Classification using censored functional data

Aurore Delaigle and Peter Hall
Department of Mathematics and Statistics, University of Melbourne, Australia.

Abstract: We consider classification of functional data when the training curves are not observed on the same interval. Different types of classifier are suggested, one of which involves a new curve extension procedure. Our approach enables us to exploit the information contained in the endpoints of these intervals by incorporating it in an explicit but flexible way. We study asymptotic properties of our classifiers, and show that, in a variety of settings, they can even produce asymptotically perfect classification. The performance of our techniques is illustrated in applications to real and simulated data.

Keywords: Bagging, cross-validation, functional principal components, incomplete curves, perfect classification, quadratic discriminant.

1 Introduction

The topic of classifying functional data has received a great deal of attention over the last decade. In the standard problem, a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is available, where, for each $i = 1, \ldots, n$, X_i is a curve coming from one of G groups and observed on a compact interval $\mathcal{I}_0$, and Y_i is its group label ($Y_i = g$ if X_i comes from group g, where $g \in \{1, 2, \ldots, G\}$). Using the training sample, the goal is to construct a classifier that can identify the group label of new curves whose group is unknown. A variety of techniques have been suggested for classifying such functional data. See for example Hall et al. (2001), Ferraty and Vieu (2003), Glendinning and Herbert (2003), Vilar and Pertega (2004), Biau et al. (2005), Fromont and Tuleau (2006), Huang and Zheng (2006), Leng and Müller (2006), López-Pintado and Romo (2006), Rossi and Villa (2006), Cuevas et al. (2007), Wang et al. (2007), Berlinet et al. (2008), Epifanio (2008), Song et al. (2008), Araki et al. (2009), Chamroukhi et al. (2010) and Delaigle et al. (2012).

The majority of existing methods rely on the fact that the data are observed on the same interval $\mathcal{I}_0$. In practice, functional data are usually observed only at a

discrete number of points, and the curves on $\mathcal{I}_0$ are obtained by joining the discretely observed points or by smoothing them using, for example, spline or kernel methods; see Ramsay and Silverman (2005) and Ferraty and Vieu (2006) for introductions to functional data analysis. A difficulty in applications is that the discrete points do not always cover the same range of values; the associated functional curves can be supported on quite different intervals, and standard methods of analysis cannot be used.

In this paper we are interested in constructing classifiers for curves of this type. More precisely, we consider classification of functions supported on a compact interval $\mathcal{I}$, in cases where the training sample consists of functions observed on other intervals, which may differ among the training curves. This classification problem was studied by James and Hastie (2001), but under somewhat restrictive parametric assumptions (see our discussion in section 2.3). We wish to develop more flexible, nonparametric techniques.

We propose several methods, depending on the nature of the curves, and make three novel contributions. First, we suggest a nonparametric approach to extending curves outside the interval where they were observed. Other extension methods were suggested by James et al. (2000), James and Hastie (2001) and Yao et al. (2005), but in contrast with our approach they rely on parametric distributional assumptions (see our discussion in section 3.2). Our function extension methodology is particularly flexible, and reflects this advantage by enjoying lower error rates when used to construct classifiers for both real and simulated data. It can be combined with bagging (Breiman, 1996) to further reduce classification error.

Second, we introduce flexible ways of combining potential differences in shapes of the curves from different populations, and potential differences between the endpoints of the intervals where the curves were observed. Indeed, in at least some applications the intervals themselves could contain information about the population where the curves originated. We provide theory showing that this information is often not used

effectively by conventional classifiers for functional data, and so should be introduced explicitly in another form. In effect, earlier contributions to the problem of classification from fragmentary functional data conditioned on the endpoints. The advantages of using endpoints explicitly are also clear from our numerical work.

Thirdly, we show that the perfect classification property of the linear discriminant classifier described by Delaigle and Hall (2012) can be extended from the context of conventional functional datasets to the setting of fragments of functions. Delaigle and Hall (2012) treated only the setting of differences in means, but in the more general context studied here we show that, with a quadratic discriminant classifier, asymptotically perfect classification can occur when there are differences between the two mean functions, or the two eigenvalue sequences, or the two eigenfunction sequences, or when there are differences simultaneously in two or three of these features.

The paper is organised as follows. In section 2 we introduce the problem; recall linear and quadratic discriminant classification procedures for standard functional data, and introduce basic methodology in cases where the function fragments overlap significantly. Section 3 details our methodology in cases where the function fragments overlap relatively little; it involves the development of a new curve extension algorithm. Section 4 introduces methods for combining data on interval endpoints; section 5 outlines theoretical properties; section 6 shows how to choose the tuning parameters and provides detail about implementation; and section 7 illustrates the methods' performance when applied to real and simulated data. The appendices give technical detail behind the results in section 5, and discuss additional results.

2 Model and challenges

2.1 Definitions of data and classification problem

For simplicity we introduce our methods in the case where the data come from two groups, or populations, i.e. when $G = 2$. Extensions to larger values of G are straight-

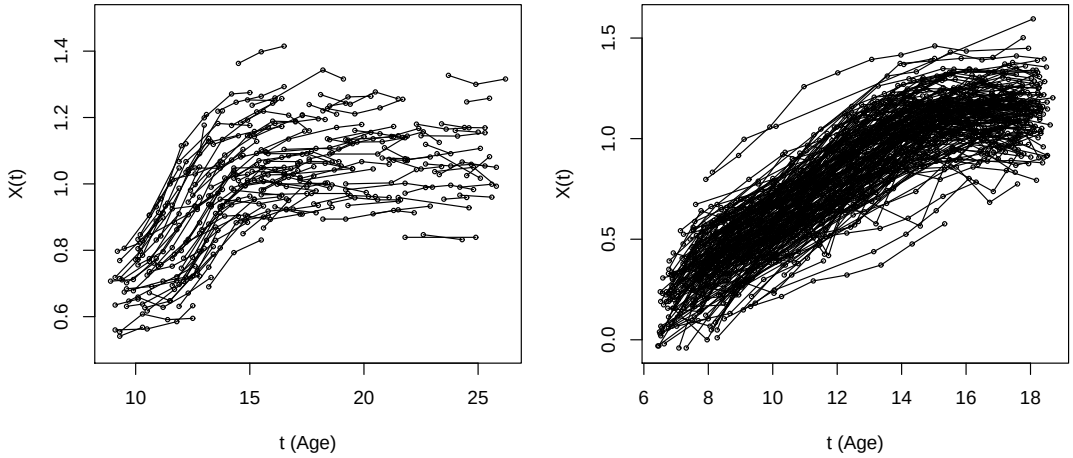

Figure 1: Left: fragments of growth curves of 153 females. Right: curves of pulmonary function for 252 US girls.

forward and will be discussed in section 7.2.2.

We observe training samples $\mathcal{X}_k = \{(X_{k1}, \mathcal{I}_{k1}), \ldots, (X_{kn_k}, \mathcal{I}_{kn_k})\}$, for $k = 0, 1$, where X_{kj} is a random function defined on a compact interval $\mathcal{I}_0$ but observed only on a compact set $\mathcal{I}_{kj} \subseteq \mathcal{I}_0$, and the $\mathcal{I}_{kj}$s are not necessarily identical. In other words, the observations are fragments of curves that remain after restricting to $\mathcal{I}_{kj}$ the full curves $X_{kj}(t)$ for $t \in \mathcal{I}_0$. We assume that, for $k = 0, 1$, the uncensored functions $X_{k1}, \ldots, X_{kn_k}$, with support $\mathcal{I}_0$, come from population Π_k, and are independent and identically distributed (i.i.d.), but are observed only in censored form, i.e. restricted to $\mathcal{I}_{kj}$, where the censoring mechanism may be different for each pair (k, j). Given a new function X, defined on the interval $\mathcal{I}_0$ but observed only on a set $\mathcal{I} \subseteq \mathcal{I}_0$, we wish to use the data pairs $(X_{kj}, \mathcal{I}_{kj})$ to classify $(X, \mathcal{I})$ as coming from Π_0 or Π_1.

Remark 1. The assumption that the "original" curves were all supported on the same interval $\mathcal{I}_0$ is used only to imply that there is a distribution of curves on $\mathcal{I}_0$, and that the distribution on $\mathcal{I}_{kj}$ corresponds to the distribution of curves supported on $\mathcal{I}_0$, restricted to $\mathcal{I}_{kj}$. In particular, curves observed on $\mathcal{I}_{kj}$ do not actually have to have existed on $\mathcal{I}_0 \setminus \mathcal{I}_{kj}$.

We consider two types of data. A first setting is that where the individual curves

are each observed on quite different intervals, which overlap only a little. For example, the left panel of Figure 1, depicting data from James and Hastie (2001), shows observed fragments of growth curves of 153 females; see section 7.2.2 for a description. Observations of the same individual are represented by the symbol •, and have been joined to form fragments of curves. A second, simpler setting is that where the curves are observed on intervals that are different, but many of which overlap at least partially. For example, the right panel of Figure 1 shows curves depicting the evolution of lung function as a function of age, for 252 US girls; see section 7.2.3 for more details. Observations of the same individual are again indicated by linked • symbols to form fragments.

Such data differ from the so-called sparse functional data studied by, for example, Yao et al. (2005). There it is typically assumed that the data come from n curves, and that, for $i = 1, \ldots, n$, the ith curve is observed only at a small number of points T_{ij}, where $j = 1, \ldots, N_j$. The N_js are assumed to be i.i.d., and the T_{ij}s are also assumed to be i.i.d. Clearly, the data shown in neither panel of Figure 1 are of this type, since the time points at which each individual is observed are not independent of each other (in both datasets, they are planned yearly examinations), and what we observe can be treated as fragments of each curve, rather than scattered, remote, and somewhat disconnected, points from each curve.

2.2 Classifiers for standard functional data

To construct classifiers based on curve fragments it would be difficult to use highly sophisticated methods, such as those based nonparametric regression estimators. Indeed, since we observe only pieces of curves, those approaches would tend to introduce too much noise to the classification decision. Therefore, as in James and Hastie (2001), our procedures are based on linear discriminant (LD) and quadratic discriminant (QD) classifiers. In a multivariate context, when the data are vectors V_i, the QD (which coincides with the Bayes' classifier if the data are normally distributed)

ascribes a new observation V to Π_0 if

$$\mathcal{S} \equiv (V - \bar{V}_0)^{\mathrm{T}} \hat{\Sigma}_0^{-1} (V - \bar{V}_0) + \log |\hat{\Sigma}_0| - (V - \bar{V}_1)^{\mathrm{T}} \hat{\Sigma}_1^{-1} (V - \bar{V}_1) - \log |\hat{\Sigma}_1| + w \quad (2.1)$$

is negative, and to Π_1 otherwise, where $\bar{V}_k$ and $\hat{\Sigma}_k$ are the empirical mean and variance in group Π_k, for $k = 0, 1$, and $w = -2 \log\{\pi/(1-\pi)\}$, with π denoting the prior probability of Π_0. The LD classifier is defined in the same way, except that $\hat{\Sigma}_k$, for $k = 0, 1$, is replaced by $\hat{\Sigma}$, an estimator of the common covariance.

A relatively conventional way of adapting the LD and QD methods to the standard functional context, where the data curves are observed on the entire interval $\mathcal{I}_0$, consists in applying these classifiers to finite dimensional projections obtained by empirical spectral analysis, as we describe next; see also Delaigle and Hall (2012) for the LD classifier. Write E_k for expectation when a data function X is drawn from Π_k, and let $\mu_k(t) = E_k\{X(t)\}$ and $K_k(s,t) = E_k\{X(s)\,X(t)\} - \mu_k(s)\,\mu_k(t)$ be, respectively, the mean and covariance in group k (here we assume that $E_k(\|X\|_0^2) < \infty$ for $k = 0, 1$, where $\|X\|_0^2 = \int_{\mathcal{I}_0} X^2$).

Consider a spectral decomposition of the covariance $K_k(s,t)$ for $(s,t) \in \mathcal{I}_0$:

$$K_k(s,t) = \sum_{j=1}^{\infty} \theta_{kj}\, \phi_{kj}(s)\, \phi_{kj}(t)\,, \quad (2.2)$$

where (θ_{kj}, ϕ_{kj}), for $k = 0, 1$ and $j = 1, \ldots, n_k$, are the respective (eigenvalue, eigenfunction) pairs for the transformation τ_k that takes a function ψ, defined on $\mathcal{I}_0$, to $\tau_k\psi$ defined by $(\tau_k\psi)(s) = \int_{\mathcal{I}_0} K_k(s,t)\,\psi(t)\,dt$. The sequence $\phi_{k1}, \phi_{k2}, \ldots$ is an orthonormal basis for the class of square-integrable functions on $\mathcal{I}_0$, and terms in the series at (2.2) are ordered so that $\theta_{k1} \geq \theta_{k2} \geq \ldots$.

Let X be a new function defined on $\mathcal{I}_0$, which we wish to classify. A Karhunen-Loève expansion of $X(t) - \mu_k(t)$, appropriate for $t \in \mathcal{I}_0$, is given by

$$X(t) - \mu_k(t) = \sum_{j=1}^{\infty} \theta_{kj}^{1/2}\, \xi_{kj}\, \phi_{kj}(t)\,, \quad (2.3)$$

where $\xi_{kj} = \theta_{kj}^{-1/2} \int_{\mathcal{I}_0} \{X(t) - \mu_k(t)\}\, \phi_{kj}(t)\,dt$ is the standardised jth principal component score when X is interpreted as coming from Π_k.

Estimators of the mean and covariance functions of each group Π_k are given by

$$\bar{X}_k(t) = n_k^{-1} \sum_{j=1}^{n_k} X_{kj}(t)\,, \tag{2.4}$$

$$\widehat{K}_k(s,t) = n_k^{-1} \sum_{j=1}^{n_k} \{X_{kj}(s) - \bar{X}_k(s)\}\,\{X_{kj}(t) - \bar{X}_k(t)\}\,. \tag{2.5}$$

Empirical approximations $(\hat{\theta}_{kj}, \hat{\phi}_{kj})$ to (θ_{kj}, ϕ_{kj}) can be derived from an expansion analogous to (2.2), for $\widehat{K}_k(s,t)$ rather than $K_k(s,t)$:

$$\widehat{K}_k(s,t) = \sum_{j=1}^{n_k} \hat{\theta}_{kj}\,\hat{\phi}_{kj}(s)\,\hat{\phi}_{kj}(t)\,,$$

where $\hat{\theta}_{k1} \geq \hat{\theta}_{k2} \geq \ldots$ and $\hat{\phi}_{k1}, \hat{\phi}_{k2}, \ldots$ is an orthonormal sequence of functions. The empirical counterpart of ξ_{kj} is $\hat{\xi}_{kj} = \hat{\theta}_{kj}^{-1/2} \int_{\mathcal{I}_0} \{X(t) - \bar{X}_k(t)\}\,\hat{\phi}_{kj}(t)\,dt$.

Motivated by the expansion in (2.3), in the functional context, a conventional approach is to use a classifier based on (2.1), where the vectors $V - \bar{V}_k$ are taken to be the first j_0 nonstandardised principal component scores $\theta_{k1}^{1/2}\,\xi_{k1}, \ldots, \theta_{kj_0}^{1/2}\xi_{kj_0}$ coming from (2.3). Since these are uncorrelated and have variances $\theta_{k1}, \ldots, \theta_{kj_0}$, then, after replacing ξ_{kj} and θ_{kj} by their empirical versions, the following QD classifier for functional data obtains: assign X to Π_0 if the statistic

$$T_{\rm fun}(X, \mathcal{I}_0 \,|\, j_0, w) = T_0(X, \mathcal{I} \,|\, j_0) - T_1(X, \mathcal{I} \,|\, j_0) + w \tag{2.6}$$

is negative, and to Π_1 otherwise, where, for $k = 0, 1$,

$$T_k(X, \mathcal{I}_0 \,|\, j_0) = \sum_{j=1}^{j_0} \left(\hat{\xi}_{kj}^2 + \log \hat{\theta}_{kj}\right), \tag{2.7}$$

and, as above, $w = -2\log\{\pi/(1-\pi)\}$. The subscript on $T_{\rm fun}$ denotes "function."

The LD classifier is defined in the same way, except that it assumes that $K_0 = K_1 \equiv K$, where K can be estimated by $\widehat{K}(s,t) = n^{-1} \sum_{k=0,1} \sum_{j=1}^{n_k} \{X_{kj}(s) - \bar{X}_k(s)\}$ $\{X_{kj}(t) - \bar{X}_k(t)\}$. Asymptotic properties of this LD classifier have been studied by Delaigle and Hall (2012), who showed that, as n increases, it can reach near perfect classification performance. In section 5.1 we show that the QD classifier can

reach asymptotic near perfect classification in more general settings, and in the more complex case of fragmental observations treated in section 2.3.

2.3 Basic fragment classifier

In our context, since the curves are not observed on the entire interval $\mathcal{I}_0$, we cannot apply the classifiers as described in section 2.2. We could instead apply them on the interval $\mathcal{I} \subseteq \mathcal{I}_0$ where the fragment of the curve X to classify was observed. The next few paragraphs discuss simple ways of doing this. However, such procedures cannot always be used, and below we explain why new methodology is required. In section 4 we shall also argue that, even in cases where the basic classifier can be applied, it can be improved by incorporating information contained in the endpoints of the intervals $\mathcal{I}_{kj}$.

For s and $t \in \mathcal{I}$, let $n_k(t)$ and $n_k(s,t)$ denote the numbers of indices j, for $1 \leq j \leq n_k$, such that $t \in \mathcal{I}_{kj}$ and $(s,t) \in \mathcal{I}_{kj} \times \mathcal{I}_{kj}$, respectively, and $\mathcal{J}_k(s)$ and $\mathcal{J}_k(s,t)$ are the respective sets of those indices. Given a new function X defined on $\mathcal{I}$, and an integer $\nu_0 \geq 1$, write $\mathcal{I}'$ and $\mathcal{I}''$ for subsets of $\mathcal{I}$ defined by

$$\mathcal{I}' = \left\{ t \in \mathcal{I} : \min[n_0(t), n_1(t)] \geq \nu_0 \right\}, \tag{2.8}$$

$$\mathcal{I}'' \times \mathcal{I}'' = \left\{ (s,t) \in \mathcal{I} \times \mathcal{I} : \min \left[n_0(s,t), n_1(s,t), n_0(t,s), n_1(t,s) \right] \geq \nu_0 \right\}, \tag{2.9}$$

respectively. The sets $\mathcal{I}'$ and $\mathcal{I}'' \times \mathcal{I}''$ are the subsets of $\mathcal{I}$ and $\mathcal{I} \times \mathcal{I}$ on which we can compute, for each k, simple estimators of μ_k and K_k, respectively, using at least ν_0 data. More precisely, for $s \in \mathcal{I}'$ and $(t,u) \in \mathcal{I}'' \times \mathcal{I}''$, we can estimate μ_k and K_k by

$$\bar{X}_k(s) = \frac{1}{n_k(s)} \sum_{j \in \mathcal{J}_k(s)} X_{kj}(s), \tag{2.10}$$

$$\widehat{K}_k(t,u) = \frac{1}{n_k(t,u)} \sum_{j \in \mathcal{J}_k(t,u)} \{X_{kj}(t) - \bar{X}_k(t)\} \{X_{kj}(u) - \bar{X}_k(u)\}. \tag{2.11}$$

These standard estimators have mean square convergence rates of at least $O(\nu_0^{-1})$.

Alternatively, we can use smoothed versions of the estimators at (2.10) and (2.11), for example Yao et al.'s (2005) local linear estimators constructed from the pooled

observations from all curves. Since it is a univariate smoother, the local linear mean estimator and its data-driven smoothing parameter are simple and fast to compute. However, the smooth covariance estimator is based on bivariate smoothing techniques, which are often time consuming for our classification problem. Moreover, this smooth covariance estimator is not guaranteed to be a covariance (the one at (2.11) is). Therefore, and since smoothing is not crucial to classification performance (see our numerical results in section 7), our preference is to use (2.11) to estimate covariances.

Then we can use the QD method described in section 2.2, but replacing $\mathcal{I}_0$ there by the interval $\mathcal{I}''$, and replacing $\bar{X}_k$ and $\widehat{K}_k$ by the definitions given above. (In the LD case, where we assume that $K_0 = K_1 \equiv K$, K can be estimated by $\widehat{K}(s,t) = \sum_{k=0,1} \sum_{j \in \mathcal{J}_k(s,t)} \{X_{kj}(s) - \bar{X}_k(s)\} \{X_{kj}(t) - \bar{X}_k(t)\} / \sum_{k=0,1} n_k(s,t)$.) In particular, here the scores $\hat{\xi}_{kj}$ and the eigenvalues $\hat{\theta}_{kj}$ are obtained from a spectral decomposition of $\widehat{K}_k$ on the interval $\mathcal{I}''$. Moreover, θ_{kj}, ϕ_{kj} and ξ_{kj} are defined as in section 2.1, but with $\mathcal{I}_0$ replaced everywhere by $\mathcal{I}''$. Choice of j_0 will be discussed in section 6.1.

We refer to this approach as the basic fragment classifier. Our main contributions to this simple setting are:

- developing theory showing that it can lead to asymptotically perfect classification in a variety of settings (see section 5.1);
- introducing a new method that can improve this basic classifier by using explicitly the endpoints of the intervals $\mathcal{I}_{kj}$ (see section 4).

While the basic fragment classifier can be very effective for data of the type of the pulmonary example in section 2.1, where the fragments are numerous and long, it cannot be used with data of the type in the growth example in section 2.1, where the fragments are short and scattered. Figure D.3 in Appendix D shows the observed fragments of curves in each of the four groups. There, we can see that the fragments within each group are sparse and overlap very little. Clearly, with such data, (2.10) and (2.11) cannot give good estimators because too few fragments overlap at each t. Moreover, to compute $\widehat{K}_k(s,t)$ reliably, we need to have a reasonable number of

fragments overlapping the interval $[s, t]$. However, only when s and t are very close to each other do some fragments overlap, which means that we can compute $\widehat{K}_k$ only on short intervals $\mathcal{I}''$. Moreover, even for short $\mathcal{I}''$ we have access to only a few fragments for computing $\widehat{K}_k(s, t)$ at each (s, t). Of course, the smoothed mean and covariance versions of Yao et al. (2005) suffer from the same difficulties. In particular, their method can estimate $K(s, t)$ only on very short intervals $\mathcal{I}''$, and even there, their estimator is usually not accurate because it is computed from too few observations.

A solution was proposed by James and Hastie (2001), who proceed by modeling the curves by splines with random coefficients; see also James et al. (2000). These coefficients are assumed to follow a multivariate normal distribution, with mean depending on the group. The number of parameters used is reduced through rank-based constraints, and the parameters are fitted by maximum likelihood using the EM algorithm, to be substituted finally into a version of the LD classifier (one that is not based on spectral decomposition). A drawback of their approach is that it relies on a number of parametric and other types of assumptions. These are perhaps necessary when the data are sparse, in the sense that the observations for each curve are "remote" points. Nevertheless, when the data can be treated as fragments it is possible to construct classifiers that rely less on parametric assumptions.

Thus, it seems that new methodology is required for data of the type in the growth example in section 2.1. In section 3 we take a new approach to this problem.

2.4 Correlated errors and classification

It might reasonably be thought that one could utilize the correlation information in experimental errors for the purpose of estimation and classification, and thereby gain improved performance. However, this expectation turns out to be particularly difficult to realise, for at least two reasons. First, in the case of longitudinal data, for example the datasets shown in either panel of Figure 1, the data are typically recorded at relatively distant, and often irregularly spaced, time points, for example by medical

personnel at a health clinic at the times of a patient's annual visits there. In such cases the errors are likely to be independent, not least because they are recorded so far apart in time. Quite reasonably, this motivates the assumption of independent errors, made by, for example, Yao et al. (2005) and James and Hastie (2001). In such cases there is not much opportunity for achieving extra performance by modeling error correlation.

Secondly, although in the case of machine recorded data, and some other data types, experimental errors may be correlated, it can be particularly difficult to access the information contained in that structure. For example, digital devices typically manipulate data internally in sophisticated ways, often nonlinearly, before outputting information, with the result that the noise in even so-called "raw" data is embedded in the signal. Fortunately, many of the devices in question have relatively low noise levels; this is typically the case for infrared spectroscopy, which is the source of the chemometric data analysed in section 7.2.1. Therefore, in such cases there is little practical opportunity for extracting useful information for classification by modeling error correlation—either that information is very difficult to access, because processing has embedded the errors nonlinearly in the signal, or the errors are so small as to make them a minor source of information.

3 Curve classifier for short and scattered fragments

3.1 Introduction

In this section we introduce new methodology for data of the type of the growth example in section 2.1, where the observed fragments are too scattered for it to be possible to construct reasonable estimators of the group means and covariances using conventional approaches such as those discussed in section 2.3. We suggest a two-stage procedure: first, extend the fragments of curves to the interval $\mathcal{I}$ where we want to perform classification; second, apply a classifier based on these extended curves.

3.2 Extending curves

Methods for reconstructing, or predicting, missing parts of curves have been suggested by several authors before, but because they were especially designed for sparsely observed curves recorded at i.i.d. time points, so far the existing techniques have focused on relatively strong parametric assumptions. As already discussed in section 2.3, in James et al. (2000) and James and Hastie (2001) the curves are modeled by splines with normally distributed random coefficients. These coefficients are determined by their best linear unbiased predictors, under constraints that reduce the number of parameters that have to be estimated. In Yao et al. (2005) the curves are expressed through their Karhunen-Loève expansions, where the means and covariances are estimated by kernel smoothers. The eigenfunction estimators are computed from the smoothed covariance estimators, and these eigenfunction estimators are used in conjunction with normality assumptions to construct predictors of the principal component scores for each curve. As already indicated in section 2.3, the mean and covariance smoothers that they employ to estimate scores cannot perform well in the context of short and sparse fragments, and therefore the associated curve predictors cannot perform well either.

As a result, these methods can produce curves which do not necessarily share the main features of the population, which in turn can affect classification performance. In this section we suggest a new nonparametric approach to extending a curve. The extension is obtained by adjoining small fragments of curves, each obtained by copying a vertically translated version of a piece of one of the observed fragments X_i.

More formally, suppose we observe a curve X_{short} on an interval $\mathcal{I}_{\text{short}} = [a_{\text{short}}, b_{\text{short}}]$, and we want to estimate the unobserved parts of the curve on an interval $\mathcal{I}_{\text{long}} = [a_{\text{long}}, b_{\text{long}}] \supset \mathcal{I}_{\text{short}}$, using n curves X_i, $i = 1, \ldots, n$, observed on respective intervals $\mathcal{I}_i = [a_i, b_i]$, such that $\mathcal{I}_{\text{short}} \subset \mathcal{I}_{\text{long}} \subseteq \bigcup_{i=1}^{n} \mathcal{I}_i \subseteq \mathcal{I}_0$. To construct the extended curve X_{ext} to the right of b_{short}, we proceed as follows:

1. For all $t \in [a_{\text{short}}, b_{\text{short}}]$, let $X_{\text{ext}}(t) = X_{\text{short}}(t)$. Let $j = 1$ and $b_{\text{ext},j} = b_{\text{short}}$.

2. For $j = 1, 2, \ldots$, repeat the following steps while $b_{\text{ext},j} < b_{\text{long}}$:

   (a) Find all the curve fragments X_i for which the associated interval $\mathcal{I}_i = [a_i, b_i]$ satisfies $a_i \leq b_{\text{ext},j}$ and $b_i > b_{\text{ext},j}$, and choose one of them (see details below), X_{i^*} say; in this notation, X_{i^*} is observed on $\mathcal{I}_{i^*} = [a_{i^*}, b_{i^*}]$.

   (b) Let $b_{\text{ext},j+1} = \min(b_{i^*}, b_{\text{long}}, b_{\text{ext},j} + \Delta)$, where $\Delta > 0$ is a tuning parameter.

   (c) For each $t \in [b_{\text{ext},j}, b_{\text{ext},j+1}]$, let $X_{\text{ext}}(t) = X_{i^*}(t) - X_{i^*}(b_{\text{ext},j}) + X_{\text{ext}}(b_{\text{ext},j})$.

The same algorithm is applied to the left of a, by adjoining small pieces one at a time in the same way as above, but from right to left.

To apply the algorithm in practice we need to choose Δ. The role of Δ is just to prevent us from copying an overly long fragment of curve (this prevents a single curve from having too much effect on the final reconstructed curve). Since our goal is to use the extended curves for classification, in principle Δ could be chosen to minimise a cross-validation estimator of classification error. However, in many cases, this is an unnecessary complication, and Δ can simply be chosen by the experimenter so that the pieces of curves used to construct X_{ext} are sufficiently short, and, overall, a fragment of length Δ contains at most a small fraction of the strong features (mode, change of concavity, etc) present on $\mathcal{I}_0$. In most cases where we are interested in extending curves, the observed fragments are short and sparse anyway, and the influence of Δ is limited. We suggest using the default value $\Delta = |\mathcal{I}_0|/10$, unless the curves have rapidly changing features, in which case Δ could be taken smaller. (Here $|\mathcal{I}_0|$ denotes the length of I_0.)

We also need to determine a way of selecting the curve X_{i^*} in step 2 (a). Suppose that, in step 2(a), we have c_j curve fragments X_i, $i = c_1, \ldots, c_j$, for which $a_i \leq b_{\text{ext},j}$ and $b_i > b_{\text{ext},j}$. We suggest two ways of choosing the curve X_{i^*}.

The first approach consists in choosing X_{i^*} at random among the c_j available fragments; each fragment is chosen with a probability $p_{ij} = 1/c_j$. The idea here is to construct a sample of extended curves which shares the main properties of a sample of full curves coming from the underlying population, and which can therefore be used

to estimate the mean and covariance functions that we need to compute discriminant classifiers. Despite its simplicity, this approach can be particularly effective for classification in a variety of settings, for example when $X_{kj}(t) = \mu_k(t) + \Delta_{kj}(t) + Z_{kj}$, where Z_{kj} is a random variable, and where Δ_{kj} is either weakly dependent, or $\Delta_{kj}(t) = \eta_{kj}\,\delta_{kj}(t)$ where η_{kj} is small and δ_{kj} is a fixed bounded process. This will be illustrated in section 7 and can also be shown by theoretical analysis; see section 5.2.

The second, more sophisticated approach can be used when the curves within a group have noticeable shape similarities, such as in the growth example in section 2.1. There, locally, the shape of a curve is similar to that of "nearby" curves. In this case a fragment can be extended by using a translated version of the "nearest" fragment. More formally, suppose we are interested in extending X_{ext} to the right of $b_{\text{ext},j}$, and let $D(X_i, X_{\text{ext}}; b_{\text{ext},j})$ denote a distance measuring the similarity between X_i and X_{ext} near $b_{\text{ext},j}$. For example, if, as in the case of the growth data, the shapes of fragments depend (at least locally) on their location on the vertical axis, we can take $D(X_i, X_{\text{ext}}; b_{\text{ext},j}) = |X_i(b_{\text{ext},j}) - X_{\text{ext}}(b_{\text{ext},j})|$. This distance can be used in many practical applications, but it can be modified to take into account other criteria, such as the derivative near $b_{\text{ext},j}$. Here, in step 2(a) of the algorithm, we take $i^* = \text{argmin}_{i=c_1,\ldots,c_j} D(X_i, X_{\text{ext}}; b_{\text{ext},j})$. See Appendix B.1 for a more formal discussion of this approach.

Remark 2. When the fragments are so sparse that there are parts of $\mathcal{I}_{\text{long}}$ where no curve fragment has been observed, the curves could be extended on those parts by a parametric method of the type described in James and Hastie (2001), or by linear extrapolation, or by the smooth mean estimator of Yao et al. (2005) shifted vertically.

3.3 Classifier based on extended curves

Let X, observed on $\mathcal{I} = [a, b] \subseteq \mathcal{I}_0$, be the new curve fragment that we wish to classify. For $k = 0, 1$ and $j = 1, \ldots, n_k$, let $\tilde{X}_{kj}$ denote the extended versions of X_{kj} obtained using the algorithm described in section 3.2, taking there $\mathcal{I}_{\text{long}} \supseteq \mathcal{I}$ and where, to

construct $\tilde{X}_{kj}$, we use only the observed fragments from group k. As we extend a curve further from the original interval $\mathcal{I}_{kj} = [a_{kj}, b_{kj}]$ on which its fragment X_{kj} was observed, the extension becomes more unreliable. Therefore, we suggest estimating the mean and covariance functions of each group Π_k by giving to each curve $\tilde{X}_{kj}$ a weight w_{kj} depending on a measure of the distance between the intervals $\mathcal{I}_{kj}$ and $\mathcal{I}$.

Specifically, let $h_k > 0$ be a bandwidth and L a kernel function, and define

$$w_{kj} = L\big(\|\mathcal{I} - \mathcal{I}_{kj}\|_{\text{int}}/\tilde{h}_k\big) \Big/ \sum_{\ell=1}^{n_k} L\big(\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\text{int}}/\tilde{h}_k\big)\,,$$

where $\tilde{h}_k = (b-a)h_k$ and $\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\text{int}} = (b - b_{kj}) \cdot 1\{b_{kj} < a\} + \{b - b_{kj} + \max(a, a_{jk}) - a\} \cdot 1\{a \le b_{kj} \le b\} + \{\max(a, a_{kj}) - a\} \cdot 1\{b_{kj} > b\}$. We take

$$\bar{X}_k(t) = \sum_{j=1}^{n_k} w_{kj}\, \tilde{X}_{kj}(t)\ ,\ \widehat{K}_k(s,t) = \sum_{j=1}^{n_k} w_{kj}\, \{\tilde{X}_{kj}(s) - \bar{X}_k(s)\}\, \{\tilde{X}_{kj}(t) - \bar{X}_k(t)\}\,,$$

and use the classification procedure described in section 2.2, i.e. the method based on $T_{\text{fun}}(X, \mathcal{I}_0 \,|\, j_0, w)$ in (2.6) but replacing there $\mathcal{I}_0$ by the interval $\mathcal{I}$ where the curve X is observed, and replacing $\bar{X}_k$ and $\widehat{K}_k$ by the definitions given above. The choice of tuning parameters will be discussed in section 6.1.

With the random version of the algorithm described in section 3.2, the function extension method includes elements of the bootstrap, in that random sampling from the data is used to select the curve fragments that are chosen when extending X_{kj} from $\mathcal{I}_{kj}$ to a larger interval containing $\mathcal{I}$. Therefore the extended versions of X_{kj} and $\mathcal{I}_{kj}$ could quite reasonably be denoted by X^*_{kj} and $\mathcal{I}^*_{kj}$, and for different versions, say N_B versions, of the set of revised training data $(X^*_{kj}, \mathcal{I}^*_{kj})$ we can potentially have different classification decisions. These decisions can be combined by assigning the new pair $(X, \mathcal{I})$ to whichever population received the majority out of these N_B decisions. This is the bagging method introduced by Breiman (1996), and it can improve performance in complex statistical problems by reducing the impact of stochastic variability.

4 Using interval endpoints

The classifiers can be improved by noting that the endpoints of the intervals themselves may contain valuable information about the population from which the curves originated. For example, they may represent the ends of a time period in a person's life when certain health-related measurements have been found necessary, and the beginning or end of that period may convey important information about the classification problem. However, this information is largely ignored by classifiers based on curves, such as the basic one in section 2.3 and the more sophisticated one in section 3.3, since these use only the information contained in the interval $\mathcal{I}$. To exploit the information potentially contained in the endpoints, we suggest combining curve-based classifiers with classifiers based explicitly on the endpoints of the intervals.

4.1 Classifier based on endpoints

First we introduce a classifier based solely on the endpoints of the intervals $\mathcal{I}_{kj}$ where the curves were observed. Write $\mathcal{I}_{kj} = [A_{kj}, B_{kj}]$, let

$$\bar{A}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} A_{kj}\,, \quad \bar{B}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} B_{kj}\,, \quad \hat{\sigma}^2_{Ak} = \frac{1}{n_k} \sum_{j=1}^{n_k} (A_{kj} - \bar{A}_k)^2\,,$$

$$\hat{\sigma}^2_{Bk} = \frac{1}{n_k} \sum_{j=1}^{n_k} (B_{kj} - \bar{B}_k)^2\,, \quad \hat{\gamma}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} (A_{kj} - \bar{A}_k)\,(B_{kj} - \bar{B}_k)$$

denote the means, variances and covariances of the interval endpoints, and let $\widehat{\Sigma}_k = (\hat{\sigma}^2_{Ak}, \hat{\sigma}^2_{Bk}; \hat{\gamma}_k)$ be the corresponding 2×2 covariance matrix.

To classify a new data curve X observed on the interval $\mathcal{I} = [A, B]$, using only the endpoints of the intervals $\mathcal{I}$ and $\mathcal{I}_{kj}$, we can use the QD which ascribes X to Π_0 if $S_{\text{int}}(\mathcal{I} \,|\, w)$ is negative, and to Π_1 otherwise, where $S_{\text{int}}(\mathcal{I} \,|\, w)$ is defined by $\mathcal{S}$ in (2.1), with V there replaced by (A, B). The subscript on S_{int} denotes "interval." Since (A, B) is of dimension only 2, here we choose w from the data (see section 6.2). For the same reason, if n were large then we could also use more sophisticated classifiers such as nonparametric Bayes methods; see section C.1 in Appendix C.

4.2 Combined classifier based on curves and endpoints

Each of the curve classifiers in sections 2.3 and 3.3, and the endpoint classifier in section 4.1, can be combined into a single approach that takes into account both curve shapes and endpoint locations. We link the two approaches in a flexible way which allows us to choose the emphasis adaptively. Our idea is to use the classifier based on $T_{\rm fun}$, unless the decision of the classifier based on $S_{\rm int}$ is sufficiently authoritative, in which case we use the latter. (Here, $T_{\rm fun}$ denotes the version of the statistic at (2.6) computed either as in section 2.3 or as in section 3.3.) Specifically, we define a new discriminating statistic T by

$$T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1) = \begin{cases} S_{\rm int}(\mathcal{I} \,|\, w_2) & \text{if } S_{\rm int} > \hat{q}_{S,\alpha_1} \\ S_{\rm int}(\mathcal{I} \,|\, w_2) & \text{if } S_{\rm int} < \hat{q}_{S,\alpha_0} \\ T_{\rm fun}(X, \mathcal{I} \,|\, j_0, w_1) & \text{otherwise}\,, \end{cases} \tag{4.1}$$

where $w_1 = -2\,\log\{\pi/(1-\pi)\}$, $0 \le w_2 < \infty$, and, for $k = 0, 1$, $\hat{q}_{S,\alpha_k}$ denotes the empirical quantile α_k of the distribution of $S_{\rm int}$ conditional on $(-1)^{k+1} S_{\rm int} \ge 0$. See section 6.4 for a detailed description of the calculation of these quantiles. (Here and below, $S_{\rm int}$ is defined as in section 4.1.)

Our suggested combined classifier is only one of the many possible ways of combining two classifiers. Formula (4.1) responds to experience gained applying various methods to real and simulated datasets, where we discovered that using $T_{\rm fun}$ alone was often best in the majority of cases, but not always in a substantial majority. In the remaining cases, $S_{\rm int}$ should be used alone. Compromise classification criteria, for example based on $\pi\, S_{\rm int} + (1-\pi)\, T_{\rm fun}$ where $\pi \in (0, 1)$ was chosen to optimise a cross-validation estimator of classification error, usually performed more poorly than the method at (4.1).

The choice of parameters will be discussed in section 6.3. We assign X to Π_0 or Π_1 according as $T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1)$ is negative or positive, respectively. In the event that other covariate information is available, for example information about the reason for parts of function curves being missing, it could be introduced explicitly to

the classifier using methods similar to those employed to include endpoint information. If the data curves were reduced to more than one fragment then the number of fragments could be treated in this way. In particular, covariate information of these types can be incorporated using the method employed to address information contained in endpoints.

The operation of adjoining, to a classifier C1, say, based on the curves (and implicitly on the endpoints), another classifier C2 based explicitly on the endpoints, thereby obtaining a new classifier C3 represented by the statistic at (4.1), does not damage the perfect classification property if C1 enjoyed that feature. In particular, C3 has the perfect classification property.

In practice we suggest using cross-validation to estimate consistently the error rates of different classifiers, and thus to compare those methods. In this way we can compare, for example, the error rate of the classifier based on $T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1)$, at (4.1), with that of the method based on $T_{\rm fun}(X, \mathcal{I} \,|\, j_0, w_1)$ alone, and adopt the technique with least estimated error. This adaptive approach allows us to experiment with a small number of different methods without risking anything more than a minor loss of performance. Consistency of standard leave-one-out cross-validation, when used to estimate error rates of methods such as the centroid classifier or Fisher's linear or quadratic discriminant, is straightforward although tedious to establish.

5 Theoretical properties

5.1 Asymptotically perfect classification of basic classifier

In this section we show that the basic QD classifier in section 2.3 can achieve asymptotic near-perfect performance. Excepting degenerate cases, this level of accuracy is not possible for conventional data, for example in the context of vector valued data. However, in the context of functional data there is potential for asymptotically perfect classification, by exploiting any of the ways in which the distribution of X can

differ from one population to another.

A perfect classification property was established by Delaigle and Hall (2012) for the LD classifier, in the case where the curves are all observed on the same interval $\mathcal{I}_0$. Below we show that this property holds in a much wider variety of settings, and when the curves are observed only on parts of $\mathcal{I}_0$. Three different sources of difference are readily accessible if we use a classifier based at least in part on the version $T_{\rm fun}(X, \mathcal{I} \,|\, j_0, w)$, used in section 2.3, of the quantity $T_{\rm fun}(X, \mathcal{I}_0 \,|\, j_0, w)$ at (2.6), i.e. with all estimators computed as in section 2.3, and $\mathcal{I}_0$ replaced by $\mathcal{I}$. These are differences between (i) the means μ_0 and μ_1, (ii) the eigenvalue sequences $\theta_{01}, \theta_{02}, \ldots$ and $\theta_{11}, \theta_{12}, \ldots$, and (iii) the eigenfunction sequences $\phi_{01}, \phi_{02}, \ldots$ and $\phi_{11}, \phi_{12}, \ldots$. We shall show in section B.2 in Appendix B that any of the differences (i), (ii) and (iii) can produce classifiers that result in classification error converging to zero as n_0 and n_1 diverge.

For simplicity, in Theorem 1 below we treat only data on the function fragments X, not on the interval endpoints, since, except in pathological cases, the endpoints alone cannot enable perfect classification. We take $\mathcal{I}_0 = [0, 1]$ and base classification on the criterion $T_{\rm fun}(X, \mathcal{I} \,|\, j_0, w)$, and we put the prior probability π equal to $1/2$, which assumption is made commonly in practice and will be imposed in our numerical work. Specifically, we assign $(X, \mathcal{I})$ to Π_0 if

$$T(X, \mathcal{I} \,|\, j_0) = \sum_{j=1}^{j_0} \left(\sum_{k=0}^{1} (-1)^k \frac{1}{\hat\theta_{kj}} \left[\int_{\mathcal{I}''} \{X(t) - \bar X_k(t)\}\, \hat\phi_{kj}(t)\, dt \right]^2 + \log \frac{\hat\theta_{0j}}{\hat\theta_{1j}} \right) \quad (5.1)$$

is negative, and to Π_1 if $T(X, \mathcal{I} \,|\, j_0)$ is positive. Here, all estimators are computed as in section 2.3.

Define $\delta_{kj} = \theta_{kj}^{-1/2} \int_{\mathcal{I}} (\mu_0 - \mu_1)\, \phi_{kj}$ and

$$V_{k_1 k_2 j} = \theta_{k_1 j}^{-1/2} \int_{\mathcal{I}} \left\{ X^{(k_2)} - E_{k_2}\big(X^{(k_2)}\big) \right\} \phi_{k_1 j}\,, \quad (5.2)$$

where $X^{(k)}$ denotes a random function drawn from population Π_k, and $k_1, k_2 \in \{0, 1\}$. In particular, V_{kkj} is, for each j, a standardised principal component score and has

zero mean and unit variance. Finally, let $Q_1(k_1, k_2 \,|\, j_0) = \sum_{j \le j_0} \{V_{k_1 k_1 j}^2 - V_{k_2 k_1 j}^2 + \log(\theta_{k_1 j}/\theta_{k_2 j})\}$ and $Q_2(k_1, k_2 \,|\, j_0) = \sum_{j \le j_0} \{(-1)^{k_1} V_{k_2 k_1 j}\, \delta_{k_2 j} + \delta_{k_2 j}^2\}$.

The following theorem summarises the main properties of the QD classifier in section 2.3. It describes in detail the large-sample properties of error rate, including but not confined to cases where error rate is asymptotically zero. The conditions of the theorem are given and discussed in section B.2 in Appendix B, and a proof is given in section B.3 there.

Theorem 1. *Assume* (B.6) *and* (B.7) *in section B.2 in Appendix B; base the classifier on the sign of* $T(X, \mathcal{I} \,|\, j_0)$*; and take* ν_0*, in the definitions of* $\mathcal{I}'$ *and* $\mathcal{I}''$ *at* (2.8) *and* (2.9)*, to diverge and to be no larger than a sufficiently small constant multiple of* n*. Then there exists a sequence of integers* $j(n)$*, diverging to infinity, such that:*

(I) For $k = 0, 1$*, if* $(X, \mathcal{I})$ *is drawn from* Π_k*, if* $j_0 \le j(n)$ *diverges as* $n \to \infty$*, and if the limit of either* $-Q_1(k, 1-k \,|\, j_0)$ *or* $Q_2(k, 1-k \,|\, j_0)$ *equals* $+\infty$*, then the probability that* X *is correctly classified converges to 1 as* $n \to \infty$*.*

(II) More generally, if X *is drawn from* Π_k*, the probability of correct classification equals*

$$P\Big\{Q_1(k, 1-k \,|\, j_0) - Q_2(k, 1-k \,|\, j_0) < 0 \;\Big|\; X, \mathcal{I}\Big\} + o(1)\,, \tag{5.3}$$

uniformly in $1 \le j_0 \le j(n)$*, as* $n \to \infty$*.*

Taking expectations in (5.3) for $k = 0, 1$, multiplying by $\pi_0 = \pi$ and $\pi_1 = 1 - \pi$, respectively, and adding, we deduce that the error rate $\mathrm{err}(j_0) \equiv P\{(X, \mathcal{I}) \text{ misclassified}\}$, of the classifier based on $T(X, \mathcal{I} \,|\, j_0)$, defined at (5.1), satisfies

$$\mathrm{err}(j_0) = \sum_{k=0}^{1} \pi_k\, P\big\{Q_1(k, 1-k \,|\, j_0) - Q_2(k, 1-k \,|\, j_0) \ge 0\big\} + o(1)\,,$$

uniformly in $1 \le j_0 \le j(n)$, as $n \to \infty$. Moreover, the error of the classifier converges to zero if $j_0 \to \infty$ and $P\{Q_1(0, 1 \,|\, \infty) - Q_2(0, 1 \,|\, \infty) = -\infty\} = P\{Q_2(1, 0 \,|\, \infty) - Q_1(1, 0 \,|\, \infty) = +\infty\} = 1$.

5.2 Theory for function extension

In the next two paragraphs we outline one of the settings where the random function extension of section 3.2 can be particularly effective for classification. Then, we provide theory for this case; see Theorem 2 below. Other examples can also be developed, for instance the one introduced in section 3.2. In Appendix B.1, we describe models where the nonrandom version of our extension procedure performs well.

Let $|\mathcal{I}|$ be the length of $\mathcal{I}$ and let $X_k = \mu_k + \Delta_k$ denote a random function drawn from Π_k, where $\mu_k = E_k(X_k)$ and the random process Δ_k has zero mean. One of the settings where random extension helps improve classification performance is when Δ_k is weakly dependent. In that setting and under mild conditions, for regular functions ψ, the variance of quantities such as $N \equiv \int_{\mathcal{I}} \Delta_k \, \psi$ increases at a strictly slower rate than $|\mathcal{I}|^2$, and often approximately in proportion to $|\mathcal{I}|$, as $|\mathcal{I}|$ increases. Therefore the "noise," N, typically satisfies $N = o_p(|\mathcal{I}|)$. On the other hand, if $|\mu_1 - \mu_2|$ remains reasonably large across most of $\mathcal{I}$, then the "signal," $s \equiv \int_{\mathcal{I}} (\mu_1 - \mu_2)^2$, increases like $|\mathcal{I}|$ rather than $o(|\mathcal{I}|)$ as $\mathcal{I}$ grows. Both N and s arise in expansions of classifiers such as LD and QD, and hence, using those expansions, we can conclude that noise has a proportionately smaller impact on the classifier, relative to the signal, if $\mathcal{I}$ is larger.

When classifying a new data function supported on $\mathcal{I}$, if we do not have access to a method such as random function extension, we can be forced to restrict attention to a significantly smaller interval than $\mathcal{I}$ where a number of fragments are available. (If n is small, this interval can even be empty and classification is not possible without extending curves.) For the reasons given in the previous paragraph, that restriction can lead to reduced performance. Put simply, classification on larger intervals results in lower levels of classification error; see the discussion following Theorem 2, below.

Next we develop theory demonstrating this property. Showing this in the most general setting would require long and complex arguments. To make our point firmly, and keep our arguments transparent, we make four simplifications, which will be captured in technical assumptions (5.4)–(5.6):

- We assume that the populations differ only in terms of means, and not in terms of covariance, or of endpoint distribution. The influence of endpoints will be studied separately in section 5.3.
- We assume that μ_1 is a simple translation of μ_0; say, $\mu_0 - \mu_1 \equiv d$, where d is a fixed, positive constant.
- We use a simplified version of the classifiers in section 4.1, namely the centroid-based classifier.
- When estimating the means μ_k from the extended curves, we do not weight the extended curves according to the amount by which they were extended. Using weights improves the estimated means, and hence the classifier's performance, but handling theory in that case requires longer arguments.

Similar results can be derived for the more sophisticated classifiers of section 4.1 which make explicit use of covariance, with or without bagging, but at the expense of longer and more complex arguments, and with stronger assumptions than those imposed here, including the weak dependence discussed earlier in this section. Nevertheless, the results derived below should help the reader understand why random function extension improves a classifier's performance.

Let X_k have the distribution of a function drawn from Π_k, before its support is censored. (Therefore the support of X_k equals $\mathcal{I}_0$.) Following the discussion in the previous paragraph, we assume that:

$$\text{the distributions of } X_k - E(X_k), \text{ for } k = 0 \text{ and } 1, \text{ do not depend on } k, \text{ and in particular are identical to the distribution of a process } Z; \tag{5.4}$$

$$\text{the process } Z \text{ satisfies} \quad E\Big\{ \sup_{t \in \mathcal{I}_0} Z(t)^2 \Big\} < \infty \,. \tag{5.5}$$

Next we stipulate the censoring mechanism that produces the intervals $\mathcal{I}_{kj} = [A_{kj}, B_{kj}]$. Define $\mathcal{I}_0 = [0, 1]$, let $\epsilon > 0$, let $\mathcal{J}_\epsilon$ be a diagonal strip down $\mathcal{I}_0 \times \mathcal{I}_0$:

$$\mathcal{J}_\epsilon = (\mathcal{I}_0 \times \mathcal{I}_0) \cap \bigcup_{-\infty < u < \infty} \Big(\{u\} \times [u - \epsilon, u + \epsilon] \Big) \,,$$

and, following the discussion in the previous paragraph again, assume that:

$$(A_{kj}, B_{kj}) \text{ is statistically independent of } X_{kj} \text{ and has a continuous distribution on } [0,1] \times [0,1], \text{ with a continuous density bounded away from 0 on } \mathcal{J}_\epsilon \text{ for some } \epsilon > 0, \text{ and not depending on } j \text{ or } k. \tag{5.6}$$

Given $a, b \in (0,1)$, with $a < b$, let $\mathcal{I} = [a,b]$. Condition (5.6) implies that, with probability converging to 1 as n_0 and n_1 diverge, the support of each interval $\mathcal{I}_{kj}$ can be extended from $\mathcal{I}_{kj}$ to $\mathcal{I}$ using random function extension.

Let X be a function drawn from Π_0 or Π_1, initially supported on $\mathcal{I}_0$ and then censored to $\mathcal{I}$. We wish to classify this curve as coming from one of the two populations. Without loss of generality, X was from Π_0. Then, $\Delta \equiv X - E(X) = X - \mu_0$ and has the distribution of Z, introduced above. Following the discussion in the before last paragraph, we consider the classifier that assigns X to Π_0 if $D(X) > 0$, and to Π_1 otherwise, where

$$D(X) = \int_{\mathcal{I}} (X - \bar{X}_1^*)^2 - \int_{\mathcal{I}} (X - \bar{X}_0^*)^2 \tag{5.7}$$

and, $\bar{X}_0^*$ and $\bar{X}_1^*$ are the respective means of the training datasets when, where possible, the functions X_{kj} are extended so that they are supported on intervals at least as large as $\mathcal{I}$.

In the setting delineated above, Theorem 2 describes the probability of committing an error when classifying X. Its proof is given in section B.4 in Appendix B.

Theorem 2. *Assume that* (5.4)–(5.6) *hold. Then, if either the censored training data are extended to* $\mathcal{I}$ *using random function extension, or if there is no censoring and the training data are all observed on* $\mathcal{I}_0$*, the following property holds:*

$$P\{D(X) \le 0\} \to P\left(d < -\frac{2}{|\mathcal{I}|} \int_{\mathcal{I}} \Delta \right) \tag{5.8}$$

as $n_0, n_1 \to \infty$.

Theorem 2 tells us two things. First, random function extension, as a way of remedying problems caused by the censoring of support intervals, achieves the same level of asymptotic classification error that would be attained if there were no censorship. Second, the theorem implies that a smaller level of error is attained by random function extension than would be achieved if we did not use that methodology and instead worked with a smaller interval, contained in $\mathcal{I}$. To appreciate the latter point, note

that since $d > 0$ then the minimum value of the right-hand side of (5.8) is zero, and that that minimum typically equals the limit of the right-hand side as $|\mathcal{I}|$ increases. (For example, this holds if Z is a stationary Gaussian process satisfying $\text{cov}\{Z(0), Z(t)\} \to 0$ as $|t| \to \infty$, $|\mathcal{I}|^{-1} \int_{\mathcal{I}} \Delta \to 0$ in probability as $|\mathcal{I}| \to \infty$.) In particular, larger intervals $\mathcal{I}$ result in lower levels of classification error.

5.3 Theory for classification using endpoints

In this section we show theoretically that specifically adjoining information about interval endpoints can lead to superior performance, relative to allowing endpoint information to "speak for itself" when it is incorporated implicitly along with the censored data function that is to be classified. Now, this comparison between two methods is complicated by the need to ensure that it is not confounded by the manner in which censored training data are extended to $\mathcal{I}_0$, and to avoid that difficulty we shall assume that the function extension step is undertaken perfectly; as we noted in section 2, there are several competing approaches to implementing it. Therefore, although each function fragment X_{kj} in the training sample was observed only on an interval $\mathcal{I}_{kj} \subseteq \mathcal{I}_0$, we shall suppose that the fragments were restored to at least $\mathcal{I}$.

To further simplify our analysis we assume that:

(a) prior to their support interval being censored, the random functions in Π_k, for $k = 0$ and 1, were Gaussian processes supported on $\mathcal{I}_0$ and with respective covariance functions $\gamma_k(t_1, t_2) = \text{cov}_k\{X(t_1), X(t_2)\}$ (with $\gamma_k(t, t)$ bounded away from zero and infinity on $\mathcal{I}$) and means $\mu_k(t) = E_k\{X(t)\}$; and (b) if the pair $(X, \mathcal{I})$ is drawn from Π_k, then the joint distribution of the endpoints (A, B) of the interval $\mathcal{I} = [A, B]$ has density f_k, independently of X, where $k = 0$ or 1. (5.9)

Assumption (5.9)(a) is imposed only because it is a convenient way of excluding pathological cases, for example where functions drawn from the two populations might take different, constant values on the interval $\mathcal{I}_0$; this would complicate our discussion below. The assumption of independence in (5.9)(b) is also made only for convenience, and, like (5.9)(a), it could be relaxed.

For additional simplicity we suppose that classification is based on the statistic $D(X)$ at (5.7). Define

$$T_k(X) = \int_{\mathcal{I}} (X - \mu_k)\,(\mu_0 - \mu_1) \tag{5.10}$$

and $d = \int_{\mathcal{I}} (\mu_0 - \mu_1)^2/2$, and let P_k denote probability measure under the assumption that X came from Π_k.

Theorem 3. *If* (5.9) *holds, if the training data are reconstructed perfectly on* $\mathcal{I}$ *as discussed three paragraphs above, and if classification is based on* $D(X)$ *defined at* (5.7)*, then the probability of misclassifying* X*, drawn randomly from* Π_0 *and* Π_1 *with respective prior probabilities* π_0 *and* π_1*, but censored to an interval* $\mathcal{I}$ *as indicated in* (5.9)*, converges to*

$$\pi_0\, P_0\{T_0(X) < -d\} + \pi_1\, P_1\{T_1(X) > d\} \tag{5.11}$$

as the training sample sizes diverge.

Theorem 3 describes the error rate of a classifier that is based directly on the pair $(X, \mathcal{I})$ and which uses endpoint information only implicitly. How does this level of error compare with that which would be obtained if we were to use endpoint information explicitly? It is clear that in some instances, for example where the densities f_0 and f_1 in (5.9)(b) are identical, we cannot improve on the error rate at (5.11) by adjoining endpoint information. Of course, this is not a problem; as noted at the end of section 4.2, by using cross-validation to compare the performances of different classifiers, we can choose among them without risking more than a minor loss of performance. Moreover, there are occasions when much is to be gained by using endpoint information explicitly, as we demonstrate below. In those cases, too, when constructing the combined classifier, cross-validation assesses consistently the advantages and disadvantages of using endpoints explicitly, relative to using them only implicitly.

To indicate the attractiveness of treating endpoint information explicitly we consider a simple endpoint classifier, S_{int}, based for example on Fisher's linear or quadratic

discriminant. For these two classifier types, and several others, as the distributions with densities f_0 and f_1 become increasingly concentrated around the pairs (a_0, b_0) and (a_1, b_1), respectively (where $a_k < b_k$ for $k = 0, 1$, and $(a_0, b_0) \neq (a_1, b_1)$), the error rate of the interval-based classifier decreases to 0. (In the case of Fisher's linear or quadratic discriminant, this is a simple corollary of the fact that the variances of the respective distributions of the endpoint pair (A, B) both decrease to zero.) That is, as the two distributions of (A, B), in the respective cases of Π_0 and Π_1, become increasingly more concentrated around the distinct pairs (a_0, b_0) and (a_1, b_1), respectively, the limit as $n_0, n_1 \to \infty$ of the probability that the endpoint classifier based on $S_{\rm int}$ commits an error, converges to 0. At the same time, the error rate at (5.11) remains bounded away from zero. (To appreciate why, note that taking $\mathcal{I} = [a_0, b_0]$ or $\mathcal{I} = [a_1, b_1]$ in the definition of T_k at (5.10), for either choice of k, does not reduce the error to zero.) It is straightforward to show from these properties that, since we use consistent quantile estimators $\hat{q}_{S,\alpha_k}$ in formula (4.1), the limit (as training sample sizes increase) of the probability that the classifier based on $T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1)$, at (4.1), is strictly less than that of the classifier based on $T_{\rm fun}(X, \mathcal{I} \,|\, j_0, w_1)$ alone, if the densities f_0 and f_1 are sufficiently concentrated.

The change between cases where the explicit use of interval endpoints has no asymptotic effect, and instances where it has a marked asymptotic effect, occurs smoothly, and in particular there is a wide variety of settings where the explicit use of endpoints has a noticeable but not extreme effect. In each setting the strength of the case for using endpoints explicitly can be assessed using cross-validation, and so the asymptotic error, when endpoints are taken into account, will not be less than that for a conventional approach that does not make explicit use of endpoints. It should also be noted that we have considered here only one instance, or reason—sufficiently high concentration of endpoint distributions about specific disjoint endpoints—for it to be attractive to include endpoint data explicitly in the classifier. There are other occasions too where using explicit endpoint data is advantageous, because it reduces

error rate.

6 Choice of parameters

6.1 Methods in sections 2.3 and 3.3

We select j_0 by cross-validation (CV) to minimise the error rate of the classifier, as follows. Let J^* (respectively, $J^*_{kj_1}$) be the minimum of the number of nonzero eigenvalues of $\hat{K}_0$ and $\hat{K}_1$ on the interval $\mathcal{I}$ (respectively, of $\hat{K}_{0,-j_1}$ and $\hat{K}_{1,-j_1}$ on $\mathcal{I}_{kj_1}$, where $\hat{K}_{k,-j_1}$ are leave-one-out estimators; see section C.3 in Appendix C). For each j_0, let n_{k,j_0} denote the number of curves X_{kj_1} for which $J^*_{kj_1} \geq j_0$. Also, let $J_{\max}$ be the largest $j_0 \leq J^*$ such that $n_{0,j_0} + n_{1,j_0} \geq n/2$. For the basic method in section 2.3 we choose j_0 between 1 and $J_{\max}$ to minimise

$$\widehat{\mathrm{err}}_{\mathrm{fun}}(j_0) = \sum_{k=0}^{1} \frac{\pi_k}{n_{k,j_0}} \sum_{j_1=1}^{n_k} I\big\{(-1)^k\, T_{\mathrm{fun},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, j_0, w_1) > 0\big\}\; I\big(J^*_{kj_1} \geq j_0\big)\,, \tag{6.1}$$

where $\pi_0 = \pi$, $\pi_1 = 1 - \pi$, and $T_{\mathrm{fun},-j_1}(X, \mathcal{I} \,|\, j_0, w_1)$ is either the QD at (2.7) or its LD analogue, constructed without using the j_1th observation. See section C.3 in Appendix C for precise definitions. As in Delaigle and Hall (2012), in case of multiple local minima we search for the global minimum among the first two local minima.

For the QD method in section 3.3 we suggest taking h_k to be the rth empirical quantile of $\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\mathrm{int}}/(b-a)$, for $\ell = 1, \ldots, n_k$, where $0 \leq r \leq 1$ (the same for each k) is chosen by CV, together with j_0, by minimising $\widehat{\mathrm{err}}_{\mathrm{fun}}$ in (6.1) with respect to j_0 and r simultaneously (of course, in this case $\widehat{\mathrm{err}}_{\mathrm{fun}}$ depends on r through $\bar{X}_k$ and $\widehat{K}_k$). For each r we resolve the issue of multiple local minima with respect to j_0 as above. For the LD method (see section C.2 in Appendix C), we take h to be the rth empirical quantile of $\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\mathrm{int}}/(b-a)$, for $\ell = 1, \ldots, n_k$ and $k = 0, 1$.

As indicated in section 2.2, we take $w = -2\,\log\{\pi/(1-\pi)\}$, where π is the prior probability of Π_0. In principle, we could choose w by CV, as we shall do in section 6.2 for the simpler method based on endpoints. However, the curve classifiers (especially

the version in section 3.3) already involve other parameters, so we prefer to fix w.

6.2 Method in section 4.1

If the distribution of (A, B) were normal then the best classification performance would be obtained by taking $w = -2\log\{\pi/(1-\pi)\}$, since the classifier then would correspond then to Bayes' rule. In more general cases, better classification could be obtained by choosing w by CV to minimise the error rate of the classifier. Therefore, we suggest selecting w to minimise

$$\widehat{\mathrm{err}}_{\mathrm{int}}(w) = \frac{\pi}{n_0} \sum_{j_1=1}^{n_0} I\bigl\{S_{\mathrm{int},-j_1}(\mathcal{I}_{0j_1} \,|\, w) > 0\bigr\} + \frac{1-\pi}{n_1} \sum_{j_1=1}^{n_1} I\bigl\{S_{\mathrm{int},-j_1}(\mathcal{I}_{1j_1} \,|\, w) \le 0\bigr\}\,,$$

where $S_{\mathrm{int},-j_1}(\,\cdot\,|\,w)$ denotes the leave-one-out version of $S_{\mathrm{int}}(\,\cdot\,|\,w)$ in section 4.1; see section C.3 in Appendix C for precise definitions.

6.3 Method in section 4.2

To choose j_0, w_1, w_2, α_0 and α_1 we could use CV, minimising the error rate of the classifier based on $T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1)$. While this would work if n were sufficiently large, for n smaller this approach can run the risk of focusing too much on the training sample, and its generalisation to more than two populations would be computationally very intensive. Therefore we use a sequential procedure, as follows.

Since our main classifier is based on T_{fun}, defined in section 2.2, but where the means and covariances are calculated as in section 2.3 or 3.3, we choose j_0 and w_1 (as well as h_0 and h_1 for the method in section 3.3) as if we were using the classifier T_{fun} alone, that is, as described in section 6.1.

In view of the way in which we use the weight w_2 in (4.1) to refine the classifier based on T_{fun}, rather than choose w_2 to optimise performance of the classifier based on S_{int}, we suggest selecting w_2 so that $|S_{\mathrm{int}}|$ is large when the classifier based on S_{int} makes a correct decision, and small otherwise. To implement this in practice, we

choose w_2 to maximise the following CV criterion:

$$CV_{\rm int}(w_2) = {\rm Dist_c}(w_2)\big/{\rm Dist_w}(w_2)\,, \tag{6.2}$$

where the subscript c stands for "correct," w stands for "wrong," and, letting $S_{{\rm int},-j_1}$ denote the leave-one-out version of $S_{\rm int}$ in section 4.1 (see section C.3 in Appendix C),

$${\rm Dist_c}(w_2) = \sum_{k=0}^{1} \frac{\pi_k}{n_k} \sum_{j_1=1}^{n_k} |S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2)|\; I\big\{(-1)^k\, S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2) < 0\big\}\,,$$

$${\rm Dist_w}(w_2) = \sum_{k=0}^{1} \frac{\pi_k}{n_k} \sum_{j_1=1}^{n_k} |S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2)|\; I\big\{(-1)^k\, S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2) \ge 0\big\}\,.$$

Finally, with j_0, w_1 and w_2 fixed, we choose α_0 and α_1 by CV to minimise the error rate of the classifier $T(X, \mathcal{I} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1)$. That is, we choose α_0 and α_1 to minimise $\widehat{\rm err}(\alpha_0, \alpha_1)$, where $\widehat{\rm err}(\alpha_0, \alpha_1)$ is equal to

$$\sum_{k=0}^{1} \frac{\pi_k}{n_{k,j_0}} \sum_{j_1=1}^{n_k} I\big\{(-1)^k\, T_{-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, j_0, w_1, w_2, \alpha_0, \alpha_1) > 0\big\}\; I\big(J^*_{kj_1} \ge j_0\big)\,, \tag{6.3}$$

with n_{k,j_0} and $J^*_{kj_1}$ as in (6.1), and T_{-j_1} defined as was T in (4.1), but with $T_{\rm fun}$ and $S_{\rm int}$ replaced by $T_{{\rm fun},-j_1}$ and $S_{{\rm int},-j_1}$.

6.4 Other details of implementation

To calculate the quantiles $\hat{q}_{S,\alpha_k}$ in section 4.2, first, we compute the n statistics $S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2)$, for $j_1 = 1, \ldots, n_k$ and $k = 0, 1$. For $k = 0, 1$, we take $\hat{q}_{S,\alpha_k}$ to be the α_k-level empirical quantile of $S_{\rm int}$ calculated from the sample of values for which $(-1)^{k+1} S_{{\rm int},-j_1}(X_{kj_1}, \mathcal{I}_{kj_1} \,|\, w_2)$ is positive. We use leave-one-out versions in the CV criterion in (6.3).

Although the value of ν_0 in section 2.3 can influence performance, it is not a crucial parameter. It is only a lower bound to the number of curves used to calculate means and covariances, and is mostly useful for n small (as n increases, the number of curves observed on most parts of $\mathcal{I}$ usually increases); and we prefer not to choose it by CV. In our numerical work we took $\nu_0 = 3$, because we had a low number of curves and short fragments.

7 Numerical properties

7.1 Simulated examples

We simulated data $X_{ik}(T_{j,ik})$, for $k = 0, 1$, $i = 1, \ldots, n_k$ and $j = 1, \ldots, N_{ik}$. In each case, $X_{ik}(T_{j,ik})$ was generated from three models:

$$X_{ik}(T_{j,ik}) = m_k(T_{j,ik}) + Z_{ik}\, f(T_{j,ik}, Z_{ik}) + \epsilon_{ikj}\,, \tag{7.1}$$

$$X_{ik}(T_{j,ik}) = m_k(T_{j,ik} - S_{ik}) + \{U_{ik} + V_{ik}\, \sin(T_{j,ik}/W_{ik})\} \times \{Z_{ik} + \sin(T_{j,ik}\, 10^{-3}\pi)\} + \epsilon_{ikj}\,, \tag{7.2}$$

$$X_{ik}(T_{j,ik}) = m_k(T_{j,ik}) + U_{ik} + V_{ik}\, \sin(T_{j,ik}/W_{ik} + Z_{ik})\,, \tag{7.3}$$

where $f(t, u) = 0.02\,\{(3t + 100)\,(u + 1)\}^{1/2}$ and with the m_ks taken to be one of:

$$m_k(t) = \sin(t/c_k)/\{(0.1\,t - d_k)^2 + 1\},\ \text{with } c_0 = 15,\ c_1 = 12,\ d_0 = 5,\ d_1 = 4, \tag{7.4}$$

$$m_k(t) = \text{logit}^{-1}\{(t - c_k)/d_k\},\ \text{with } (c_0, d_0) = (50, 20) \text{ and } (c_1, d_1) = (40, 12) \tag{7.5}$$

or $(c_1, d_1) = (50, 5)$, where $\text{logit}^{-1}(x) = e^x/(1 + e^x)$. In model (7.1), the groups means μ_k are equal to m_k, and in models (7.2) and (7.3), $\mu_k = m_k + \delta_k$ for some function δ_k that can easily be obtained by taking the expectation of $X_{ik}(t)$, but whose expression is cumbersome, whence our implicit definition of μ_k. Except otherwise stated, we took the variables N_{ik}, S_{ik}, $T_{j,ik}$, U_{ik}, V_{ik}, W_{ik}, Z_{ik} and ϵ_{ikj} (or the subset of them appearing in each model) to be totally independent, and we generated them according to different settings.

In each setting, we generated $B = 200$ pairs of test and training samples, and applied, to the test sample, the classifiers constructed from the training sample. In the training sample, for several values of n_k, we generated n_k curves from group k, for $k = 0, 1$. As before we use the notation $n = n_0 + n_1$. In the test sample, we generated 100 curves which came with equal probability from Π_0 or from Π_1. Tables 1 and 2 report the percentage of test curves that were correctly classified, averaged over the B test samples, for various classifiers.

Table 1: Percentage of correctly classified observations for the simulated data of section 7.1.1, in settings (1) to (8), using the methods in §3.3 or §4.2, or the procedure of James and Hastie (2001) (JH).

	(1)			(2)			(3)			(4)		
n	§3.3	§4.2	JH	§3.3	§4.2	JH	§3.3	§4.2	JH	§3.3	§4.2	JH
70	80.3	79.3	72.3	84.9	89.7	73.5	61.0	60.1	61.7	67.0	87.0	63.1
100	83.5	82.8	72.9	88.1	92.4	73.2	63.2	62.5	62.8	69.4	88.6	63.5
	(5)			(6)			(7)			(8)		
n	§3.3	§4.2	JH	§3.3	§4.2	JH	§3.3	§4.2	JH	§3.3	§4.2	JH
70	66.5	65.6	60.5	70.8	69.7	72.1	78.2	87.9	68.4	70.2	68.8	61.2
100	69.0	67.9	62.0	74.4	73.9	72.0	81.2	90.2	69.9	73.7	73.0	61.5

7.1.1 Comparison of methods for short fragments

First we consider the most challenging case, where the fragments are short, as in section 3. There, the only available methods are those based on curve extension with (section 4.2) or without (section 3.3) the use of endpoints, and the procedure of James and Hastie (2001). Since the fragments are short and not very numerous, we use the LD version of the curve-based classifier. We generated short fragments from eight settings, denoted by (1) to (8) and described in section A.2 in Appendix A. These examples were to chosen to illustrate several aspects of the classification problem:

- In settings (1), (3), (5), (6) and (8), the distributions of the endpoints of the intervals do not differ among the groups, whereas in settings (2), (4) and (7), the endpoints contain valuable information for classification.
- The curves m_k in (7.4) (settings (1)–(5)) have pronounced features, whereas those in (7.5) (settings (6)–(8)) are monotone, as in the growth example.
- In model (7.1) (setting (1), (2), (6) and (7)), it is easily seen that the shape of fragments depends heavily on their vertical location, hence we can apply the nonrandom extension algorithm. This is not the case for models (7.2) and (7.3) (settings (3)–(5) and (8)), where we apply the random extension algorithm.

For the training sample sizes, we took $(n_0, n_1) = (30, 40)$ or $(45, 55)$ in settings (1) to (5), and $(n_0, n_1) = (35, 35)$ or $(50, 50)$ in settings (6) to (8). Table 1 reports the average percentage of correctly classified test curves, for the curve classifier of section 3.3 based on extended curves, for the combined classifier of section 4.2 which also uses endpoint information, and for the method of James and Hastie (2001).

For the latter, we need to choose three tuning parameters: the number of knots

q, and two parameters (p and h) used for reducing dimension. Choosing all three by CV appears to be too time consuming. Therefore, we took $h = 1$ as in James and Hastie (2001), and $p = q - 1$. This choice of p resulted from experimentation with several examples; we cannot guarantee that it is always the best choice, but it seems to be a reasonable compromise between numerical problems and good performance. In settings (1)–(5) we chose q by minimising a CV estimate of classification error. In settings (6)–(8), numerical problems prevented us from choosing q by CV, and we chose the value ($q = 3$ or 4) that gave the best results over the B simulations.

For the methods of sections 3.3 and 4.2, we chose the tuning parameters as in section 6. When we used the random extension procedure of section 3.2, we also used the bagging approach described at the end of section 3.3, with $N_B = 25$. That is, we randomly extended the training curves 25 times, applied the full classification procedure for these 25 versions, and assigned the new fragment to the population chosen the most frequently out of the 25 replicates.

The results are reported in Table 1. In most cases, our approach outperformed that of James and Hastie (2001), although the latter worked reasonably well. We can see that including endpoint information can be very valuable for classification, sometimes improving performance by 20%, without degrading it much otherwise.

7.1.2 Comparison of methods for long fragments

In the simpler case of long fragments, we can apply several classifiers: the basic procedure of section 2.3 with (see section 4.2) or without incorporating endpoint information, and the method of James and Hastie (2001). Recall that our main methodological contribution in this setting is the idea of incorporating endpoint information, and our main goal in this section is to illustrate the improvements it can bring. Comparison with James and Hastie's (2001) approach is given for illustration. It is also of interest here to compare performance of the classifiers when the means are estimated using the empirical estimator or by the smooth version of Yao et al. (2005).

Table 2: Percentage of correctly classified observations for the simulated data of section 7.1.2, in settings (1) to (4), using the method of §2.3 with empirical means (A) or Yao et al.'s (2005) means (B), the method of §4.2 with empirical means (C) or Yao et al.'s (2005) means (D) or the procedure of James and Hastie (2001) (E).

n	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
	(1)					(2)					(3)					(4)				
40	70.7	72.6	69.0	70.8	65.7	71.1	72.8	80.6	81.2	65.6	67.9	67.8	66.8	66.8	65.0	69.0	69.2	80.2	80.1	66.1
80	72.1	73.6	71.1	72.3	68.0	73.6	75.1	83.6	83.6	66.9	71.3	71.0	70.4	70.3	71.0	71.2	71.8	82.1	82.5	71.6

For brevity we focus on the LD classifier; for QD versions, see section 7.2.3.

We generated long fragments from four settings, denoted by (1) to (4) and described in section A.3 in Appendix A. These examples were chosen to illustrate two pairs of contrasting settings:

- curves with a lot of features (settings (3) and (4)) or simpler curves as in the pulmonary example (settings (1) and (2));
- endpoints which contain information for classification (settings (2) and (4)) or are not informative for classification purposes (settings (1) and (3)).

In each case the training samples were of size $n = n_0 + n_1$, where $n_0 = n_1 = 20$ or 40, and the 100 test samples came with equal probability from Π_0 or Π_1. Table 2 reports the mean percentages of correctly classified observations for the basic LD classifier in section 2.3 (using the empirical means or the smooth means of Yao et al., 2005), the combined classifier of section 4.2 (again using either version of the mean estimator), and the procedure of James and Hastie (2001). As in section 7.1.1, for the three tuning parameters required for James and Hastie's (2001) method, we took $h = 1$ and $p = q - 1$, where, in settings (3) and (4), q was chosen by CV, and in settings (1) and (2) q was chosen as for settings (6)–(8) in section 7.1.1.

The results indicate that smoothing is not crucial when computing the mean of the basic classifier: it can either improve or worsen classification a little. In most cases, the basic classifier performed a little better than the method of James and Hastie (2001). A significant advantage of the former is that it is fully data-driven and fast to compute. The combined classifier of section 4.2 boosted performance significantly when the endpoints contained valuable information, without degrading

Table 3: Percentage of correctly classified observations for the Wheat data using the method in section 2.3.

n	Full curves		Case (a)		Case (b)		Case (c)		Case (d)	
	LD	QD	LD	QD	LD	QD	LD	QD	LD	QD
30	95.7%	94.9%	95.6%	93.9%	92.4%	91.7%	91.3%	90.8%	94.2%	90.2%
50	99.5%	97.9%	97.9%	97.4%	95%	94.9%	93.9%	94.4%	96.8%	96.1%
80	99.9%	98.7%	98.9%	98.2%	97.5%	97.7%	97%	98.3%	97%	97.5%

it much otherwise.

7.2 Real data examples

7.2.1 Illustration of the perfect classification property

We illustrate the near perfect classification property discussed in section 5.1, through the wheat data described by Kalivas (1997). The data X_i, for $i = 1, \ldots, 100$, consist of the first derivatives of near infrared spectra of 100 wheat curves, measured at 700 equispaced wavelengths, denoted by 1 through 700. As in Delaigle and Hall (2012), we put in Π_0 the 41 observations whose moisture level is more than 15, and put the other 59 curves in Π_1. Delaigle and Hall (2012) showed that, in this example, the LD classifier applied to the full (non fragmented) curves, reaches near perfect classification.

To examine the impact that observing only fragments of curves has on the classifier, we kept only a fragment $[A_i, B_i]$ of the ith curve, for $i = 1, \ldots 100$, where we generated the A_is randomly, and took $B_i = \min(A_i + \delta_i, 700)$ with δ_i chosen randomly. We did this in four ways, denoted by (a) through (d) and described in section A.1 in Appendix A. Examples of fragments from cases (c) and (d) are shown in Figures D.1 and D.2 in Appendix D.

In all cases we created such fragments $B = 100$ times and divided the data into a training sample of size $n = 30$, 50 or 80, and a test sample of size $100 - n$. For each pair of samples created this way, we calculated the LD and QD classifiers of section 2.3 from the training sample, which we applied to the curves in the test sample. Table 3 reports the mean percentage (averaged over the B test samples) of correctly classified test curves. We also show the near perfect results obtained with the LD and

QD applied to the full curves. It can be seen that, from cases (a) to (d), the intervals become more scattered, but classification performance remains excellent. For small n, LD slightly outperforms QD, but as n increases, the opposite tends to be true, especially in more complex settings.

7.2.2 Short fragments: the growth data

The growth dataset described by Bachrach et al. (1999) consists of measurements of growth through spinal bone mineral density, for individuals from four ethnic groups (referred to as Asians, Blacks, Hispanics and Caucasians), taken at ages ranging from 8 to 25 years. For each individual, only 2 to 4 measurements were taken, and only over a period of a few years. Growth curves are usually smooth and monotone, and since the measurements per individual are taken close in time, the linearly joined observations give good approximations to fragments of curves. We base our classification procedure on the X_is, where X_i is the approximate fragment for the ith individual.

We use the same subset of $n = 153$ female individuals as James and Hastie (2001), comprised of $n_1 = 35$ Asians, $n_2 = 43$ Blacks, $n_3 = 27$ Hispanics and $n_4 = 48$ Caucasians. In each group, these fragments of curves do not overlap much (see Figure D.3 in Appendix D). Therefore we use the method of section 3. Moreover, some of the group sizes are quite small, and therefore we use the LD version of the classifier.

The classifiers introduced in the previous sections for the case of two groups can be extended to the case of $G > 2$ groups ($G = 4$ for these data) in the standard way. Specifically, we calculate the group means for each group, extend the definition of the discriminant T_k in (2.7) to $k = 1, \ldots, G$, and classify a new observation as coming from the group having the smallest value of T_k. We choose the bandwidths h and $h_1, \ldots, h_G$ as in section 6.1, which amounts to choosing only one parameter r.

As in James and Hastie (2001) we applied the classifier to each of the 153 observations, taking each time the training sample to consist of the remaining 152 observations. We took each prior probability $\pi_1, \ldots, \pi_4$ equal to $1/4$. Proceeding in this way,

our procedure in section 3.3 classified 67 individuals correctly. Our combined classifier, based on the curves and the endpoints, increased this number to 73. Indeed, for these data, useful classification information appears to be available in the endpoints, since the number of correct classifications, using LD based on the endpoints alone is 59, thus significantly higher than would be expected from random guessing. Using James and Hastie's (2001) method with $p = 2$ and $h = 1$, as recommended there; and with q chosen by CV; we correctly classified 61 observations.

7.2.3 Long fragments: the forced expiratory volume data

In a pollution study reported by Dockery et al. (1983), 13 379 children from six US cities were examined for several years. We use the subset of 300 girls described by Fitzmaurice et al. (2004). For each girl the data consist of yearly measurements of age, height and log(FEV1), where FEV1 is the forced expiratory volume in one second, a measure of pulmonary function. Not all girls started in and dropped from the study at the same age, and for each girl we have access to only a fragment of the curve $X_i(t)$, where t denotes age and $X_i(t)$ denotes log(FEV1) at age t. Let $\mathcal{I}_i$ be the age interval during which the ith girl was observed.

We kept only the 252 girls who were examined at least twice, and divided the data into two groups of equal size. We did this in two ways called settings (1) and (2), and described in section A.4 in Appendix A. To assess performance in both cases, as in section 7.2.1 we randomly split the data $B = 200$ times into a training sample of size $n = 50$, 100 or 200, and a test sample of size $252 - n$, and calculated the percentage of correctly classified observations as in section 7.2.1. As can be seen from Figure D.4 in Appendix D, the observed fragments are long and overlap significantly in each group. Therefore we can use the basic classifier of section 2.3, where, as in the simulated examples, we can use the empirical mean or the smooth mean of Yao et al. (2005). We tried to use the method of James and Hastie (2001), but were unable to make it work with these data because of numerical problems.

Table 4: Percentage of correctly classified observations for the FEV data using LD or QD classifiers with empirical mean or with smooth means of Yao et al. (2005), the latter being indicated by a subscript sm.

		§2.3		§4.1	§4.2		§2.3		§4.2	
setting	n	LD	LD_{sm}		LD	LD_{sm}	QD	QD_{sm}	QD	QD_{sm}
(1)	50	64.3	65.2	56.7	67.1	67.8	62.1	63.1	64.7	65.4
	100	64.6	65.4	57.7	67.0	67.8	63.0	64.1	66.1	67.0
	200	65.7	65.8	58.2	68.7	68.7	65.7	66.7	68.8	69.7
(2)	50	62.7	62.3	69.6	69.1	70.0	59.5	58.9	67.6	67.8
	100	62.5	62.4	71	70.4	72.1	59.7	59.7	69.4	69.9
	200	64.6	63.7	72.8	73.8	75.8	59.3	59.5	71.8	72.5

The proportion of correctly classified observations is reported in Table 4; for the methods in sections 2.3 and 4.2 we show the results of the LD and QD versions, using both the empirical mean or the smooth mean of Yao et al. (2005). In setting (1), the endpoints are much less informative then the curves, and the combined classifier in section 4.2 does not improve on the classifier in section 2.3, but does not degrade it much either. In setting (2), the endpoints are much more informative than the curves, and the combined classifier improves significantly on the classifier in section 2.3. In general, LD outperformed QD, which is often the case in practice unless n is rather large or the covariances of the two groups differ significantly. In setting (1), using a smooth mean improved classification a little, but degraded it in setting (2).

To choose which of the methods in sections 2.3 to 4.2 to use, one can be guided by a comparison of the CV estimates of classification error for each method. Since the method in section 4.2 is more complex, in general one can use the method in section 2.3, unless the CV error for the classifier in section 4.2 is significantly smaller, in which case we use the latter. For example, in setting (1) the CV error of the classifier in section 4.2 was only 2% smaller than that of the classifier in section 2.3, whereas in setting (2), it was 10% smaller.

Appendices

Appendices are given in the supplementary file.

Acknowledgements

Research supported by grants and fellowships from the Australian Research Council. We thank the editor, the associate editor, and two referees for their valuable comments that helped improve a previous version of the manuscript.

References

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. *J. Clinical Endocrinology & Metabolism*, **84**, 4702–4712.

Biau, G., Bunea, F. and Wegkamp, M.H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory* **51**, 2163–2172.

Berlinet, A., Biau, G. and Rouvière, L. (2008) Functional classification with wavelets. *Annales de l'Institut de statistique de l'université de Paris*, **52**, 61–80.

Breiman, L. (1995). Bagging predictors. *Machine Learning* **24**, 123–140.

Chamroukhi, F., Same, A., Govaert, G. and Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing* **73**, 1210–1221.

Cuevas, A., Febrero, M. and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comput. Statist.* **22**, 481–496.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional. *J. Roy. Satist. Soc.* Ser. B **74**, 267–286.

Delaigle, A., Hall, P. and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99**, 299–313.

Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E. and Ferris, B.G. (1983). Distribution of FVC and FEV1 in children 6 to 11 years old. *American Review of Respiratory Disease*, **128**, 405–412.

Epifanio, I. (2008). Shape descriptors for classification of functional data. *Technometrics* **50**, 284–294.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Comput. Statist. Data Anal.* **4**, 161–173.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York.

Fitzmaurice, G., Laird, N. and Ware, J. (2004). Applied Longitudinal Analysis. Wiley Series in Probability and Statistics.

Fromont, M. and Tuleau, C. (2006). Functional classification with margin conditions. In *Learning Theory–Proceedings of the 19th Annual Conference on Learning Theory, Pittsburgh, PA, USA, June 22-25, 2006*, Eds J.G. Carbonell and J. Siekmann. Springer, New York.

Glendinning, R.H. and Herbert, R.A. (2003). Shape classification using smooth principal components. *Patt. Recognition Lett.* **24**, 2021–2030.

Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33**, 284–306.

Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.

Huang, D.-S. and Zheng, C.-H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22**, 1855–1862.

James, G. and Hastie, T. (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *J. Roy. Statist. Soc., Ser. B*, **63**, 533–550.

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

James, G. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98**, 397–408.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **37**, 255–259.

Leng, X. and Müller, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22**, 68–76.

López-Pintado, S. and Romo, J. (2006). Depth-based classification for functional data. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72**, 103–120.

Peng, J. and Müller, H-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Statist.* **2**, 1056–1077.

Preda, C., Saporta, G. and Leveder, C. (2007). PLS classification of functional data. *Comput. Statist.* **22**, 223–235.

Ramsay, J.O. and Silverman, B.W. (2005). Functional Data Analysis, second edn. Springer, New York.

Song, J.J., Deng, W., Lee, H.-J. and Kwon, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Comp. Biology and Chemistry* **32**, 426–432.

Wang, X.H., Ray, S. and Mallick, B.K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102**, 962–973.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577–590.

NOT FOR PUBLICATION APPENDICES FOR THE PAPER: CLASSIFICATION USING CENSORED FUNCTIONAL DATA

A More details on simulation settings

A.1 Settings for section 7.2.1

The four settings considered in section 7.2.1 are:

(a) for X_i in Π_0 (respectively, Π_1), $A_i \sim U[1, 50]$ (respectively, $A_i \sim U[1, 100]$), and for each i, $\delta_i \sim U[600, 700]$;

(b) for X_i in Π_0 (respectively, Π_1), $A_i \sim U[1, 150]$ (respectively, $A_i \sim U[100, 250]$), and $\delta_i \sim U[200, 300]$ (respectively, $\delta_i \sim U[150, 250]$);

(c) for each i, $A_i \sim U[100, 250]$ and $\delta_i \sim U[150, 250]$;

(d) for each i, $A_i \sim U[100, 450]$ and $\delta_i \sim U[150, 250]$.

A.2 Settings for section 7.1.1

The eight settings considered in section 7.1.1 are:

(1) Model (7.1) with m_k in (7.4), $Z_{ik} \sim U[-2/3, 2/3]$, $N_{ik} \sim \big[U[5, 10]\big]$, $T_{1,ik} \sim U[1, 95]$ and $\epsilon_{ikj} \sim N(0, 10^{-4})$. For $j > 1$, $T_{j,ik} = T_{j-1,ik} + U_{j,ik}$, where $U_{j,ik} \sim U[0.75, 1.25]$.

(2) Same as (1), except that $N_{i0} \sim \big[U[5, 10]\big]$, $N_{i1} \sim \big[U[8, 20]\big]$, $T_{1,i0} \sim 0.5\, f_{U[1,30]} + 0.5 f_{U[31,95]}$, and $T_{1,i1} \sim U[1, 80]$.

(3) Model (7.2) with m_k in (7.4), $S_{ik} \sim U[-5, 10]$, $U_{ik} \sim U[-1, 1]$, $V_{ik} \sim U[0.02, 0.05]$, $W_{ik} \sim U[2, 3]$, $Z_{ik} \sim U[0.1, 0.5]$, $\epsilon_{ikj}|T_{j,ik} \sim N\big(0, \{0.005 \log(T_{j,ik} + 2)\}^2\big)$, $N_{ik} \sim \big[U[5, 10]\big]$, $T_{1,ik} \sim U[1, 95]$, and for $j > 1$, $T_{j,ik} = T_{j-1,ik} + 1$.

(4) Same as (3) but with $N_{i0} \sim \big[U[5, 10]\big]$, $N_{i1} \sim \big[U[8, 20]\big]$, $T_{1,i0} \sim 0.5\, f_{U[1,30]} + 0.5 f_{U[31,95]}$, and $T_{1,i1} \sim U[1, 80]$.

(5) Model (7.3) with m_k in (7.4), $U_{ik} \sim U[-1, 1]$, $V_{ik} \sim U[0.025, 0.05]$, $W_{ik} \sim U[2, 3]$, $Z_{ik} \sim N(0, 0.04)$, $N_{ik} \sim \big[U[5, 10]\big]$, $T_{1,ik} \sim U[1, 95]$, and for $j > 1$, $T_{j,ik} = T_{j-1,ik} + 1$.

(6) Same as (1), but with m_k in (7.5), where $(c_1, d_1) = (40, 12)$.

(7) Same as (2), but with m_k in (7.5), where $(c_1, d_1) = (40, 12)$.

(8) Model (7.2) with m_k in (7.5), where $(c_1, d_1) = (50, 5)$, $S_{ik} \sim U[-5, 10]$, $U_{ik} \sim U[-0.75, 0.75]$, $V_{ik} \sim U[0.02, 0.05]$, $W_{ik} \sim U[2, 3]$, $Z_{ik} \sim U[0.1, 0.5]$, $\epsilon_{ikj}|T_{j,ik} \sim N\big(0, \{0.005 \log(T_{j,ik}/2 + 2)\}^2\big)$, $N_{ik} \sim \big[U[7, 10]\big]$, $T_{1,ik} \sim U[1, 95]$, and for $j > 1$, $T_{j,ik} = T_{j-1,ik} + 1$.

A.3 Settings for section 7.1.2

The four settings in section 7.1.2 are:
(1) Model (7.1), with μ_k as in (7.5), where $(c_1, d_1) = (50, 5)$ and $Z_{ik} \sim U[-2/3, 2/3]$, $N_{ik} \sim \big[U[5, 14]\big]$, $T_{1,ik} \sim U[1, 40]$ and $\epsilon_{ikj} \sim N(0, 0.15)$. For $j > 1$, $T_{j,ik} = T_{j-1,ik} + U_{j,ik}$, where $U_{j,ik} \sim U[5.5, 8.5]$.

(2) Same as (1), but with $N_{i0} \sim \big[U[5, 10]\big]$, $N_{i1} \sim \big[U[7, 14]\big]$, $T_{1,i0} \sim U[1, 50]$, and $T_{1,i1} \sim U[1, 30]$.

(3) Same as (1), but with μ_k as in (7.4), where $c_0 = 5$, $c_1 = 4.8$, $d_0 = 5$, $d_1 = 4.9$, and with $\epsilon_{ikj} \sim \text{Exp}(3.5) - 1/3.5$, where $\text{Exp}(\lambda)$ denotes the exponential distribution with mean $1/\lambda$.

(4) Same as (3), but with $N_{i0} \sim \big[U[5, 10]\big]$, $N_{i1} \sim \big[U[7, 14]\big]$, $T_{1,i0} \sim U[1, 50]$, and $T_{1,i1} \sim U[1, 30]$.

A.4 Settings for section 7.2.3

In setting (1) we created the groups according to height at the age value (10.7) that belongs to the largest number of the intervals $\mathcal{I}_i$. In particular, Π_0 was comprised of girls whose height at age 10.7 was less than 1.41, and Π_1 contained the other girls. In setting (2), we put into group Π_0 the girls whose height at the start of the study was less than 1.27, and we put the others into group Π_1.

B Theoretical discussion

B.1 Theoretical discussion relating to function extension

Here we describe models where the nonrandom approach (i.e. the second approach in section 3.2, where the curve is extended by its "nearest" fragment) performs well in reconstructing missing parts of curves. Our discussion is under the assumption that the closeness of a fragment to the curve being extended is measured by the vertical distance D introduced in section 3.2. Since we apply the extension method groupwise, it suffices to treat cases where all curves come from a single population.

We start with a simple model where the curves are all vertical translations of one another. There, if X_i denotes the ith observed random function, we have $X_i(t) = \mu(t) + V_i$, where μ denotes the mean function of the population and the V_is are i.i.d. random variables. In this model our curve extension procedure results in perfect reconstruction. To explain this point, let us suppose that X_{i_1} and X_{i_2} are observed on intervals $\mathcal{I}_{i_1} = [a_{i_1}, b_{i_1}]$ and $\mathcal{I}_{i_2} = [a_{i_2}, b_{i_2}]$, respectively, and that $a_{i_2} < b_{i_1} < b_{i_2}$. Then, if we use X_{i_2} to extend X_{i_1} from $[a_{i_1}, b_{i_1}]$ to $[a_{i_1}, b_{i_2}]$, and align $X_{i_1}(t)$ and $X_{i_2}(t)$ at $t = b_{i_1}$, we obtain the extended function $X_{i_1}^{\rm ext}$ which is defined, for $t \in [b_{i_1}, b_{i_2}]$, by

$$X_{i_1}^{\rm ext}(t) = X_{i_1}(b_{i_1}) - X_{i_2}(b_{i_1}) + X_{i_2}(t)\,. \tag{B.1}$$

The latter function equals $X_{i_1}(t)$, exactly, on $t \in [b_{i_1}, b_{i_2}]$.

More general models can be based on other approaches to curve extension, for example matching on the basis of gradients, but here we focus on vertical distance D, which is often relevant in practice. For many datasets, including the growth example depicted in Figure 1, the variation of random functions about their mean, for a given population, is captured well by treating the functions as approximate vertical translations, or scale multiples, of one another. To reflect this we model X_i by

$$X_i(t) = \mu(t) + V_i\, g(t) + W_i(t)\,, \tag{B.2}$$

where the random variable V_i and the random process W_i have zero mean, the V_is

are i.i.d., the W_is are i.i.d., and μ and g are deterministic functions. Moreover, we assume that $W_i(t)$ is small in size relatively to $V_i\, g(t)$. A simple way to model this mathematically is to assume that g is bounded away from zero, that the distribution of V_i does not depend on n, and that $W_i = \delta\, U_i$ for a fixed process U_i and a quantity $\delta = \delta(n)$ that decreases to zero slowly as $n \to \infty$. Of course, in practice δ is fixed, but its finite-sample smallness can be modeled asymptotically by allowing δ to decrease to 0.

Suppose we extend to $[a_{i_1}, b_{i_2}]$ the function X_i recorded on $\mathcal{I}_i = [a_{i_1}, b_{i_1}] \subset \mathcal{I}_0$, using a fragment X_{i_2} recorded on $\mathcal{I}_{i_2} = [a_{i_2}, b_{i_2}]$, where $a_{i_2} < b_{i_1} < b_{i_2}$. The function extension, $X_{i_1}^{\text{ext}}(t) = X_{i_1}(b_{i_1}) - X_{i_2}(b_{i_1}) + X_{i_2}(t)$, is given, for $t \in [b_{i_1}, b_{i_2}]$, by

$$\begin{aligned} X_{i_1}^{\text{ext}}(t) = X_{i_1}(t) &+ (V_{i_2} - V_{i_1})\, g(t) - (V_{i_2} - V_{i_1})\, g(b_{i_1}) \\ &+ W_{i_1}(b_{i_1}) - W_{i_2}(b_{i_1}) + W_{i_2}(t) - W_{i_1}(t)\,. \end{aligned}$$

More generally, suppose we extend X to the right ℓ times, in the same manner, via intervals $\mathcal{I}_{i_j} = [a_{i_j}, b_{i_j}]$ for $1 \le j \le \ell + 1$, where $a_{i_{j+1}} < b_{i_j} < b_{i_{j+1}}$ for $1 \le j \le \ell$. Then it can be shown that, for $t \in [b_{i_\ell}, b_{i_{\ell+1}}]$,

$$\begin{aligned} X_{i_1}^{\text{ext}}(t) =& X_{i_1}(t) + (V_{i_{\ell+1}} - V_{i_1})\, g(t) - \sum_{j=1}^{\ell} (V_{i_{j+1}} - V_{i_j})\, g(b_{i_j}) \\ &+ \sum_{j=1}^{\ell} \{W_{i_j}(b_{i_\ell}) - W_{i_{\ell+1}}(b_{i_j})\} + W_{i_{\ell+1}}(t) - W_{i_1}(t)\,. \end{aligned} \tag{B.3}$$

Recall that the random process $W_i(t) = O_P(\delta)$ is small. Since the W_is are independent for $1 \le i \le n$, then the series

$$\sum_{j=1}^{\ell} \{W_{i_j}(b_{i_\ell}) - W_{i_{+1}}(b_{i_j})\} + W_{i_{\ell+1}}(t) - W_{i_1}(t)\,, \tag{B.4}$$

appearing on the right-hand side of (B.3), is also small; more precisely, if ℓ is bounded then it is $O_P(\delta)$.

Note too that the successive X_{i_j}s are chosen by minimising D, in order that $|X_{i_{j+1}}(b_{i_j}) - X_{i_1}^{\text{ext}}(b_{i_j})|$ be as small as possible. Since the W_is are of order δ where

$\delta \to 0$, then (B.2) entails that X_{i_2} is chosen so that V_{i_2} is as close as possible to V_{i_1}, modulo a small amount of variation arising from the term $W_{i_1}(b_{i_1}) - W_{i_2}(b_{i_1})$; that i_3 is chosen so that V_{i_3} is as close as possible to V_{i_2}, modulo a small variation arising from $W_{i_2}(b_{i_2}) - W_{i_3}(b_{i_2})$; and so on. Therefore, noting that adjacent values of the variables V_i are distant $O_p(n^{-1})$ apart, except in the tails (here we are assuming that the V_is have a common continuous distribution), the quantities

$$V_{i_2} - V_{i_1}\,,\; V_{i_3} - V_{i_2}\,,\; \ldots\,,\; V_{i_{\ell+1}} - V_{i_\ell} \tag{B.5}$$

are also small, and in fact equal $O_p(\delta + n^{-1})$. Hence, provided again that ℓ is bounded, $V_{i_{\ell+1}} - V_{i_1}$ will also equal $O_p(\delta + n^{-1})$. Of course, $V_{i_{\ell+1}} - V_{i_1}$ is the sum of the quantities in (B.5), and so, using again the assumption that ℓ is fixed, $V_{i_{\ell+1}} - V_{i_1} = O_p(\delta + n^{-1})$. For all these reasons, formula (B.3) implies that $X_{i_1}^{\rm ext}(t)$ is approximately equal to $X_{i_1}(t)$.

The theory outlined above can be extended in several ways, for example by generalising the model at (B.2). One generalisation is to replace the product $V_i\, g(t)$ by the more complex form $g(t \,|\, Z_i)$, where Z_i is a random variable, $g(\,\cdot\,|\,\theta)$ is a function that depends monotonically and smoothly on a scalar parameter θ. Analogues of the results discussed above can be derived in this case. Results such as these show that, in cases of practical interest, function extension can come close to recovering the true functions X_i, even though their support intervals were only fragments of the original interval $\mathcal{I}_0$, and therefore lead to classification performance that approaches that which would have been obtained if there had been no censorship.

In practice, when n is small and ℓ is large or δ is not small enough, noise can accumulate as we extend a fragment to a larger and larger interval. However, since, our mean and covariance estimators put less weight on fragments that require more extension (see section 3.3), this noise does not have too much influence on classification performance. This can be proved rigorously, but requires longer arguments.

B.2 Conditions for Theorem 1

We assume that:

(a) for $k = 0, 1$, n/n_k is bounded as $n \to \infty$;
(b) $E_k\{X(t)^4\}$ is bounded uniformly in $k = 0, 1$ and $t \in \mathcal{I}$;
(c) the ratio θ_{0j}/θ_{1j} is bounded away from zero and infinity as $j \to \infty$;
(d) for both $k = 0$ and $k = 1$ there are no ties in the sequence $\theta_{k1}, \theta_{k2}, \ldots$;
(e) for $(k_1, k_2) = (0, 1)$ and $(1, 0)$ the series $Q_1(k_1, k_2 \,|\, j_0)$ either converges in probability to $Q_1(k_1, k_2 \,|\, \infty)$ as $j_0 \to \infty$, in which case $Q_1(k_1, k_2 \,|\, \infty)$ is assumed to be finite with probability 1, or diverges to $-\infty$ in probability as $j_0 \to \infty$, in the sense that, $\forall C > 0$, $P\{Q_1(k_1, k_2 \,|\, j_0) \le -C\} \to 1$ as $j_0 \to \infty$;
(f) for $(k_1, k_2) = (0, 1)$ and $(1, 0)$ the series $Q_2(k_1, k_2 \,|\, j_0)$ either converges in probability to $Q_2(k_1, k_2 \,|\, \infty)$ as $j_0 \to \infty$, in which case $Q_2(k_1, k_2 \,|\, \infty)$ is assumed to be finite with probability 1, or diverges to $+\infty$ in probability as $j_0 \to \infty$, in the sense that, $\forall C > 0$, $P\{Q_2(k_1, k_2 \,|\, j_0) > C\} \to 1$ as $j_0 \to \infty$; (B.6)
(g) if, for $(k_1, k_2) = (0, 1)$ or $(1, 0)$, and $\ell = 1$ or 2, $Q_\ell(k_1, k_2 \,|\, \infty)$ is finite with probability 1, then the distribution of $Q_\ell(k_1, k_2 \,|\, j_0)$ is continuous for $j_0 = 1, 2, \ldots, \infty$;
(h) the bivariate distributions of the interval endpoints, for data from Π_0 and Π_1, are continuous with respective probability densities f_0 and f_1, the supports of which are identical.

To avoid having to give a model describing the relationship between the functions and the respective intervals on which they are defined, we impose the following simplifying assumption; it can be dispensed with at the cost of a longer proof:

for $k = 0, 1$ the pairs (A_{kj}, B_{kj}), representing the endpoints of the intervals $\mathcal{I}_{kj} = [A_{kj}, B_{kj}]$ for $1 \le j \le n_k$, are independent and identically distributed and are independent too of the independent and identically distributed functions X_{kj}, for $1 \le j \le n_k$, defined on the interval $\mathcal{I}_0$. (B.7)

Conditions (B.6)(a)–(d) are conventional, and (B.6)(g) is mild, for example holding when X is a Gaussian process. Condition B.6(h) eliminates pathological cases, where we could construct a classifier based solely on the interval endpoints and for which the probability of correct classification would converge to 1 geometrically fast as sample size increased. Assumptions (B.6)(e) and (B.6)(f) hold quite generally, and lead to perfect classification when (i) the two means, or (ii) the two eigenvalue se-

quences, or (iii) the two eigenfunction sequences, are sufficiently different, or when there are sufficient differences in two or three of the features (i)–(iii). More detail is given below, where for simplicity we treat cases where just one of (i)–(iii) holds.

Case (i): Differences between mean functions alone. For simplicity we assume here that $\theta_{kj} = \theta_j$ and $\phi_{kj} = \phi_j$, not depending on k. Then $Q_1(k_1, k_2 \,|\, j_0) \equiv 0$ for $(k_1, k_2) = (0, 1)$ or $(1, 0)$, and moreover, $\delta_{kj} = \delta_j \equiv \theta_j^{-1/2} \int_{\mathcal{I}} (\mu_0 - \mu_1)\, \phi_j$, a sequence of constants not depending on k. It follows that if $s(\delta) \equiv \sum_{j \geq 1} \delta_j^2 < \infty$ then, for $(k_1, k_2) = (0, 1)$ or $(1, 0)$, the random variable $Q_2(k_1, k_2 \,|\, \infty)$ is finite with probability 1, whereas if $\sum_{j \geq 1} \delta_j^2 = \infty$ then $Q_2(k_1, k_2 \,|\, j_0) \to +\infty$ in probability as $j_0 \to \infty$. (Here we have used the fact that principal component scores are uncorrelated, and so $\mathrm{var}\{Q_2(k_1, k_2 \,|\, j_0)\} = \sum_{j \leq j_0} \delta_j^2$.) In summary, (B.6)(e) and (B.6)(f) hold with $Q_1(k_1, k_2 \,|\, \infty)$ and $Q_2(k_1, k_2 \,|\, \infty)$ always being well defined and finite with probability 1, if $s(\delta) < \infty$ but $Q_2(k_1, k_2 \,|\, j_0) \to \infty$ otherwise.

Case (ii): Differences between eigenvalue sequences alone. Here, to simplify our analysis we assume that $\phi_{kj} = \phi_j$, not depending on k, and $\mu_0 = \mu_1$. However, we allow the eigenvalue sequences to differ, and in particular we write $\theta_{1j} = \theta_{0j}\, (1 - c_j)^{-1}$ where the c_js are constants. To simplify our argument we suppose too that the ratio θ_{1j}/θ_{0j} is bounded away from zero and infinity. Then $Q_2(k_1, k_2 \,|\, j_0) \equiv 0$,

$$Q_1(0, 1 \,|\, j_0) = \sum_{j=1}^{j_0} \{c_j + \log(1 - c_j)\} + \sum_{j=1}^{j_0} \frac{c_j}{\theta_{0j}}\, (1 - E_0) \left\{ \int_{\mathcal{I}} (X - \mu_1)\, \phi_j \right\}^2, \quad \text{(B.8)}$$

$$Q_1(1, 0 \,|\, j_0) = \sum_{j=1}^{j_0} \left\{1 - (1 - c_j)^{-1} - \log(1 - c_j)\right\} + \sum_{j=1}^{j_0} \frac{1 - (1 - c_j)^{-1}}{\theta_{1j}}\, (1 - E_1) \left\{ \int_{\mathcal{I}} (X - \mu_0)\, \phi_j \right\}^2, \quad \text{(B.9)}$$

where in the first instance X is drawn from Π_0, and in the second, from Π_1. Observe too that

$$\begin{aligned} &\text{each of the series } \textstyle\sum_j \{c_j + \log(1 - c_j)\} \text{ and } \sum_j \{1 - (1 - c_j)^{-1} - \log(1 - c_j)\} \\ &\text{converges if } \textstyle\sum_j c_j^2 < \infty, \text{ and diverges to } -\infty \text{ in asymptotic proportion to} \\ &\textstyle\sum_{j \leq j_0} c_j^2 \text{ otherwise.} \end{aligned} \quad \text{(B.10)}$$

Now, the random variables $\int_{\mathcal{I}}(X - EX)\,\phi_j$ are uncorrelated, and in numerical work are often taken to be independent; they are always independent if X is Gaussian. Therefore, in a great many cases the second series on the far right-hand side of (B.8), and on the far right-hand side of (B.9), converges if and only if its mean square is finite, and in particular if and only if $t(c) \equiv \sum_{j\geq 1} c_j^2 < \infty$, and is of smaller order than $\sum_{j\leq j_0} c_j^2$ otherwise. Combining this property with (B.10) we deduce that, in such instances, (B.6)(e) and (B.6)(f) hold with $Q_2(k_1, k_2 \,|\, \infty)$ being well defined and finite with probability 1, $Q_1(k_1, k_2 \,|\, \infty)$ being finite with probability 1 if $t(c) < \infty$ holds, and $Q_1(k_1, k_2 \,|\, j_0) \to -\infty$ in probability if $t(c) = \infty$.

Case (iii): Differences between eigenfunction sequences alone. We assume here that the differences are only in terms of the eigenfunction sequences, and the eigenvalue sequences and the means are the same for both populations. Many examples in this setting can be treated as in case (ii). For instance, if we initially take the set of eigenfunctions $\phi_1, \phi_2, \ldots$ to be the same for both populations, and perturb the sequence of eigenvalues in one population to obtain those in another, then, after we have reordered the (eigenvalue, eigenfunction) pairs so that the eigenvalues are arranged in decreasing order, as is conventional before indexing the pairs, we have a setting that can be viewed as one where the eigenvalue sequence is common to both populations but the eigenfunction sequences differ. Since our treatment of case (ii) above did not require the eigenvalues for either population to be arranged in decreasing order, it therefore applies also to the present setting, with the same conclusions being drawn; see the last sentence in case (ii). Many other instances can be treated with only a little more effort, for example those where the eigenvalue sequences are identical and the function spaces generated by the pairs $(\phi_{k,2j-1}, \phi_{k,2j})$, for $j \geq 1$, are identical for both populations but the eigenfunctions $\phi_{k,2j-1}$ and $\phi_{k,2j}$ differ for $k = 0, 1$ and each $j \geq 1$. In these settings the conclusions noted in the last paragraph of case (ii) continue to apply.

B.3 Proof of Theorem 1

Step 1: Showing that $\mathcal{I}'$ and $\mathcal{I}''$ can be taken equal to $\mathcal{I}$. Assumption (B.6)(h) implies that, if (a, b) is in the interior of the support of f_0 or f_1, then it is in the interior of the support of both and hence, for some $\epsilon_1, \epsilon_2 > 0$, the values of $n_k(t)$ and $n_k(s, t)$, defined immediately below (2.11), satisfy

$$P\Big\{ \min_{k=0,1} \min_{a-\epsilon_1<t<b+\epsilon_1} n_k(t) > \epsilon_2\, n \Big\} \to 1\,,$$
$$P\Big\{ \min_{k=0,1} \min_{a-\epsilon_1<s<t<b+\epsilon_1} n_k(s,t) > \epsilon_2\, n \Big\} \to 1\,.$$

as $n \to \infty$. Therefore, if ν_0 diverges and is no larger than $\epsilon_2 n$ then $P(\mathcal{I} = \mathcal{I}' = \mathcal{I}'') \to 1$. This property allows us to work below with $\mathcal{I}$ rather than $\mathcal{I}'$ or $\mathcal{I}''$.

Step 2: Deterministic approximation to $T_{\rm fun}(x, \mathcal{I}\,|\,j_0, 0)$, at (2.6). Given a function g_1 that is square-integrable on $\mathcal{I}$, and a function g_2 of two variables that is square integrable on $\mathcal{I} \times \mathcal{I}$, define $\|g_1\|^2 = \int_{\mathcal{I}} g_1^2$ and $|\!|\!| g_2 |\!|\!|^2 = \int_{\mathcal{I}}\int_{\mathcal{I}} g_2^2$. It is known (see Theorem 1 of Hall and Hosseini-Nasab, 2006) that

$$\sup_{j\ge1} \big|\hat\theta_j - \theta_j\big| \le \widehat\Delta\,, \quad \big\|\hat\phi_j - \phi_j\big\| \le 8^{1/2}\, \widehat\Delta/\delta_j\,, \tag{B.11}$$

where $\widehat\Delta = |\!|\!| \widehat K - K |\!|\!|$ and $\delta_j = \min_{1\le k\le j} (\theta_k - \theta_{k+1})$. Since $E_k\{X(t)^4\}$ is bounded uniformly in $k = 0, 1$ and $t \in \mathcal{I}$ (see (B.6)(b)) then $E\{\widehat K(s,t) - K(s,t)\}^2 \to 0$ for each $(s, t) \in \mathcal{I}$, and therefore, $\widehat\Delta \to 0$ in probability. This property, (B.11) and the fact that there are no ties among the eigenvalue sequences (see (B.6)(d)) imply that there exist a sequence of positive integers $j(n)$, increasing to infinity, and a sequence of positive numbers $\epsilon(n) \le \frac12$, decreasing to zero, such that

$$j(n) \le \epsilon(n)^{-1}\,, \quad \min_{k=0,1} \min_{1\le j\le j(n)} \theta_{kj} \ge \epsilon(n)\,, \quad \max_{k=0,1} \|\mu_k - \bar X\| = O_p\big\{\epsilon(n)^4\big\} \tag{B.12}$$

and $P(\mathcal{E}_n) \to 1$ as $n \to \infty$, where $\mathcal{E}_n$ is the event defined by

$$\mathcal{E}_n = \max_{k=0,1} \max_{1\le j\le j(n)} \big(\big|\hat\theta_{kj} - \theta_{kj}\big| + \big\|\hat\phi_{kj} - \phi_{kj}\big\|\big) \le \epsilon(n)^4\,.$$

Then, if $\mathcal{E}_n$ obtains and $1 \le j_0 \le j(n)$,

$$\sum_{j=1}^{j_0} \left| \log \hat{\theta}_{kj} - \log \theta_{kj} \right| \le \sum_{j=1}^{j_0} \left| \log \left(1 - \left| \frac{\hat{\theta}_{kj} - \theta_{kj}}{\theta_{kj}} \right| \right) \right| \le \sum_{j=1}^{j(n)} \left| \log \left\{ 1 - \epsilon(n)^3 \right\} \right| = O\left\{ \epsilon(n)^2 \right\}, \tag{B.13}$$

and also,

$$\max_{k=0,1} \max_{1 \le j \le j(n)} \left| \hat{\theta}_{kj}^{-1} - \theta_{kj}^{-1} \right| \le \max_{k=0,1} \max_{1 \le j \le j(n)} \frac{|\hat{\theta}_{kj} - \theta_{kj}|}{(\theta_{kj} - |\hat{\theta}_{kj} - \theta_{kj}|)\, \theta_{kj}} \le \frac{\epsilon(n)^4}{\{\epsilon(n) - \epsilon(n)^4\}\, \epsilon(n)} \le 2\, \epsilon(n)^2, \tag{B.14}$$

and moreover, if g is a square-integrable function of one variable,

$$\int_{\mathcal{I}} |g| \left| \hat{\phi}_{kj} - \phi_{kj} \right| \le \|g\| \left\| \hat{\phi}_{kj} - \phi_{kj} \right\| \le \|g\|\, \epsilon(n)^4 . \tag{B.15}$$

Recall the definition of $T_{\rm fun}(x, \mathcal{I} \,|\, j_0, 0)$ at (2.6), and put

$$\begin{aligned} T_k^0(X, \mathcal{I} \,|\, j_0) &= \sum_{j=1}^{j_0} \left(\xi_{kj}^2 + \log \theta_{kj} \right) \\ &= \sum_{j=1}^{j_0} \left(\frac{1}{\theta_{kj}} \left[\int_{\mathcal{I}} \{X(t) - \mu_k(t)\}\, \phi_{kj}(t)\, dt \right]^2 + \log \theta_{kj} \right), \\ T_{\rm fun}^0(X, \mathcal{I} \,|\, j_0, 0) &= T_0^0(X, \mathcal{I} \,|\, j_0) - T_1^0(X, \mathcal{I} \,|\, j_0), \end{aligned}$$

where (θ_{kj}, ϕ_{kj}) and ξ_{kj} are as defined in section 2.3. Combining (B.12)–(B.15) we deduce that if $\mathcal{E}_n$ holds and $j_0 \le j(n)$ then:

$$\begin{aligned} &\left| T_{\rm fun}(X, \mathcal{I} \,|\, j_0, 0) - T_{\rm fun}^0(X, \mathcal{I} \,|\, j_0, 0) \right| \\ &\qquad \le j(n) \max_{k=0,1} \max_{1 \le j \le j(n)} \left[\left| \hat{\theta}_{kj}^{-1} - \theta_{kj}^{-1} \right| \left(\left\| X - \bar{X}_k \right\| \left\| \hat{\phi}_{kj} \right\| \right)^2 \right. \\ &\qquad\qquad + \epsilon(n)^{-1} \left| \int_{\mathcal{I}} \left\{ \left(X - \bar{X}_k \right) \hat{\phi}_{kj} - \left(X - \mu_k \right) \phi_{kj} \right\} \right| \\ &\qquad\qquad \left. \times \left| \int_{\mathcal{I}} \left\{ \left(X - \bar{X}_k \right) \hat{\phi}_{kj} + \left(X - \mu_k \right) \phi_{kj} \right\} \right| \right] + O_p\{\epsilon(n)\} \end{aligned}$$

$$
\begin{aligned}
&\le \epsilon(n)^{-1} \max_{k=0,1} \max_{1 \le j \le j(n)} \Big\{ 2\,\epsilon(n)^2 \left(\left\| X - \bar{X}_k \right\| \left\| \hat{\phi}_{kj} \right\| \right)^2 \\
&\qquad + \epsilon(n)^{-1} \left(\left\| X - \bar{X}_k \right\| \left\| \hat{\phi}_{kj} - \phi_{kj} \right\| + \left\| \bar{X}_k - \mu_k \right\| \left\| \phi_{kj} \right\| \right) \\
&\qquad\qquad \times \left(\left\| X - \bar{X}_k \right\| \left\| \hat{\phi}_{kj} \right\| + \left\| X - \mu_k \right\| \left\| \phi_{kj} \right\| \right) \Big\} + O_p\{\epsilon(n)\} \\
&= O_p\big\{\epsilon(n)^{1/3}\big\}\,,
\end{aligned}
$$

where the latter bound holds uniformly in functions X, defined on $\mathcal{I}$, for which $\|X\| \le \epsilon(n)^{-1/3}$. (We have used the fact that, for all j and k, $\|\phi_{kj}\| = \|\hat{\phi}_{kj}\| = 1$.) Hence, for all $\epsilon > 0$,

$$
\max_{k=0,1} P\Big\{ \Big| T_{\rm fun}(X, \mathcal{I} \,|\, j_0, 0) - T^0_{\rm fun}(X, \mathcal{I} \,|\, j_0, 0) \Big| > \epsilon \;\Big|\; \mathcal{I}\,;\, (X, \mathcal{I}) \in \Pi_k \Big\} \to 0 \qquad \text{(B.16)}
$$

as $n \to \infty$, where, here and below, the notation $P\{\ldots \,|\, \mathcal{I};\, (X, \mathcal{I}) \in \Pi_k\}$ denotes probability measure conditional on the event that $(X, \mathcal{I})$ was drawn from population Π_k, and also conditional on the interval $\mathcal{I}$. Note too that, by (B.7), X is independent of $\mathcal{I}$.

Step 3: Expanding $T^0_{\rm fun}(x, \mathcal{I} \,|\, j_0, 0)$. If $(X, \mathcal{I})$ is drawn from Π_0 then $T^0_{\rm fun}(X, \mathcal{I} \,|\, j_0, 0)$ can be expanded as follows:

$$
T^0_{\rm fun}(X, \mathcal{I} \,|\, j_0, 0) = U_1(X, \mathcal{I} \,|\, j_0) - U_2(X, \mathcal{I} \,|\, j_0)\,, \qquad \text{(B.17)}
$$

where

$$
\begin{aligned}
U_1(X, \mathcal{I} \,|\, j_0) &= \sum_{j=1}^{j_0} \Bigg[\frac{1}{\theta_{0j}} \left\{ \int_{\mathcal{I}} (X - \mu_0)\, \phi_{0j} \right\}^2 - \frac{1}{\theta_{1j}} \left\{ \int_{\mathcal{I}} (X - \mu_0)\, \phi_{1j} \right\}^2 \\
&\qquad\qquad + \log(\theta_{0j}/\theta_{1j}) \Bigg] = Q_1(0, 1 \,|\, j_0)\,, \\
U_2(X, \mathcal{I} \,|\, j_0) &= \sum_{j=1}^{j_0} \frac{1}{\theta_{1j}} \Bigg[2 \left\{ \int_{\mathcal{I}} (X - \mu_0)\, \phi_{1j} \right\} \left\{ \int_{\mathcal{I}} (\mu_0 - \mu_1)\, \phi_{1j} \right\} \\
&\qquad\qquad + \left\{ \int_{\mathcal{I}} (\mu_0 - \mu_1)\, \phi_{1j} \right\}^2 \Bigg] = Q_2(0, 1 \,|\, j_0)\,.
\end{aligned}
$$

Therefore, noting (B.6)(e) and (B.6)(f),

$$U_1(X, \mathcal{I} \,|\, j_0) \to \begin{cases} Q_1(0,1) & \text{if } Q_1(0,1) \text{ is finite with probability 1} \\ -\infty & \text{otherwise} \end{cases} \tag{B.18}$$

$$U_2(X, \mathcal{I} \,|\, j_0) \to \begin{cases} Q_2(0,1) & \text{if } Q_2(0,1) \text{ is finite with probability 1} \\ +\infty & \text{otherwise} \end{cases} \tag{B.19}$$

where the convergence is in probability.

On the other hand, if X is drawn from Π_1 then $T^0_{\text{fun}}(X, \mathcal{I} \,|\, j_0, 0)$ can be expanded as follows:

$$T^0_{\text{fun}}(X, \mathcal{I} \,|\, j_0, 0) = U_3(X, \mathcal{I} \,|\, j_0) + U_4(X, \mathcal{I} \,|\, j_0)\,, \tag{B.20}$$

where

$$U_3(X, \mathcal{I} \,|\, j_0) = \sum_{j=1}^{j_0} \left[\frac{1}{\theta_{0j}} \left\{ \int_{\mathcal{I}} (X - \mu_1)\, \phi_{0j} \right\}^2 - \frac{1}{\theta_{1j}} \left\{ \int_{\mathcal{I}} (X - \mu_1)\, \phi_{1j} \right\}^2 + \log(\theta_{0j}/\theta_{1j}) \right] = -Q_1(1, 0 \,|\, j_0)\,,$$

$$U_4(X, \mathcal{I} \,|\, j_0) = \sum_{j=1}^{j_0} \frac{1}{\theta_{0j}} \left[2 \left\{ \int_{\mathcal{I}} (X - \mu_1)\, \phi_{0j} \right\} \left\{ \int_{\mathcal{I}} (\mu_1 - \mu_0)\, \phi_{0j} \right\} + \left\{ \int_{\mathcal{I}} (\mu_1 - \mu_0)\, \phi_{0j} \right\}^2 \right] = Q_2(1, 0 \,|\, j_0)\,.$$

Therefore, in view of (B.20),

$$U_3(X, \mathcal{I} \,|\, j_0) \to \begin{cases} -Q_1(1,0) & \text{if } Q_1(1,0) \text{ is finite with probability 1} \\ \infty & \text{otherwise} \end{cases} \tag{B.21}$$

$$U_4(X, \mathcal{I} \,|\, j_0) \to \begin{cases} Q_2(1,0) & \text{if } Q_2(1,0) \text{ is finite with probability 1} \\ \infty & \text{otherwise} \end{cases} \tag{B.22}$$

where the convergence is in probability.

Step 4: Completion. It follows from (B.16) that the criterion $T(X, \mathcal{I} \,|\, j_0)$, at (5.1), satisfies

$$\max_{k=0,1} P\Big\{ \Big| T(X, \mathcal{I} \,|\, j_0) - T^0(X, \mathcal{I} \,|\, j_0) \Big| > \epsilon \;\Big|\; \mathcal{I}\,;\, (X, \mathcal{I}) \in \Pi_k \Big\} \to 0 \tag{B.23}$$

for all $\epsilon > 0$.

Part (I) of Theorem 1 follows on combining (B.17), (B.18), (B.19) and (B.23), and part (II) follows on combining (B.17), (B.21), (B.22), and (B.23). The uniformity claimed in Theorem 1 is trivial if $Q_1(0,1)$ or $Q_2(0,1)$ is infinite (when sampling from Π_0), or if $Q_1(1,0)$ or $Q_2(1,0)$ is infinite (in the case of sampling from Π_1); and, when those quantities are finite, the uniformity follows from the convergences in the first cases of (B.18), (B.19), (B.21) and (B.22).

B.4 Proof of Theorem 2

We treat only the case where the support intervals of the censored training data are extended to $\mathcal{I}$ using random function extension. Let $n = \min(n_0, n_1)$ and define the event

$$\mathcal{E}_k = \left\{ \mathcal{I} \subseteq \bigcup_{j=1}^{n_k} \mathcal{I}_{kj} \right\}.$$

It follows from (5.6) that $P(\mathcal{E}_k) \to 1$ as $n \to \infty$. Conditional on $\mathcal{E}_k$ being true, let X^*_{kj} denote a particular version of the extension of X_{kj} to an interval that contains $\mathcal{I}$, and define $\bar{X}^*_k = n_k^{-1} \sum_j X^*_{kj}$ and $\bar{\Delta}^*_k = n_k^{-1} \sum_j (X^*_{kj} - \mu_k)$. Since $\mathcal{E}_k$ holds then the supports of both $\bar{X}^*_k$ and $\bar{\Delta}^*_k$ include $\mathcal{I}$, and moreover, $E\{\bar{\Delta}^*_k(t)^2\, I(\mathcal{E}_k)\} \to 0$ uniformly in $t \in \mathcal{I}$, where $I(\mathcal{E}_k)$ denotes the indicator function of $\mathcal{E}_k$. Therefore,

$$\int_{\mathcal{I}} \left\{ (\bar{\Delta}^*_k)^2 + |\bar{\Delta}^*_k|\,|\Delta| + |\bar{\Delta}^*_k|\,|d| \right\} \to 0 \tag{B.24}$$

in probability as $n \to \infty$.

Recall that $(X, \mathcal{I})$ came from Π_0, and note that

$$\int_{\mathcal{I}} (X - \bar{X}^*_k)^2 = \begin{cases} \int_{\mathcal{I}} (\Delta - \bar{\Delta}^*_0)^2 & \text{if } k = 0 \\ \int_{\mathcal{I}} (\Delta - \bar{\Delta}^*_1 + d)^2 & \text{if } k = 1\,. \end{cases}$$

Hence, noting the definition of $D(X)$ at (5.7),

$$D(X) = \int_{\mathcal{I}} (\bar{\Delta}^*_1 - \bar{\Delta}^*_0)\,(\bar{\Delta}^*_1 + \bar{\Delta}^*_0 - 2\,\Delta) + 2 \int_{\mathcal{I}} (\Delta - \bar{\Delta}^*_1)\,d + \int_{\mathcal{I}} d^2$$

$$= 2 \int_{\mathcal{I}} \Delta \, d + \int_{\mathcal{I}} d^2 + o_p(1) \,, \tag{B.25}$$

where the last identity follows from (B.24). Since $d > 0$ is fixed then (5.8) follows from (B.26).

B.5 Proof of Theorem 3

Under the assumptions imposed in the theorem, if $\bar{X}_k$ represents the mean of the training data from Π_k, with their support restored to $\mathcal{I}_0$; if μ_k is the corresponding population mean; and if X is drawn from Π_k; then the statistic $D(X)$, at (5.7), assumes the form

$$\begin{aligned} D(X) &= \int_{\mathcal{I}} (X - \bar{X}_1)^2 - \int_{\mathcal{I}} (X - \bar{X}_0)^2 = \int_{\mathcal{I}} (X - \mu_1)^2 - \int_{\mathcal{I}} (X - \mu_0)^2 + O_p\big(n^{-1/2}\big) \\ &= T_k(X) + d_k + O_p\big(n^{-1/2}\big) \,. \end{aligned} \tag{B.26}$$

The theorem follows directly from (B.26).

C Other classifiers

C.1 Classifier for section 4.1

Since the vectors (A_{kj}, B_{kj}) are only of dimension two, if sample size were large enough we could also use the more sophisticated nonparametric form of the Bayes classifier. Specifically, for $k = 0$ and 1, let $\hat{f}_k$ be a nonparametric kernel estimator of the bivariate density f_k of the joint distribution of the endpoints of $\mathcal{I} = [A, B]$ when X is drawn from Π_k, based on the data (A_{kj}, B_{kj}), for $1 \le j \le n_k$. That is,

$$\hat{f}_k(s,t) = \frac{1}{n_k h_1 h_2} \sum_{j=1}^{n_k} K\left(\frac{s - A_{kj}}{h_1} \,,\, \frac{t - B_{kj}}{h_2} \right) , \tag{C.1}$$

where K is a bivariate kernel function integrating to one, and $h_1 > 0$ and $h_2 > 0$ are two bandwidths. The nonparametric Bayes classifier assigns X to Π_0 if

$$S_{\text{int}}(\mathcal{I} \,|\, w) = \log \hat{f}_1(A, B) - \log \hat{f}_0(A, B) + w \tag{C.2}$$

is negative, and to Π_1 otherwise, with w as in section 4.1. (Here S_{int} is as defined in section 4.1.)

C.2 Classifier for section 3.3

The linear discriminant version of the method in section 3.3 amounts to replacing there each occurrence of $\hat{K}_k$ by $\hat{K}$, where

$$\widehat{K}(s,t) = \sum_{k=0,1} \sum_{j=1}^{n_k} \tilde{w}_{kj} \{\tilde{X}_{kj}(s) - \bar{X}_k(s)\} \{\tilde{X}_{kj}(t) - \bar{X}_k(t)\}. \tag{C.3}$$

Here we take

$$\tilde{w}_{kj} = \frac{L\big(\|\mathcal{I} - \mathcal{I}_{kj}\|_{\text{int}}/\{(b-a)h\}\big)}{\sum_{k=0,1} \sum_{\ell=1}^{n_k} L\big(\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\text{int}}/\{(b-a)h\}\big)},$$

where h is a bandwidth. As for h_0 and h_1, we let h be the rth empirical quantile of $\|\mathcal{I} - \mathcal{I}_{k\ell}\|_{\text{int}}/(b-a)$, for $k = 0, 1$ and $\ell = 1, \ldots, n_k$, where $0 \le r \le 1$ is the same for h, h_0 and h_1 and is chosen by CV as described in section 3.3. Moreover, we replace each occurrence of $\hat{\theta}_{kj}$ and $\hat{\phi}_{kj}$ by $\hat{\theta}_j$ and $\hat{\phi}_j$, the empirical eigenvalues and eigenfunctions of $\hat{K}$. The rest of the procedure remains unchanged.

C.3 Leave-one-out quantities

C.3.1 Classifier based on T_{fun}

Analogously to the definitions at (2.6) and (2.7), if we leave out X_{0j_1} or if we leave out X_{1j_1}, respectively, then $T_{\text{fun},-j_1}(X, \mathcal{I} \,|\, j_0, w)$ is defined by, respectively,

$$T_{\text{fun},-j_1}(X, \mathcal{I} \,|\, j_0, w) = T_{0,-j_1}(X, \mathcal{I} \,|\, j_0) - T_1(X, \mathcal{I} \,|\, j_0) + w$$
$$T_{\text{fun},-j_1}(X, \mathcal{I} \,|\, j_0, w) = T_0(X, \mathcal{I} \,|\, j_0) - T_{1,-j_1}(X, \mathcal{I} \,|\, j_0) + w,$$

where

$$T_{k,-j_1}(X, \mathcal{I} \,|\, j_0) = \sum_{j=1}^{j_0} \left(\frac{1}{\hat{\theta}_{kj,-j_1}} \left[\int_{\mathcal{I}''} \{X(t) - \bar{X}_{k,-j_1}(t)\} \, \hat{\phi}_{kj,-j_1}(t) \, dt \right]^2 + \log \hat{\theta}_{kj,-j_1} \right),$$

and, letting $n_{k,-j_1}(t)$ and $n_{k,-j_1}(s,t)$ denote the numbers of elements in the sets $\mathcal{J}_k(t) \setminus \{j_1\}$ and $\mathcal{J}_k(s,t) \setminus \{j_1\}$, respectively, and assuming that $n_{k,-j_1}(s)$, $n_{k,-j_1}(t)$ and $n_{k,-j_1}(s,t)$ are all strictly positive,

$$\bar{X}_{k,-j_1}(t) = \frac{1}{n_{k,-j_1}(t)} \sum_{j \in \mathcal{J}_k(t) \setminus \{j_1\}} X_{kj}(t)\,,$$

with $\hat{\theta}_{k1,-j_1} \geq \hat{\theta}_{k2,-j_2} \geq \ldots$ and $\hat{\phi}_{k1,-j_1}, \hat{\phi}_{k2,-j_1}, \ldots$ coming from the spectral decomposition

$$\widehat{K}_{k,-j_1}(s,t) = \sum_{j=1}^{\infty} \hat{\theta}_{kj,-j_1}\, \hat{\phi}_{kj,-j_1}(s)\, \hat{\phi}_{kj,-j_1}(t)$$

of

$$\widehat{K}_{k,-j_1}(s,t) = \frac{1}{n_{k,-j_1}(s,t)} \sum_{j \in \mathcal{J}_k(s,t) \setminus \{j_1\}} \{X_{kj}(t) - \bar{X}_{k,-j_1}(t)\}\, \{X_{kj}(s) - \bar{X}_{k,-j_1}(s)\}\,.$$

C.3.2 Classifier based on S_{int}

Recall the definition of S_{int} in section 4.1. In the quadratic discriminant case, let

$$\begin{aligned}
\bar{A}_{k,-j_1} &= \frac{1}{n_k - 1} \sum_{j\,:\,j \neq j_1} A_{kj}\,, \quad \bar{B}_{k,-j_1} = \frac{1}{n_k - 1} \sum_{j\,:\,j \neq j_1} B_{kj}\,, \\
\hat{\sigma}^2_{Ak,-j_1} &= \frac{1}{n_k - 1} \sum_{j\,:\,j \neq j_1} (A_{kj} - \bar{A}_{k,-j_1})^2\,, \quad \hat{\sigma}^2_{Bk,-j_1} = \frac{1}{n_k - 1} \sum_{j\,:\,j \neq j_1} (B_{kj} - \bar{B}_{k,-j_1})^2\,, \\
\hat{\gamma}_{k,-j_1} &= \frac{1}{n_k - 1} \sum_{j\,:\,j \neq j_1} (A_{kj} - \bar{A}_{k,-j_1})\,(B_{kj} - \bar{B}_{k,-j_1}).
\end{aligned}$$

Then if we leave out X_{0j_1}, we define

$$\begin{aligned}
S_{\text{int},-j_1}(\mathcal{I} \,|\, w) =& (A - \bar{A}_{0,-j_1}, B - \bar{B}_{0,-j_1})\, \widehat{\Sigma}^{-1}_{0,-j_1}\, (A - \bar{A}_{0,-j_1}, B - \bar{B}_{0,-j_1})^{\mathrm{T}} + \log |\widehat{\Sigma}_{0,-j_1}| \\
& - (A - \bar{A}_1, B - \bar{B}_1)\, \widehat{\Sigma}^{-1}_1\, (A - \bar{A}_1, B - \bar{B}_1)^{\mathrm{T}} - \log |\hat{\Sigma}_1| + w,
\end{aligned}$$

and if we leave out X_{1j_1}, we define

$$\begin{aligned}
S_{\text{int},-j_1}(\mathcal{I} \,|\, w) =& (A - \bar{A}_0, B - \bar{B}_0)\, \widehat{\Sigma}^{-1}_0\, (A - \bar{A}_0, B - \bar{B}_0)^{\mathrm{T}} + \log |\hat{\Sigma}_0| \\
& - (A - \bar{A}_{1,-j_1}, B - \bar{B}_{1,-j_1})\, \widehat{\Sigma}^{-1}_{1,-j_1}\, (A - \bar{A}_{1,-j_1}, B - \bar{B}_{1,-j_1})^{\mathrm{T}} \\
& - \log |\widehat{\Sigma}_{1,-j_1}| + w.
\end{aligned}$$

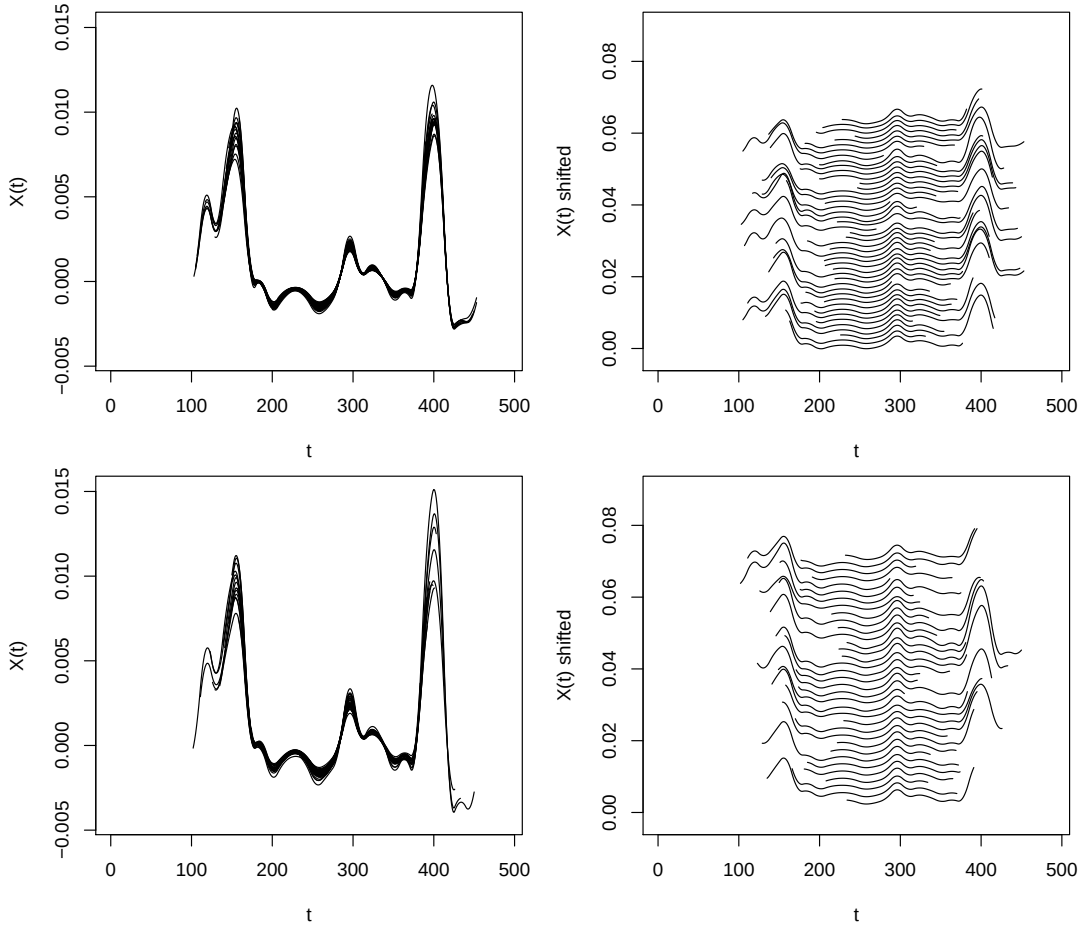

Figure D.1: Wheat curves, case (c). Left panel: the fragment curves in Π_0 (top) and Π_1 (bottom); right panel: shifted fragment curves in Π_0 (top) and Π_1 (bottom).

D Additional figures

The left panels of Figures D.1 and D.2 show a typical example of 100 fragments of wheat curves obtained in cases (c) and (d), respectively (see section 7.2.1). The right panels show the same curves, but to make the endpoints of the fragments more apparent we have translated each curve vertically by a small amount.

Figure D.3 shows, for each ethnic group, the growth curves introduced in section 7.2.2. Figure D.4 shows the curves of the FEV data by group, in Cases I and II introduced in section 7.2.3.

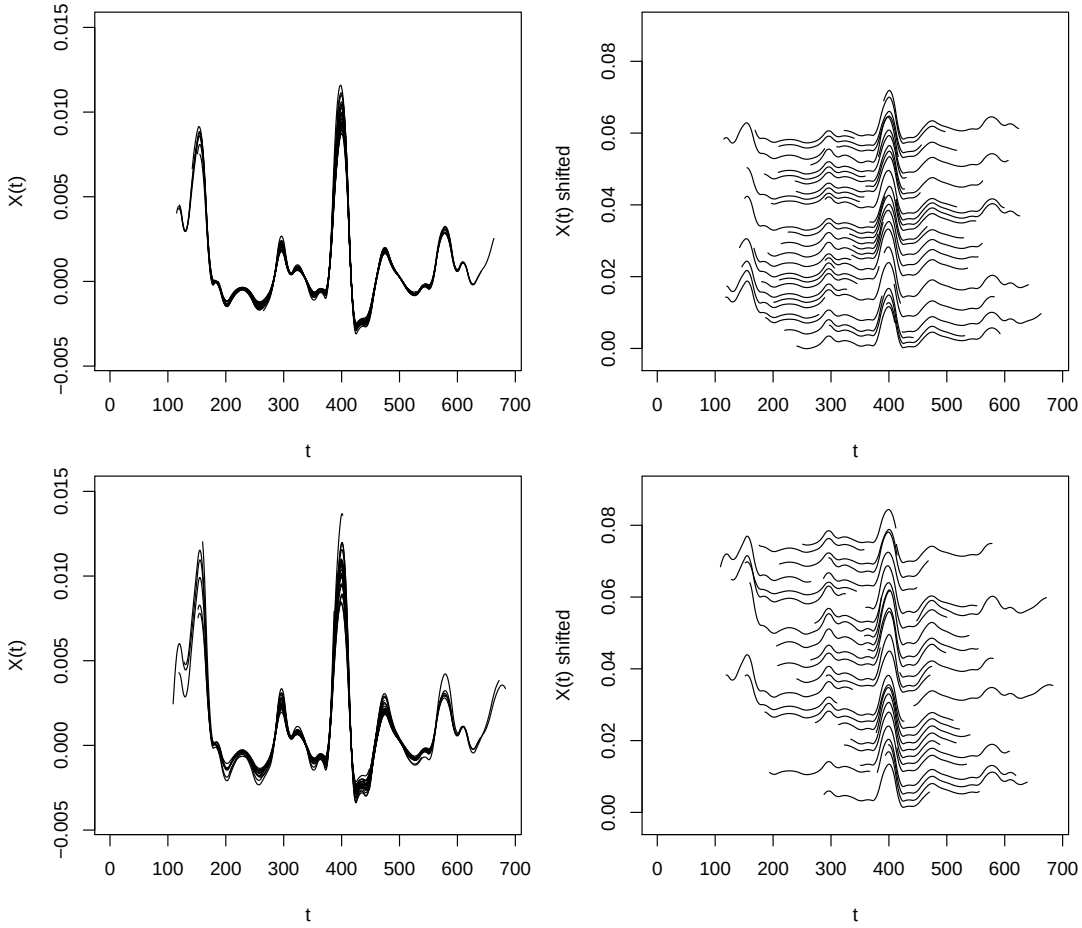

Figure D.2: Wheat curves, case (d). Left panel: the fragment curves in Π_0 (top) and Π_1 (bottom); right panel: shifted fragment curves in Π_0 (top) and Π_1 (bottom).

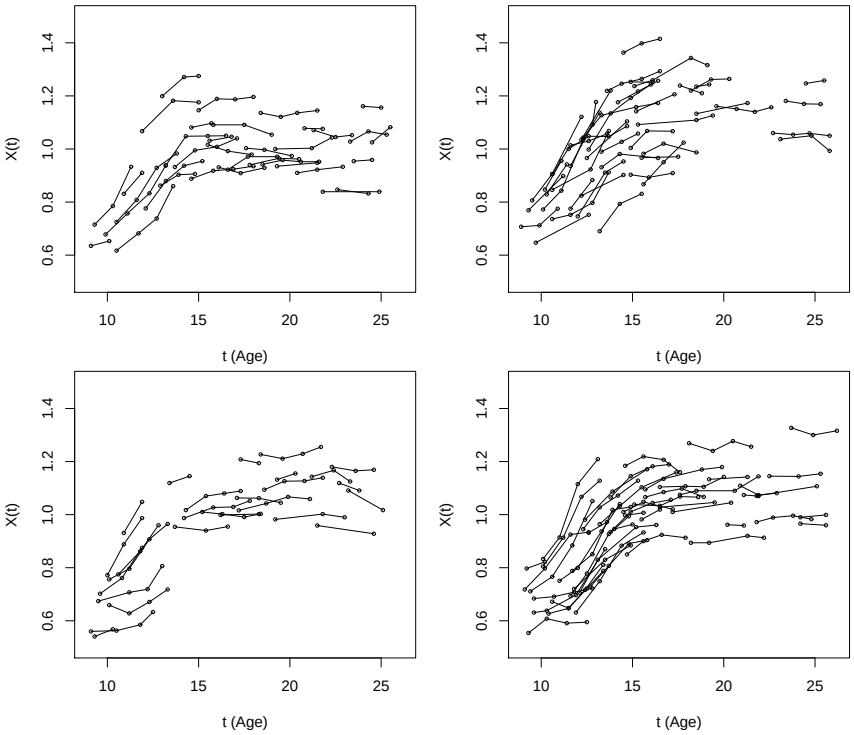

Figure D.3: Growth curves of 153 females from four ethnic groups: Asians (top left), Blacks (top right), Hispanics (bottom left) and Caucasians (bottom right).

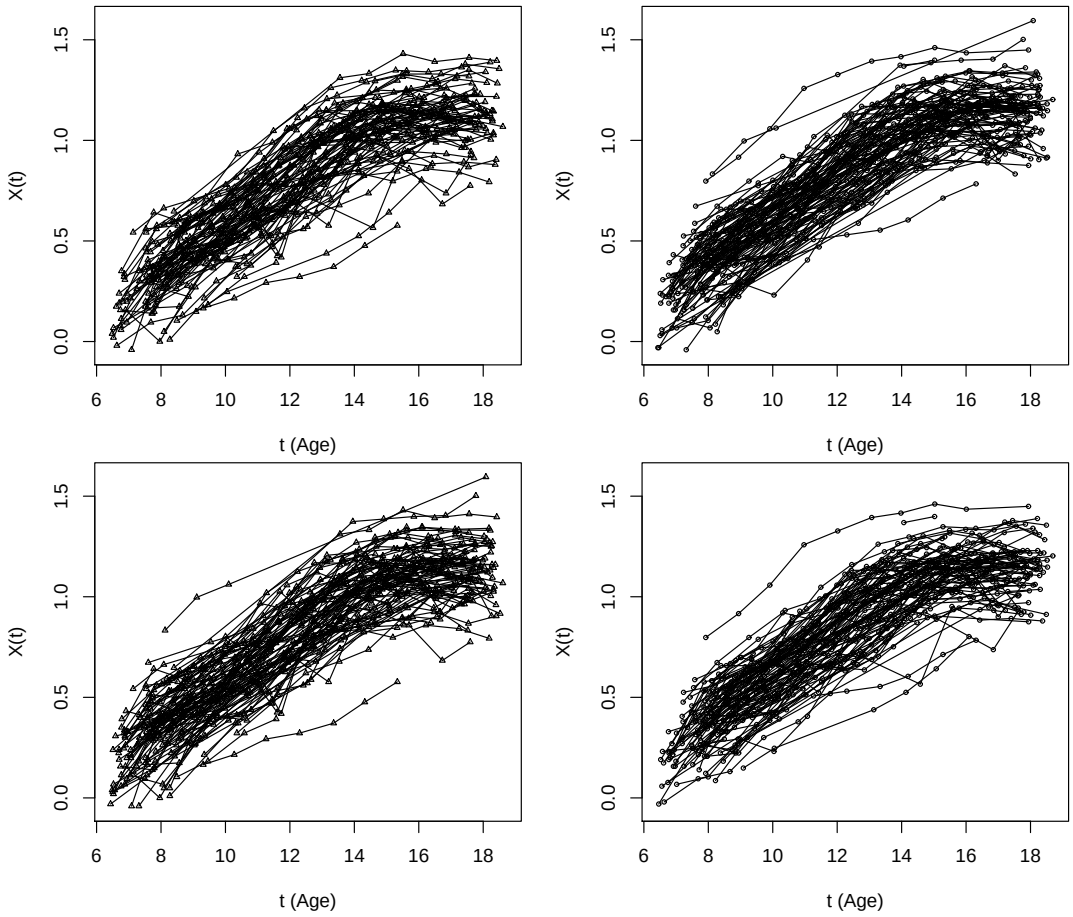

Figure D.4: Curves of log(FEV1) for 252 US girls. Left: Curves for group Π_0; right: curves for group Π_1; top: case I, bottom: case II.