# Functional Linear Discriminant Analysis for Irregularly Sampled Curves

GARETH M. JAMES
*Marshall School of Business, University of Southern California*
gareth@usc.edu
and TREVOR J. HASTIE
*Department of Statistics, Stanford University*
hastie@stat.stanford.edu

February 13, 2001

**Abstract**

We introduce a technique for extending the classical method of Linear Discriminant Analysis to data sets where the predictor variables are curves or functions. This procedure, which we call *functional linear discriminant analysis (FLDA)*, is particularly useful when only fragments of the curves are observed. All the techniques associated with LDA can be extended for use with FLDA. In particular FLDA can be used to produce classifications on new (test) curves, give an estimate of the discriminant function between classes, and provide a one or two dimensional pictorial representation of a set of curves. We also extend this procedure to provide generalizations of quadratic and regularized discriminant analysis.

*Some key words*: Classification; Filtering; Functional data; Linear discriminant analysis; Low dimensional representation; Reduced rank; Regularized discriminant analysis; Sparse curves.

## 1 Introduction

Linear discriminant analysis (LDA) is a popular procedure which dates back as far as Fisher (1936). Let $\mathbf{X}$ be a $q$-dimensional vector representing an observation from one of several possible classes. Linear discriminant analysis can be used to classify $\mathbf{X}$ if the class is unknown. Alternatively, it can be used to characterize the way that classes differ via a *discriminant function*. There are several different ways of describing LDA. One is using probability models. Suppose that the $i$th class has density $f_i(\mathbf{x})$ and prior probability $\pi_i$. Then Bayes' formula tells us that

$$P(\text{Class} = i|\mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{\sum_j f_j(\mathbf{x})\pi_j}\cdot \tag{1}$$

It is relatively simple to show that the rule that classifies to the largest conditional probability will make the smallest expected number of misclassifications. This is known as the Bayes' rule or classifier. If we further assume that the $i$th class has a Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance $\Sigma$ then it can be shown that classifying to the maximum conditional probability is equivalent to classifying to

$$\arg\max_i L_i, \tag{2}$$

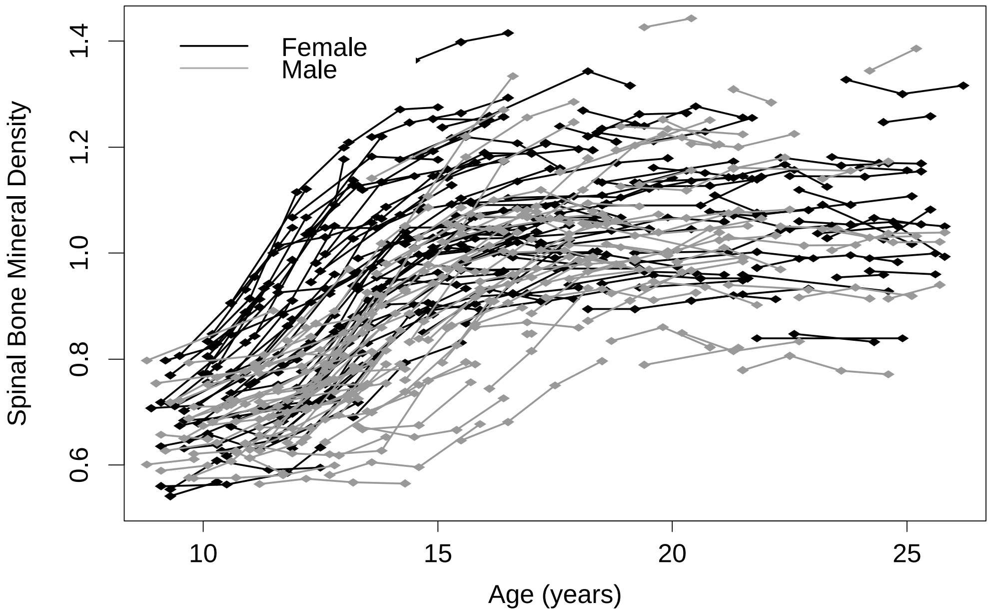



Figure 1: *Data measurements of spinal bone mineral density ($g/cm^2$) for 280 individuals. The black lines represent females and the grey lines males.*

where $L_i$ is the discriminant function

$$L_i = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i / 2 + \log \pi_i.$$

Note that $L_i$ is a linear function of $\mathbf{x}$. When the maximum likelihood estimates for $\boldsymbol{\mu}_i$ and $\Sigma$ are used we arrive at the linear discriminant analysis procedure.

### 1.1 LDA on functional data

LDA can be implemented on data of any finite dimension but cannot be directly applied to infinite-dimensional data such as functions or curves. Provided the entire curve has been observed, this can be overcome by discretizing the time interval. However, this generally results in highly correlated high-dimensional data which makes the within-class covariance matrix difficult to estimate. There are two common solutions to this problem. The first is to use some form of regularization, such as adding a diagonal matrix to the covariance matrix. See for example DiPillo (1976), DiPillo (1979), Campbell (1980), Friedman (1989) and Hastie *et al.* (1995). We call this the *regularization method*. The second is to choose a finite-dimensional basis, $\phi(x)$, and find the best projection of each curve onto this basis. The resulting basis coefficients can then be used as a finite-dimensional representation making it possible to use LDA, or any other procedure, on the basis coefficients. We call this the *filtering method*.

Unfortunately, it is often the case that only a fragment of each curve has been observed. Consider, for example, the data illustrated in Figure 1. These data consist of measurements of spinal bone mineral density for 280 individuals taken at various ages, a subset of the data presented in Bachrach *et al.* (1999). Even though, in aggregate, there are 860 observations measured over a period of almost two decades, we only have 2-4 measurements for each individual, typically measured over no more than a couple of years. In this situation both of the common approaches to discriminant analysis can break down. The regularization method is not feasible because discretizing would result in a large number of missing observations in each dimension. The filtering method also has several potential problems. The first is that an assumption is made of a common covariance matrix for each curves' basis coefficients. However, if the curves are measured at different time points, as is the case in the growth curve data of Figure 1, the coefficients will all have different covariances. One would ideally like to put more weight on accurate basis coefficients but the filtering method does not allow such an approach. A second problem is that with extremely sparse data sets some of the basis coefficients may have infinite variance, making it impossible to estimate the entire curve. For example, with the spinal bone density data, each individual curve has so few observations that it is not possible to fit any reasonable basis. In this case the method fails and there is no way to proceed. For the sparse data considered in this paper these are serious problems.

### 1.2 General functional model

The regularization and filtering approaches can both be viewed as methods for fitting the following general functional model. Let $g(t)$ be the curve of an individual randomly drawn from the $i$th class. Assume that, if $g(t)$ is in Class $i$, it is distributed as a Gaussian process with

$$E\{g(t)\} = \mu_i(t), \quad Cov\{g(t), g(t')\} = \omega(t, t')$$

One typically never observes an individual over the entire curve; rather one samples the curve with error at distinct time points $t_1, \ldots, t_n$. We assume that the measurement errors are uncorrelated with mean zero and constant variance $\sigma^2$. Let $\mathbf{Y}$ be the vector of observations of $g(t)$ at times $t_1, \ldots, t_n$. Then

$$\mathbf{Y} \sim N(\boldsymbol{\mu}_i, \Omega + \sigma^2 I)$$

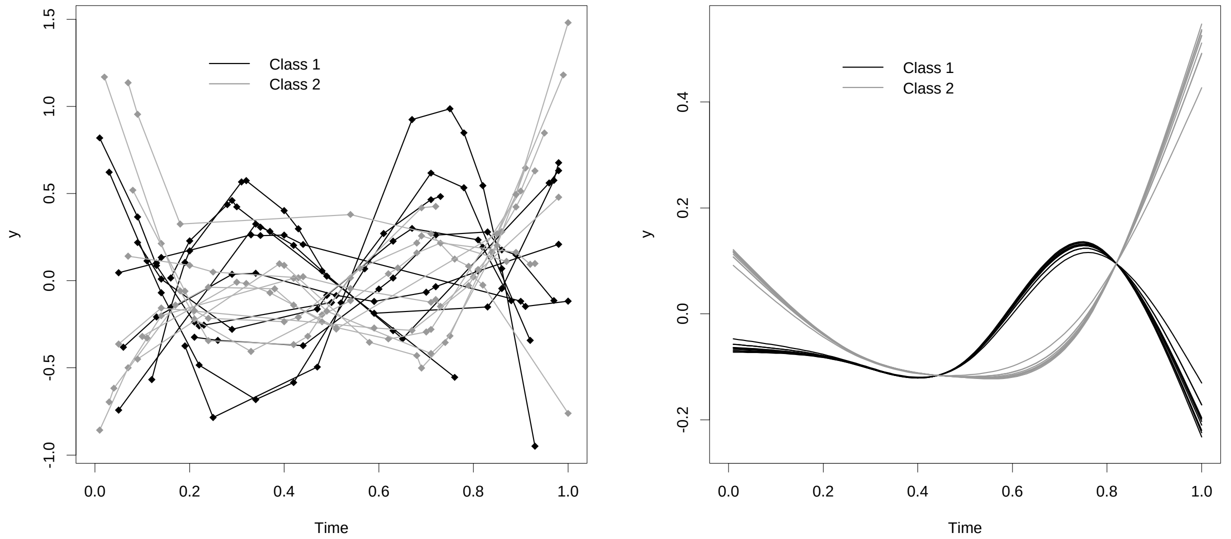
where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i(t_1) \\ \mu_i(t_2) \\ \vdots \\ \mu_i(t_n) \end{bmatrix}, \qquad \Omega = \begin{bmatrix} \omega(t_1, t_1) & \omega(t_1, t_2) & \cdots & \omega(t_1, t_n) \\ \omega(t_2, t_1) & \omega(t_2, t_2) & \cdots & \omega(t_2, t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega(t_n, t_1) & \omega(t_n, t_2) & \cdots & \omega(t_n, t_n) \end{bmatrix} \tag{3}$$

The Bayes rule for classifying this curve is given by (2) with $\boldsymbol{\mu}_i$ as given in (3) and $\Sigma = \Omega + \sigma^2 I$. Many functional classification procedures are simply methods for fitting this model, i.e. estimating $\mu_i(t)$ and $\omega(t, t')$, and then using the classification rule given by (2). For example, the regularization approach attempts to estimate $\mu_i(t)$ and $\omega(t, t')$ by producing sample estimates along a fine lattice of time points. Alternatively, the filtering method forms estimates by modeling $\mu_i(t)$ and $\omega(t, t')$ using basis functions. However, we saw in Section 1.1 that, when confronted with sparse data sets, both methods can produce poor fits to the model.

### 1.3 The FLDA approach

In this paper we present an alternative method for fitting the functional model of §1.2, which copes well with sparse data. We call this method "functional linear discriminant analysis" (FLDA). The procedure uses a spline curve plus random error to model observations from each individual. The spline is parameterized using a basis function multiplied by a $q$-dimensional coefficient vector. This




(a) (b)

Figure 2: *(a) A simulated data set of* 20 *curves from 2 different classes. (b) The transformed curves after removing the random components.*

effectively transforms all the data into a single $q$-dimensional space. Finally, the coefficient vector is modeled using a Gaussian distribution with common covariance matrix for all classes, in analogy with LDA. The observed curves can then be pooled to estimate the covariance and mean for each class. This makes it possible to form accurate estimates for each individual curve based on only a few observations.

This has several advantages over the regularization and filtering methods. First, as it does not rely on forming individual estimates for each curve, it can be used on sparse data sets such as the growth curve data. Second, by producing an estimate for the covariance kernel it is possible to estimate the variance of the basis coefficient for each curve and automatically put more weight on the more accurate coefficients.

As a simple illustration of the effectiveness of FLDA in separating curves from different classes consider Figure 2(a). This is a plot of curves from a simulated data set. Class 1 curves are plotted as black lines and Class 2 as grey. Each class has a different mean function and the curves are generated by combining the class mean with a random curve plus random normal noise. From visual inspection alone there is no obvious separation between classes. However, Figure 2(b) provides a plot of the transformed curves after using the FLDA procedure to remove the "random component" from each curve. Now the separation is clear so when the same procedure is applied to a new curve it will clearly be identifiable which group it falls into and it can be classified with high accuracy. The level of accuracy depends on the signal to noise ratio, which is high in this case. However, the key point is that the strong signal is not apparent in Figure 2(a) and only emerges as a result of using the FLDA procedure. In the classical two-class LDA setting this transformation amounts to projecting observations onto the line segment spanned by the means.

The FLDA model and classification procedure are presented in Sections 2 and 3. One of the reasons for the popularity of LDA is that it can be used for a variety of tasks. It can be used to project high-dimensional data into a low dimension and hence produce a graphical representation.

Furthermore it can also be used for classification and to produce a discriminant function to identify areas of discrimination between classes. In Section 4 we show how FLDA can be used to generalize each of these tools to functional data. Section 5 explains how the standard FLDA framework can be extended to include rank-reduced and non-identical within-class covariance matrices. The latter of these extensions provides a functional generalization of quadratic discriminant analysis.

# 2 The FLDA model

In this section we develop the FLDA model by generalizing the LDA model given in §1 to handle functional data.

## 2.1 A generalization to functional data

Let $g_{ij}(t)$ be the true value at time $t$ for the $j$th individual or curve from the $i$th class. Let $\mathbf{Y}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ be the corresponding vectors of observations and measurement errors at times $t_{ij1}, \ldots, t_{ijn_{ij}}$. Then we begin with

$$\mathbf{Y}_{ij} = \mathbf{g}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, m_i,$$

where $K$ is the number of classes and $m_i$ is the number of individuals in the $i$th class. The measurement errors are assumed to have mean zero, constant variance $\sigma^2$ and be uncorrelated with each other and $\mathbf{g}_{ij}$. These assumptions implicitly mean that we are assuming that the time points that we have failed to observe are missing at random. As we only have a finite amount of data we need to place some restrictions on $g_{ij}$ in order to fit this model. A common approach to modeling functional data is to represent the functions using a flexible basis (Ramsay and Silverman 1997, Chapter 3). We choose to use natural cubic spline functions because of their desirable mathematical properties and easy implementation (de Boor, 1978; Green and Silverman, 1994). Let

$$g_{ij}(t) = \mathbf{s}(t)^T \boldsymbol{\eta}_{ij},$$

where $\mathbf{s}(t)$ is a spline basis with dimension $q$ and $\boldsymbol{\eta}_{ij}$ is a $q$-dimensional vector of spline coefficients. This leads to a more restricted model,

$$\mathbf{Y}_{ij} = S_{ij}\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij} \quad i = 1, \ldots, K, \quad j = 1, \ldots, m_i,$$

where

$$S_{ij} = (\mathbf{s}(t_{ij1}), \ldots, \mathbf{s}(t_{ijn_{ij}}))^T.$$

Notice that the problem of modeling $\mathbf{Y}_{ij}$ has reduced to one of modeling $\boldsymbol{\eta}_{ij}$. However, $\boldsymbol{\eta}_{ij}$ is a $q$-dimensional variable, so a natural approach is to model it using the Gaussian distribution assumed for the standard LDA model,

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_{ij}, \quad \boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, \Gamma).$$

If we assume that the error terms are also normally distributed this gives

$$\mathbf{Y}_{ij} = S_{ij}(\boldsymbol{\mu}_i + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots K, \quad j = 1, \ldots, m_i, \tag{4}$$

$$\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \sigma^2 I), \quad \boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, \Gamma).$$

### 2.2 Rank reduced LDA

A reduced-rank version of LDA is often performed by transforming or projecting the variables into a lower-dimensional subspace and classifying in this subspace. The subspace is chosen to maximize the between-class covariance relative to the within-class covariance. These transformed variables are called *linear discriminants* or *canonical variables.* Anderson (1951) and Hastie and Tibshirani (1996) outline an alternative procedure using the constraint

$$\boldsymbol{\mu}_i = \boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i, \quad \Lambda^T\Sigma^{-1}\Lambda = I, \quad \sum_i \boldsymbol{\alpha}_i = 0, \tag{5}$$

where $\boldsymbol{\lambda}_0$ and $\boldsymbol{\alpha}_i$ are respectively $q$- and $h$-dimensional vectors, and $\Lambda$ is a $q \times h$ matrix, $h < \min(q, K)$. Both sets of authors show that using maximum likelihood to fit the Gaussian LDA model of §1 with the added constraint (5) and classifying to the maximum posterior probability is identical to the classification from the reduced rank LDA procedure.

The same rank constraint can be placed on the means in (4). This gives the final form of the FLDA model,

$$\mathbf{Y}_{ij} = S_{ij}(\boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots K, \quad j = 1, \ldots, m_i, \tag{6}$$

$$\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \sigma^2 I), \quad \boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, \Gamma),$$

in which $\lambda_0$, $\Lambda$ and $\boldsymbol{\alpha}_i$ are confounded if no constraint is imposed. Therefore we place the following restrictions on $\Lambda$ and the $\boldsymbol{\alpha}_i$'s,

$$\Lambda^T S^T \Sigma^{-1} S \Lambda = I, \qquad \sum_i \boldsymbol{\alpha}_i = 0, \tag{7}$$

where $\Sigma = \sigma^2 I + S\Gamma S^T$ and $S$ is the basis matrix evaluated over a fine lattice of points. The constraint provides a form of normalization for the linear discriminants. More details will be given in the following section. In practice the lattice should include, at least, all time points in the data set. For example the spinal bone density data was measured in 1/10th of a year increments from age 8.8 to 26.2 years so the lattice covered the same period. This model is identical to the general functional model of §1.2 with

$$\mu_i(t) = \mathbf{s}(t)^T(\boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i)$$

and

$$\omega(t, t') = \mathbf{s}(t)^T\Gamma\mathbf{s}(t').$$

## 3 Classifying curves

In this section we first detail a maximum likelihood procedure for fitting (6) and then a method for forming classifications by combining (1) and (6) to form an estimate of the Bayes classifier.

### 3.1 Fitting the model

Fitting the FLDA model involves estimating $\boldsymbol{\lambda}_0, \Lambda, \boldsymbol{\alpha}_i, \Gamma$ and $\sigma^2$. Notice that (6) implies

$$\mathbf{Y}_{ij} \sim N\left(S_{ij}(\boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i), \Sigma_{ij}\right),$$

where

$$\Sigma_{ij} = \sigma^2 I + S_{ij}\Gamma S_{ij}^T.$$

Since observations from different individuals are assumed to be independent, the joint distribution of the observed curves is

$$\prod_{i=1}^{K}\prod_{j=1}^{m_i}\frac{1}{(2\pi)^{n_{ij}/2}|\Sigma_{ij}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0+\Lambda\boldsymbol{\alpha}_i])^T\Sigma_{ij}^{-1}(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0+\Lambda\boldsymbol{\alpha}_i])\right]. \qquad (8)$$

A natural approach to fitting the model is to maximize (8) over $\boldsymbol{\lambda}_0, \Lambda, \boldsymbol{\alpha}_i, \Gamma$ and $\sigma^2$. Unfortunately, directly maximizing this likelihood is a difficult non-convex optimization problem. If the $\boldsymbol{\gamma}_{ij}$ had been observed, however, the joint likelihood of $\mathbf{Y}_{ij}$ and $\boldsymbol{\gamma}_{ij}$ would simplify to

$$\prod_{i=1}^{K}\prod_{j=1}^{m_i}\frac{1}{(2\pi)^{(n_{ij}+q)/2}\sigma^{n_{ij}}|\Gamma|^{1/2}}$$
$$\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0+\Lambda\boldsymbol{\alpha}_i+\boldsymbol{\gamma}_{ij}])^T(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0+\Lambda\boldsymbol{\alpha}_i+\boldsymbol{\gamma}_{ij}])-\frac{1}{2}\boldsymbol{\gamma}_{ij}^T\Gamma^{-1}\boldsymbol{\gamma}_{ij}\right].$$

Maximizing this likelihood is much less complex, and suggests treating the $\boldsymbol{\gamma}_{ij}$ as missing data and implementing the EM algorithm (Dempster *et al.*, 1977; Laird and Ware, 1982). The EM algorithm involves alternately calculating the expected value of the missing data $\boldsymbol{\gamma}_{ij}$ and maximizing the joint likelihood. The E step is performed using the equation

$$E(\boldsymbol{\gamma}_{ij}|\mathbf{Y}_i,\boldsymbol{\gamma}_0,\Lambda,\boldsymbol{\alpha}_i,\Gamma,\sigma^2) = (\sigma^2\Gamma^{-1}+S_{ij}^TS_{ij})^{-1}S_{ij}^T(\mathbf{Y}_{ij}-S_{ij}\boldsymbol{\lambda}_0-S_{ij}\Lambda\boldsymbol{\alpha}_i),$$

while the M step involves maximizing

$$Q = -\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{m_i}E\Big\{(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0-\Lambda\boldsymbol{\alpha}_i-\boldsymbol{\gamma}_{ij}])^T(\mathbf{Y}_{ij}-S_{ij}[\boldsymbol{\lambda}_0-\Lambda\boldsymbol{\alpha}_i-\boldsymbol{\gamma}_{ij}])/\sigma^2$$
$$+\, n_{ij}\log(\sigma^2)+\boldsymbol{\gamma}_{ij}^T\Gamma^{-1}\boldsymbol{\gamma}_{ij}+\log|\Gamma|\Big\}$$

holding $\boldsymbol{\gamma}_{ij}$ fixed. Further details can be obtained from the web site www-rcf.usc.edu/~gareth. As with all EM algorithms the likelihood will increase at each iteration but it is possible to reach a local rather than global maximum. This can be a problem for very sparse data sets such as the bone mineral density data. However, the problem is generally eliminated by enforcing a rank constraint on $\Gamma$ as discussed in §5.1.

Other model selection questions arise in practice, such as the choice of $q$, the dimension of the spline basis. There are several possible procedures that have been applied to models of this type. One is to calculate the cross-validated likelihood for various dimensions and choose the model corresponding to the maximum (James *et al.*, 2000). AIC and BIC are two other, less computationally expensive, procedures that have also proved successful on this sort of data (Rice and Wu, 2000). In practice the final classification appears to be relatively robust to any reasonable choice of dimension but this is an area of ongoing research.

### 3.2 Classification

Notice that under the standard reduced-rank LDA model,

$$\mathbf{X}|\text{Class} = i \sim N(\boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i, \Sigma).$$

Hence, using Bayes' formula, the probability of Class $i$ given $\mathbf{X}$ is proportional to

$$\begin{aligned}(\mathbf{X} - \boldsymbol{\lambda}_0 - \Lambda\boldsymbol{\alpha}_i)^T\Sigma^{-1}(\mathbf{X} - \boldsymbol{\lambda}_0 - \Lambda\boldsymbol{\alpha}_i) - 2\log\pi_i &= ||\mathbf{X} - \boldsymbol{\lambda}_0 - \Lambda\widehat{\boldsymbol{\alpha}}_{\mathbf{X}}||^2_{\Sigma^{-1}} + ||\Lambda\widehat{\boldsymbol{\alpha}}_{\mathbf{X}} - \Lambda\boldsymbol{\alpha}_i||^2_{\Sigma^{-1}} - 2\log\pi_i \\ &= C(\mathbf{X}) + ||\widehat{\boldsymbol{\alpha}}_{\mathbf{X}} - \boldsymbol{\alpha}_i||^2 - 2\log\pi_i\end{aligned}$$

where $\widehat{\boldsymbol{\alpha}}_{\mathbf{X}} = \Lambda^T\Sigma^{-1}(\mathbf{X} - \boldsymbol{\lambda}_0)$. Note that the second line follows from the fact that $\Lambda^T\Sigma^{-1}\Lambda = I$. This means that classifying an observation $\mathbf{X}$ using reduced-rank LDA is identical to classifying to

$$\arg\min_i \left(||\widehat{\boldsymbol{\alpha}}_{\mathbf{X}} - \boldsymbol{\alpha}_i||^2 - 2\log\pi_i\right).$$

It can be shown that $\widehat{\boldsymbol{\alpha}}_{\mathbf{X}}$ and $\boldsymbol{\alpha}_i$ are equal to the linear discriminants of $\mathbf{X}$ and $\boldsymbol{\mu}_i$ that LDA produces, up to an additive constant.

The same approach is used for FLDA. By combining (1) and (6) we see that the posterior probability that a curve $\mathbf{Y}$ was generated from Class $i$ is proportional to

$$(\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 - S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i)^T{\Sigma_{\mathbf{Y}}}^{-1}(\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 - S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i) - 2\log\pi_i,$$

where $S_{\mathbf{Y}}$ is the spline basis matrix for $\mathbf{Y}$ and

$$\Sigma_{\mathbf{Y}} = \sigma^2 I + S_{\mathbf{Y}}\Gamma {S_{\mathbf{Y}}}^T.$$

So $\mathbf{Y}$ will be classified to

$$\arg\min_i \left(||\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 - S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i||^2_{\Sigma_{\mathbf{Y}}^{-1}} - 2\log\pi_i\right). \tag{9}$$

Notice, however, that if one lets

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} = \left(\Lambda^T {S_{\mathbf{Y}}}^T\Sigma_{\mathbf{Y}}^{-1}S_{\mathbf{Y}}\Lambda\right)^{-1}\Lambda^T{S_{\mathbf{Y}}}^T\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0), \tag{10}$$

then

$$||\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 - S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i||^2_{\Sigma_{\mathbf{Y}}^{-1}} = ||\mathbf{Y}_{\mathbf{Y}} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 - S_{\mathbf{Y}}\Lambda\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}||^2_{\Sigma_{\mathbf{Y}}^{-1}} + ||S_{\mathbf{Y}}\Lambda\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} - S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i||^2_{\Sigma_{\mathbf{Y}}^{-1}}.$$

Hence (9) is equivalent to

$$\arg\min_i \left(||\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} - \boldsymbol{\alpha}_i||^2_{(\Lambda^T{S_{\mathbf{Y}}}^T\Sigma_{\mathbf{Y}}^{-1}S_{\mathbf{Y}}^T\Lambda)} - 2\log\pi_i\right) = \arg\min_i \left(||\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} - \boldsymbol{\alpha}_i||^2_{Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}})^{-1}} - 2\log\pi_i\right) \tag{11}$$

since

$$Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}) = \left(\Lambda^T{S_{\mathbf{Y}}}^T\Sigma_{\mathbf{Y}}^{-1}S_{\mathbf{Y}}\Lambda\right)^{-1}.$$

Just as with standard LDA, $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$ and $\boldsymbol{\alpha}_i$ are, up to an additive constant, the linear discriminants of $\mathbf{Y}$ and $\boldsymbol{\mu}_i$. Therefore (11) corresponds to classifying to the class whose mean is closest to our test point in the reduced space where distance is measured using the inverse covariance of $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$. Notice

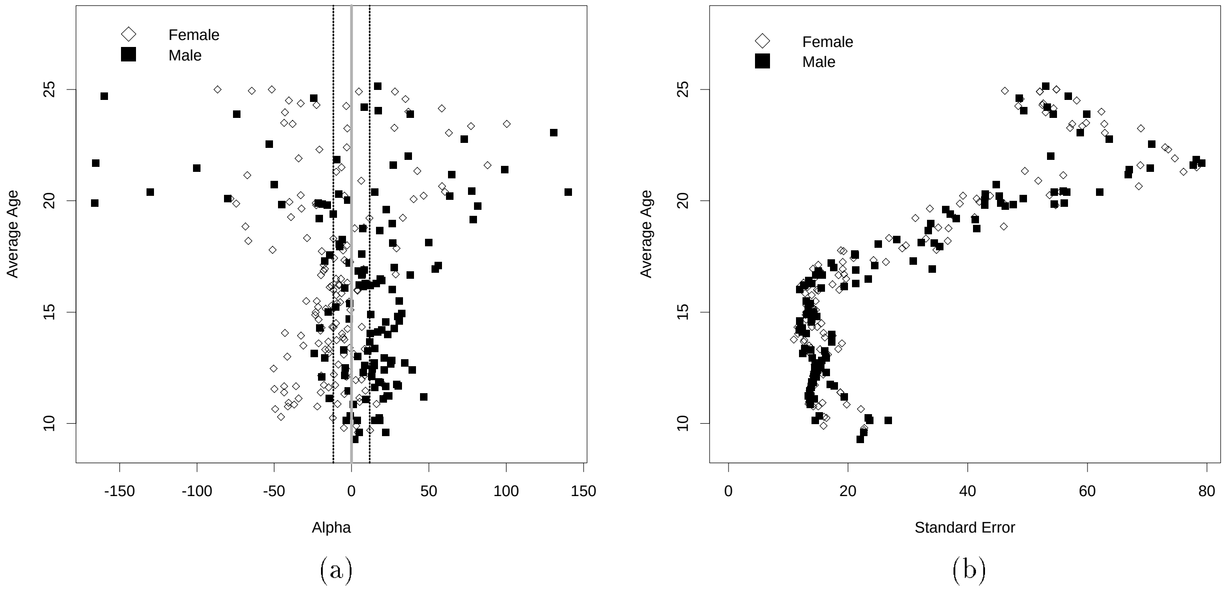



Figure 3: *(a) Linear discriminants for each curve of the spinal bone density data, plotted against the average age people were measured at. (b) Estimates of the standard error of the linear discriminants for each curve plotted against the average age. There is a clear trend of increasing variability with age.*

also that if $\mathbf{Y}$ has been measured over the entire time period, so that $S_{\mathbf{Y}} = S$, then

$$Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}) = I$$

and (11) reduces to

$$\arg\min_i \left(||\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} - \boldsymbol{\alpha}_i||^2 - 2\log\pi_i\right).$$

# 4 Applications of FLDA

In this section we show how three of the most important tools that LDA provides, namely, low dimensional representation, discrimination functions and classification, can be replicated using FLDA.

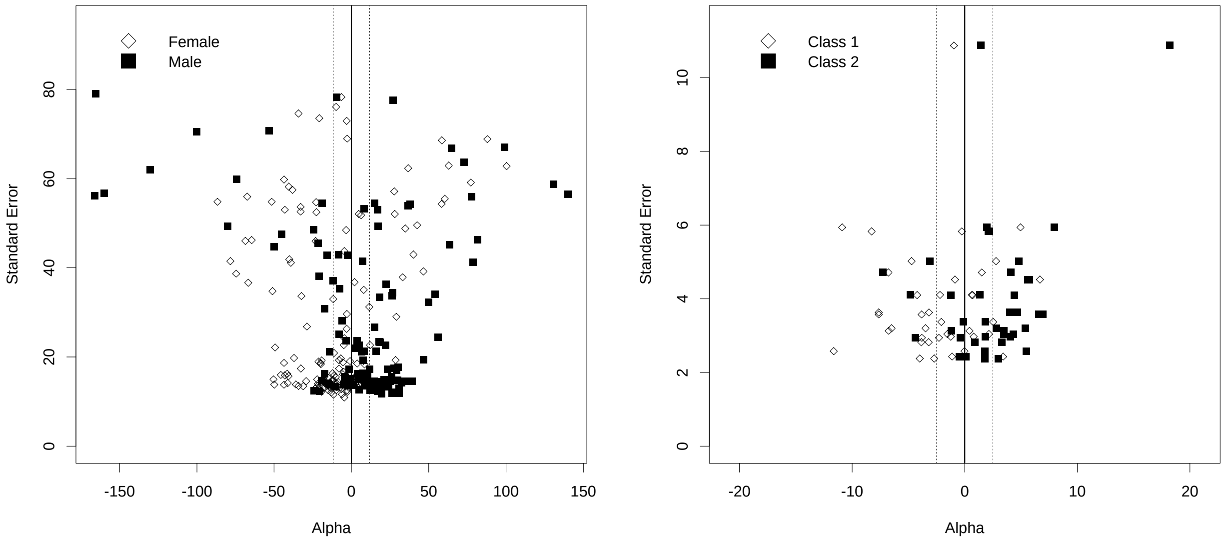
## 4.1 Low dimensional representation of curves

One of the reasons for the popularity of LDA is that it provides the ability to view high-dimensional data by projecting it onto a low-dimensional space. This allows one to visually determine the discrimination between classes. As mentioned in §3.2, in a standard finite-dimensional setting the linear discriminant of $\mathbf{X}$ equals

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{X}} = \Lambda^T\Sigma^{-1}(\mathbf{X} - \boldsymbol{\lambda}_0),$$

up to an additive constant. As $Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{X}}) = I$, the transformed variables all have identity covariance, so the distance between different observations is Euclidean and can be easily calculated by visual inspection.

This provides a natural approach to projecting functional data into a low-dimensional space. In




(a) (b)

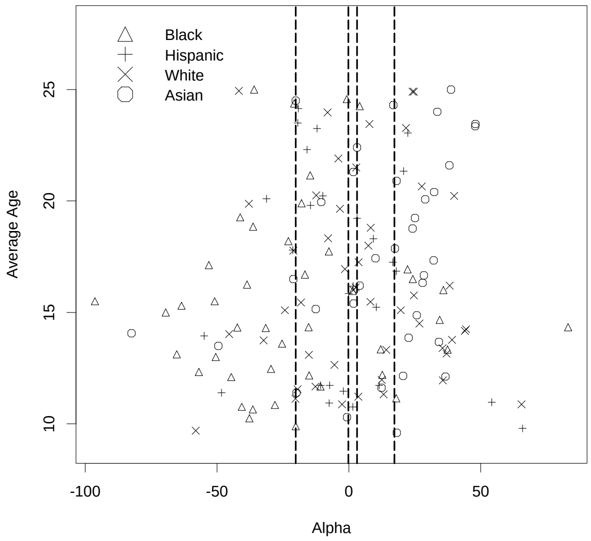
Figure 4: *(a) Plot of the linear discriminants for each curve on the spinal bone density data versus the estimated standard errors. (b) A linear discriminant plot for a simulated data set. It shows a fairly clear difference between the classes but also a significant overlap.*

the FLDA model the analogue of $\widehat{\boldsymbol{\alpha}}_{\mathbf{X}}$ is $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$, given in (10). Recall that $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$ is the linear discriminant for $\mathbf{Y}$ and that if $\mathbf{Y}$ has been observed over the entire interval $Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}) = I$. However, if only fragments of the curve have been observed,

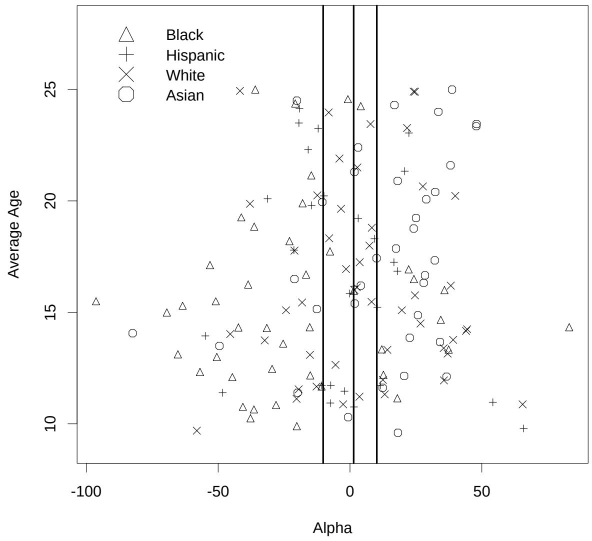
$$Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}) = \left(\Lambda^T S_{\mathbf{Y}}{}^T \Sigma_{\mathbf{Y}}^{-1} S_{\mathbf{Y}} \Lambda\right)^{-1}.$$

This makes direct comparison of points more difficult because the covariance structure may no longer be diagonal and, in general, curves measured at different time points will have different covariances. However, if $h = 1$, so the linear discriminant is a scalar, the only effect this has is that each point has a different standard error. Figures 3-6 provide examples of this.

Figure 3(a) shows linear discriminants for each curve from the growth curve data of Figure 1, plotted versus the average age for observations from each individual. For a two-class situation, such as this, the plot also provides a simple classification rule; curves with positive linear discriminant are classified as male and curves with negative linear discriminant as female. The plot reveals some interesting properties of the data. Curves measured at ages below eighteen years are relatively well-separated while individuals measured at older ages have little discernible separation. This trend is also apparent upon close examination of the original curves. The two solid vertical lines either side of zero give the fitted values for the $\boldsymbol{\alpha}_i$'s. They represent the "class centroids" in the transformed space. Their close proximity to each other relative to the variability of the linear discriminants indicate little overall separation between classes. However, recall that, as a result of (7), the linear discriminant of $\mathbf{Y}$ will have a standard error of one if the curve is measured over the entire interval. In fact the separation between the centroids is about 23.5 standard deviations. This implies that the confusion between genders is a result of the small number of observations per individual, which cause the standard error to increase dramatically. A clear separation could be achieved with more



(a)



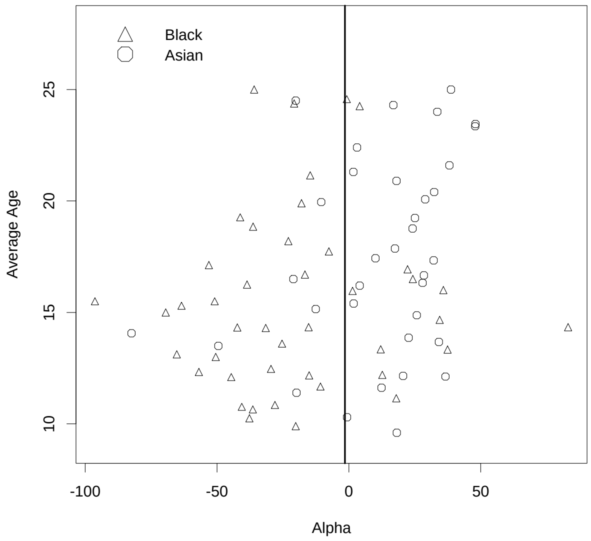
(b)

Figure 5: *Linear discriminants for the spinal bone density data using ethnicity as the class variable. (a) The dotted lines represent, from left to right, class centroids for Blacks, Hispanics, Whites, and Asians. Notice that Blacks and Asians are fairly well separated while Hispanics and Whites are not. (b) The solid lines represent decision boundaries for classifying a given curve.*

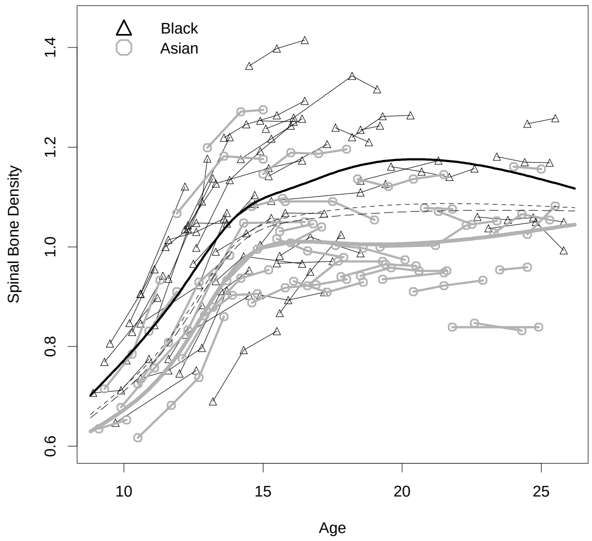
observations. Figure 3(b) gives the standard error for each linear discriminant versus average age of observation. A curve with measurements over the entire time period would have standard error 1, so this plot gives an indication of the amount of information lost by only observing the curve at a limited number of time points. The standard errors range from over ten to eighty, indicating that a great deal of accuracy has been sacrificed. The increased variability also explains the poor separation at older ages where the standard errors are large relative to the distance between the class centroids. Once the model has been fit, the standard error can be calculated for a curve observed at an arbitrary set of time points. This provides a method for deciding on an "optimal" design in terms of locating a finite number of observations for an individual to minimize the standard error.

Figure 4 gives two plots which combine the linear discriminants and their standard errors together. This gives an easy method for deciding on the reliability of a given observation. For example points with high standard error should be treated with caution. We call these *linear discriminant plots*. The left and right vertical dotted lines indicate the class centroids while the center lines are the class discriminators. Figure 4(a) shows that the points with relatively low standard error have far better separation than those with a large standard error. Figure 4(b) provides a similar plot for a simulated data set consisting of 80 curves measured at the same set of time points as that of the data set illustrated in Figure 2. Notice that even though each curve has been measured at fairly evenly spaced points throughout the time interval the standard errors still range up to ten. The two classes are relatively well separated but there is still some clear overlap. The distance between the class centroids is 5 standard deviations, indicating that one could achieve near perfect separation by sampling the curves at a wider range of time points.

Figures 5 and 6 are further linear discriminant plots. They were produced by using ethnicity as the class variable on a subset of the growth curve data, all females. A plot of the growth curves for



(a)



(b)

Figure 6: *(a) Linear discriminants for Blacks and Asians. While there is still some overlap the separation is far clearer. The vertical line gives the classification boundary. (b) A plot of the raw data for Blacks and Asians. Notice that Blacks tend to have a higher bone density. The two solid lines represent the means for Blacks and Asians while the dotted lines are for Hispanics and Whites.*

each ethnicity (not shown) indicates that there may be no clear separation between classes. This is borne out by Figure 5(a) which gives a plot of linear discriminant versus average age. There is significant overlap between the classes. However, it is still possible to gain some information. The four vertical dotted lines represent the class centroids for, from left to right, Blacks, Hispanics, Whites and Asians. It is clear that there is very little separation between Hispanics and Whites while Blacks and Asians are relatively well separated. This is highlighted by Figure 6 which shows linear discriminants and a plot of the raw data for Blacks and Asians alone. The differences are now clear. Figure 5(b) is identical to 5(a) except that the three discrimination boundaries are plotted in place of the class centroids. The discrimination boundaries divide the space into four regions. Points in the leftmost region are classified as Black, the next as Hispanic, followed by White and finally Asian.

## 4.2 Classification

The ability of LDA to perform classification is of equal importance to its ability to explain discrimination between classes. In §3.2 we showed that to classify a curve using FLDA one need only produce $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$ using (10) and classify using (11). When all classes have equal weight and $\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}$ is one dimensional this classification rule simplifies to

$$\arg\min_i (\widehat{\alpha}_{\mathbf{Y}} - \alpha_i)^2 \tag{12}$$

While classification is not the primary goal on the spinal bone density data, we apply (12) to it to illustrate the procedure. When using gender as the class variable the overall training error rate comes out at 29.3%. However, the rate increases substantially to 44.1% for ages over 18 and

| | | True Ethnicity | | | | |
|---|---|---|---|---|---|---|
| | | Asian | Black | Hispanic | White | Total |
| Prediction | Asian | 22(62.9) | 9(20.9) | 8(30.0) | 19(39.6) | 58 |
| | Black | 7(20.0) | 30(69.8) | 10(37.0) | 13(27.1) | 60 |
| | Hispanic | 1(2.9) | 2(4.7) | 5(18.5) | 7(14.6) | 15 |
| | White | 5(14.3) | 2(4.7) | 4(14.8) | 9(18.8) | 20 |
| | Total | 35(100) | 43(100) | 27(100) | 48(100) | 153 |

Table 1: *Confusion matrix of classifications for the four ethnicities. The numbers in parentheses give the percentages of each ethnicity receiving the corresponding classification. Asians and Blacks have relatively little confusion while Hispanics and Whites have a great deal.*

decreases to 22.0% for ages under 18. This conforms to our expectations from Figure 3(a) which shows much better discrimination for lower ages.

Table 1 gives the confusion matrix when ethnicity is used as the class variable. It shows the true ethnicity and the corresponding classification for each of the 153 individuals. For example, 22 of the 35 Asians were classified as Asian while 10 of the 27 Hispanics were classified as Black. While the overall training error rate is 56.9%, a large fraction of the errors are among Hispanics and Whites, while Asians and Blacks are relatively well classified. When Asians and Blacks are considered alone the error rate drops to 25%.

In Table 2 we present the results from a simulation study where the FLDA procedure is compared with two other classifiers. The first is the filtering method of §1.1. Recall that the filtering method consists of fitting a flexible basis, in this case cubic splines, to each curve and then classifying by using LDA on the basis coefficients. The filtering method provides a simple comparison to FLDA. The second is the Bayes classifier which is optimal if the true distribution of the classes is known. It provides the best case error rate. The study consisted of a three-class problem. For each class, 50 curves were generated according to the FLDA model (6) and sampled on a fine grid of 100 equally-spaced points. Then, for each curve, a random subset, 50-94%, of the observations, were removed to replicate curve fragments. Multiple data sets were created and the FLDA and filtering procedures were applied to each. Error rates were then calculated on a separate, test set, of 300 curves. Furthermore the Bayes error rate, which is the lowest possible, was also calculated on this test set. Over the 100 time points the average deviation of mean curves between Classes 1 and 2 and between Classes 2 and 3 was 0.066 while it was twice this number between Classes 1 and 3. The standard deviation of the error terms was $\sigma = 0.1$. Finally the average standard deviation over the 100 time points of the random curves $S\gamma$ was approximately 0.368. The first figure gives a guide as to the signal while the last two indicate the noise.

Table 2 provides a summary of the test error rates. As one would expect, all three sets of error rates increase with the fraction of missing data. Of far more interest is the similarity between the FLDA and Bayes error rates. Even with 90% of the data removed the difference is only 0.5%. With less than 80% of the data removed the filtering and FLDA methods give comparable results. However, the filtering method deteriorates rapidly until at 94% its error rate is close to that of the naive classifier which randomly assigns a class label based on the prior probability for each class. Note that at 94% the filtering method could not even be applied to several of the simulated data sets because individual curves could not be fitted.

| % Missing | Filtering | FLDA | Bayes |
|---|---|---|---|
| 50 | 0(0) | 0(0) | 0 |
| 80 | 9.5(0.4) | 8.6(0.2) | 8.3 |
| 84 | 18.8(0.5) | 12.6(0.3) | 12.7 |
| 90 | 40.8(1.9) | 17.2(0.3) | 16.7 |
| 94 | 60.4(1.1) | 28.1(0.5) | 26.7 |

Table 2: *Test error rates from the simulation study for various different fractions of missing data. The numbers in parentheses indicate estimated standard errors for the error rates.*

### 4.3 Class discrimination functions

In a standard two-class LDA setting the discriminant function is defined as

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1}$$

where $\Sigma$ is the within group covariance matrix. This function gives the weight put on each dimension in determining the classification of a point. In the FLDA setting $\boldsymbol{\mu}_i = S(\boldsymbol{\lambda}_0 + \Lambda\boldsymbol{\alpha}_i)$ so the functional analogue is

$$(S\Lambda\boldsymbol{\alpha}_1 - S\Lambda\boldsymbol{\alpha}_2)^T \Sigma^{-1} = (\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)^T \Lambda^T S^T \Sigma^{-1} \qquad (13)$$

where $\Sigma = \sigma^2 I + S\Gamma S^T$ and $S$ is the spline basis matrix evaluated on a fine grid of points over the entire time period. Equation (13) can be used to produce a discriminant function for any set of data. Figure 7 provides examples from the growth curve data. Figure 7(a) gives the discriminant function using gender as the class variable. There is a strong negative peak before age 15 and a large positive peak afterwards; this indicates a phase shift between genders and explains why there is far better separation for the earlier ages. Figure 7(b) gives a similar plot using ethnicity as the class variable. Again most of the discrimination appears to be in the early years.

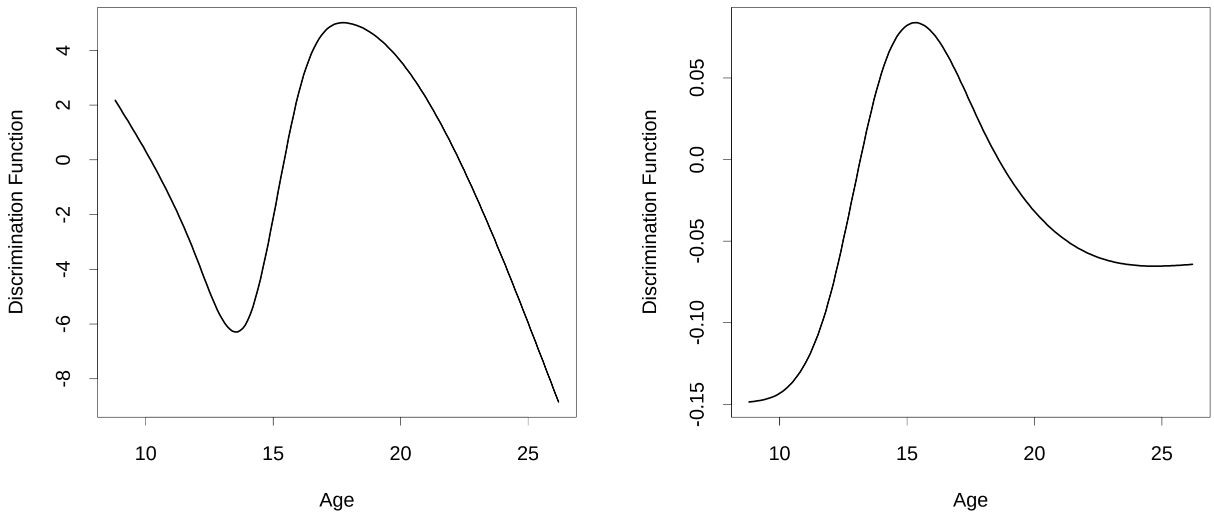
A comparison of discriminant functions produced from the simulation study of §4.2, using both the FLDA and filtering approaches, is given in Figure 8. The population discriminant function is shown in black while its estimates are in grey. Figures 8(a) and (b) present results from the study with 84% of the data removed while Figures 8(c) and (d) present results with 90% removed. For each, the top plot shows the discriminant functions from 40 different simulations using FLDA, while the bottom plot gives the corresponding graph for the filtering approach. As adding or multiplying the discriminant function by a constant leaves the classification unchanged, the estimates have been transformed to produce the least squares fit to the population discriminant function. It is clear that in both cases the FLDA approach is producing more accurate discriminant functions, reflected in the decreased error rates of Table 2.

## 5 Extensions

Under the standard FLDA model

$$Cov(\mathbf{Y}_{ij}) = \sigma^2 I + S_{ij}\Gamma S_{ij}^T. \qquad (14)$$

In this section we consider a number of possible extensions to this model by exploring different assumptions for $\Gamma$ and hence the covariance matrix of $\mathbf{Y}_{ij}$.




(a) (b)

Figure 7: *Discriminant functions for the growth curve data of Figure 1. (a) Using gender as the class variable. There is a strong indication of phase shift. (b) Using ethnicity as the class variable.*

## 5.1 Reduced Rank Covariance Matrices

Under (14), no restrictions are placed on the structure of $\Gamma$. In practice, the likelihood function for data sets such as the spinal bone density data has a large number of local maxima which make this model difficult to fit. As a result, for even moderate $q$, one can produce a highly variable fit to the data. James *et al.* (2000) show that such problems can be reduced by enforcing a rank constraint on $\Gamma$. A rank $p$ constraint is equivalent to setting $\Gamma = \Theta D \Theta^T$ and

$$Cov(\mathbf{Y}_{ij}) = \sigma^2 I + S_{ij} \Theta D \Theta^T S_{ij}^T$$

where $\Theta$ is a $q \times p$ $(p < q)$ matrix and $D$ is a diagonal matrix.

James *et al.* (2000) suggest several methods for choosing the rank and conclude that the optimal rank for these data is $p = 2$. The results from Section 4 were produced using such a reduced rank model with $p = 2$ because this gave a far better fit to the data.

## 5.2 Functional Quadratic Discriminant Analysis

The FLDA procedure, in analogy with LDA, makes an assumption of a common covariance matrix, $\Gamma$, for the $\gamma_{ij}$ vectors from each class. This can result in a considerable reduction in variance for classes with a small sample size. However, the assumption can cause a considerable increase in bias if the covariances are not common. In a standard setting *quadratic discriminant analysis (QDA)* provides a less restrictive procedure by allowing different covariance matrices. In a similar manner the FLDA model of §2.2 can be generalized by removing the assumption of a common covariance term, $\Gamma$. This gives the covariance structure

$$Cov(\mathbf{Y}_{ij}) = \sigma^2 I + S_{ij} \Gamma_i S_{ij}^T,$$

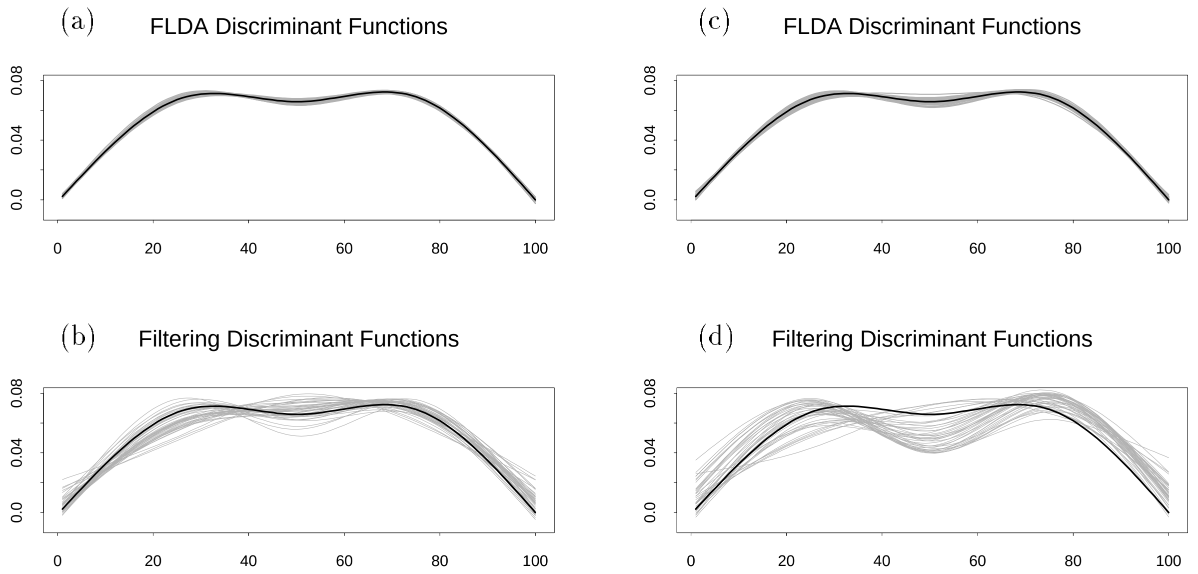



Figure 8: *Results from two different sets of simulations. Plots (a) and (b) show the true discriminant function (black line) and estimates (grey lines) from 40 different simulations using FLDA (a) and filtering (b) methods with* 84% *of each curve unobserved. Plots (c) and (d) show the equivalent results with* 90% *of each curve unobserved.*

where $\Gamma_i$ is the covariance matrix for Class $i$. The posterior probability is now proportional to

$$d_i(\mathbf{Y}) = ||\mathbf{Y} - S_{\mathbf{Y}}\boldsymbol{\lambda}_0 + S_{\mathbf{Y}}\Lambda\boldsymbol{\alpha}_i||^2_{\Sigma^{-1}_{i_{\mathbf{Y}}}} + \ln|\Sigma^{-1}_{i_{\mathbf{Y}}}| - 2\log\pi_i.$$

where

$$\Sigma_{i_{\mathbf{Y}}} = \sigma^2 I + S_{\mathbf{Y}}\Gamma_i {S_{\mathbf{Y}}}^T$$

By fitting this model and classifying to $\arg\min_i d_i(\mathbf{Y})$ a generalization of QDA, which we call *functional quadratic discriminant analysis (FQDA)* is produced.

## 5.3 Functional Regularized Discriminant Analysis

It is well known that LDA can perform badly if the assumption of a common within-class covariance matrix is violated, while QDA generally requires a larger sample size (Wald and Kronmal, 1977). A small sample size causes a covariance matrix to be produced which is close to singular and hence excessive weight is placed on the directions corresponding to small eigenvalues. Regularization has been highly successful in this sort of poorly-posed inverse problem (Titterington (1985), O'Sullivan (1985)). Friedman (1989) suggests the following regularization approach. Let $S_i/n_i$ be the within class sample covariance matrix for class $i$ and let $S = \sum_{i=1}^K S_i$. Then $S/n$ is the pooled covariance matrix which is used for LDA while $S_i/n_i$ is used for QDA. A compromise between the two approaches can be achieved by setting the within-class covariance matrix equal to

$$\hat{\Gamma}_i(w_1) = S_i(w_1)/n_i(w_1),$$

where

$$S_i(w_1) = (1 - w_1)S_i + w_1 S \quad \text{and} \quad n_i(w_1) = (1 - w_1)n_i + w_1 n.$$

A second level of regularization, namely shrinkage towards the identity matrix, is provided through

$$\hat{\Gamma}_i(w_1, w_2) = (1 - w_2)\hat{\Gamma}_i(w_1) + \frac{w_2}{q} tr[\hat{\Gamma}_i(w_1)]I, \tag{15}$$

where $q$ is the dimension of the space. $\hat{\Gamma}_i(w_1, w_2)$ is used as the within-class covariance matrix for the $i$th class. Friedman calls this approach *regularized discriminant analysis (RDA)*. RDA has been shown to outperform both LDA and QDA in a large variety of situations.

A generalization to *functional regularized discriminant analysis (FRDA)* can be achieved using the following covariance structure

$$Cov(\mathbf{Y}_{ij}) = \sigma^2 I + S_{ij}\hat{\Gamma}_i(w_1, w_2)S_{ij}^T$$

where $\hat{\Gamma}_i(w_1, w_2)$ is defined as in (15). The choice of $w_1$ and $w_2$ is made using cross-validation. Applying cross-validation to the FRDA model is potentially computationally expensive. However, in the RDA setting, an algebraic update allows for a significantly faster implementation (Friedman, 1989). This update can be used in the FLDA setting by fitting the FQDA model, treating the resulting $\gamma_{ij}$'s as $q$ dimensional data and fitting RDA to estimate $w_1$ and $w_2$, and finally fitting the FRDA model with $w_1$ and $w_2$ fixed.

FRDA contains both FLDA and FQDA as sub models. By setting both $w_1$ and $w_2$ equal to zero the FQDA model is produced. While setting $w_1 = 1$ and $w_2 = 0$ produces the FLDA model. Furthermore, by setting $w_1 = w_2 = 1$ a functional generalization of the nearest-means classifier is produced where an observation is assigned to the closest class mean, in Euclidean distance.

# 6 Conclusion

We have presented a method, which we call functional linear discriminant analysis (FLDA), for generalizing linear discriminant analysis to functional data. FLDA possesses all the usual LDA tools, including a low-dimensional graphical summary of the data, and classification of new curves. When the functional data have been measured over a large number of time points the procedure provides similar results to the filtering method introduced in Section 1.1. However, when only fragments of the function are available the FLDA approach can still produce favorable outcomes while the filtering method fails completely. FLDA can also be generalized in a number of ways. A reduced rank version can be implemented when the data are very sparse and a quadratic version can be used when an assumption of a common covariance matrix is inappropriate. A regularized version, which is a compromise between FLDA and FQDA, is also available.

Another possible generalization, which we have not considered, is to model heterogeneous variance and autocorrelation in the error terms. This may well improve the classification accuracy of the method provided enough time points have been observed per curve to provide accurate estimates. Unfortunately, allowing $\sigma^2$ to vary would make it impossible to enforce the constraint,

$$\Lambda^T S^T \Sigma^{-1} S\Lambda = I, \tag{16}$$

which ensures that, if a curve is measured at all time points, $Cov(\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}}) = I$. In turn (16) allows figures such as 3(b) to provide a measure of the amount of information lost through missing observations. If $\sigma^2$ was allowed to vary this interpretation would no longer be feasible.

**Acknowledgments**

The authors would like to thank the Editor and referees for many constructive suggestions. Trevor Hastie was partially supported by grants from the National Science Foundation and the National Institutes of Health.

## References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* **22**, 327–351.

Bachrach, L. K., Hastie, T. J., Wang, M. C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth; a longitudinal study. *Journal of Clinical Endocrinology & Metabolism* **84**, 4702–4712.

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

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Ser. B* **39**, 1–22.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics, Part A - Theory and Methods* **A5**, 843–854.

DiPillo, P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of classification. *Communications in Statistics, Part A - Theory and Methods* **A8**, 1447–1457.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

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

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach.* Chapman and Hall: London.

Hastie, T. J., Buja, A., and Tibshirani, R. J. (1995). Penalized discriminant analysis. *Annals of Statistics* **23**, 73–102.

Hastie, T. J. and Tibshirani, R. J. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B, Methodological* **58**, 155–176.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

O'Sullivan, F. (1985). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–527.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer.

Rice, J. A. and Wu, C. O. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics (Under review)* .

Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review* **53**, 141–170.

Wald, P. W. and Kronmal, R. A. (1977). Discriminant functions when covariances are unequal and sample sizes moderate. *Biometrics* **33**, 479–484.