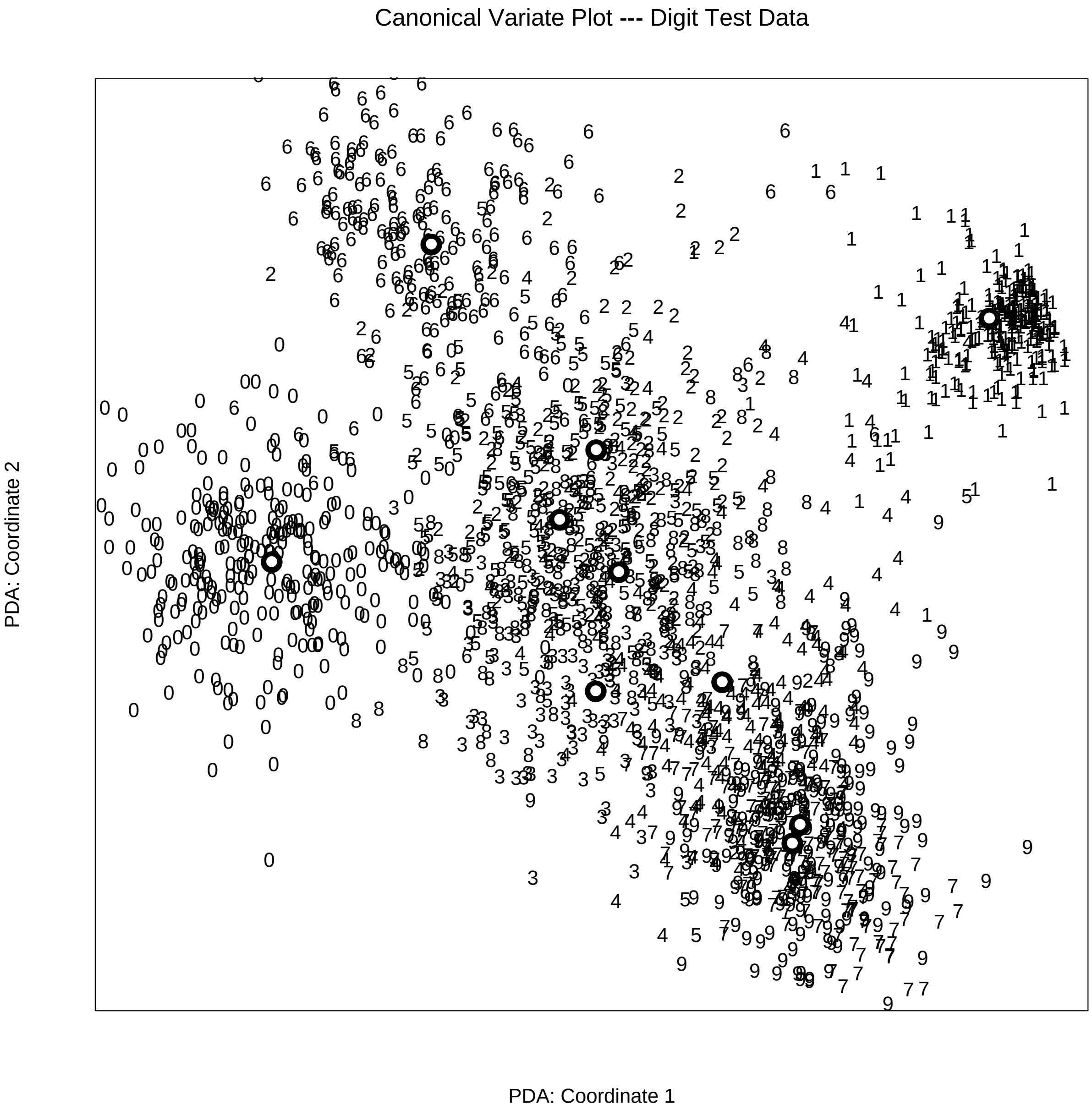

Figure 9: *The first two penalized canonical variates, evaluated for the test data. The circles indicate the class centroids. The first coordinate contrasts mainly 0s and 1s, while the second contrasts 6s and 7/9s.*