Chapter 4

Nonparametric Tests and Prediction Sets for Functional Data

4.1 Introduction

This chapter studies data depth methodology in functional data analysis, for (i) detecting "outliers", (ii) constructing prediction sets, and (iii) performing two-sample tests. The motivation comes from emerging types of functional data in biotech manufacturing. The following are examples of routine tasks:

1. assess if the spectroscopy profile of a newly-manufactured batch is 'within expectation' relative to the profiles of historical batches. A spectroscopy profile captures chemical and/or biophysical characteristics of a material. A typical example is the absorption of light of different wavelengths over a prefixed range.
2. assess the difference or similarity between two sets of time series representing a key process attribute measured on a cell culture (or purification) before and after a critical process change.
3. compare particle size distributions before and after a change in starting material or process.

Currently these tasks are performed visually by well-trained specialists following a detailed protocol. A more data-driven approach would remove human bias.

In functional data analysis the observations are functions $t \mapsto X_i(t)$ of a continuous argument t. This setup fits the preceding examples, in which wavelength, time or particle size are conceptually continuous variables. In practice, one observes the values of these functions only at a finite grid of values t, and hence the observations are vectors $\mathbf{X}_i = \big(X_i(t_1), \cdots, X_i(t_d)\big)$. If we observe N curves at d grid points, then the full data will be an $N \times d$ data matrix $\mathbf{X}$. A data matrix resulting from functional data is different than a data matrix resulting from just any N observational vectors, as the columns are ordered and usually have a numerical reference. When visualizing the datamatrix, it is more intuitive to plot N curves against this numerical index, than to think of N points in d-dimensional space. Also the

number of columns, resulting from the sampling frequency of the grid points $t_1, \cdots, t_d$, is to a certain extent a subjective choice. Ideally it should reflect our expectation on the variability of the underlying functions. Less smooth functions should be observed on a finer grid, as otherwise important local features may be missed. To make the distinction explicit, we shall refer to the observations $\mathbf{X}_i$ as "functional vectors" .

The chapter is organized as follows. We start with an overview of functional data depth (FDD) in Section 4.2, and propose two new constructions of FDD, with a view to outlier detection when the population is multi-modal. Next we examine the sensitivity of different FDDs to different types of outliers in Section 4.3, and show the potential advantage of centering the curves. In Section 4.4 we discuss three approaches to construct a prediction set with functional data. We end the chapter in Section 4.5 with a discussion of three approaches to two-sample testing with functional data.

Pre-processing of the functional data, such as landmark registration, may be needed to remove contamination (e.g. a slight horizontal shift or stretching of a curve), and to ensure that the inference can be focused on functional features of interest. This is usually case specific; it is beyond the scope of this chapter. One may consult Chapters 3-5 in [45] for a detailed account.

4.2 Functional Data Depth (FDD)

4.2.1 Definition and Overview

We restrict to functions $x: V \mapsto \mathbb{R}$ on an interval $V = [a,b] \subset \mathbb{R}$. If ς is the set of all functions of interest and $\mathscr{P}$ is the set of all probability measures on ς, then a statistical depth functional is a mapping $D: \varsigma \times \mathscr{P} \mapsto \mathbb{R}$. A value $D(x,P)$ is meant to measure centrality of the function x within the distribution P, a higher value indicating that x is closer to the center of P. The 'population depth' $D(x,P)$ will be estimated with the help of a random sample $x_1, \ldots, x_n$ of functions from some $P \in \mathscr{P}$, and then leads to a measure of centrality 'within a sample'.

For a discussion on functional depth when the functions x have more than one argument, we refer to [11].

By discretizing a function x over a grid with d points, it becomes a vector in $\mathbb{R}^d$, to which we can apply any notion of data depth for random vectors. Several fundamental differences between such a functional vector and a general multivariate observation need to be taken into account. Following the summary in [48], we note:

1. A functional vector $\mathbf{X}_i$, when visualized as a curve divides the two-dimensional plane into upper and lower regions. This provides intuition on possible definitions of a "central" curve in a sample. For instance, roughly 50% of the sample curves should lay entirely or mostly in the upper or lower region defined by the "central" curve.

2. The notion of *centrality* (or *outlyingness*) of a curve should not solely focus on addressing its *location* relative to the population, but also on its *shape*. The vertical location of a curve x over support $[a,b]$ is commonly defined as

$$\frac{1}{b-a}\int_a^b x(t)\,dt. \tag{4.1}$$

   For depicting a functional pattern there are many possibilities.

3. A notion of centrality of a curve can involve a trade-off between global aspects across its full support V and local aspects in the neighborhood of points in V.
4. Not all desirable properties of multivariate data depth (see [72]) are meaningful for functional depth, such as full affine invariance. See [38], [20], [48] for a discussion. A list of desirable properties of functional data depth was given in [38] and further refined in [20].

The prevalent functional depths may be classified into:

(i) Integrated depth (ID), proposed by [16],
(ii) Projection-based depth (PBD), proposed in [71] and in [12],
(iii) Geometrical depths, including band depth (BD) from [32], half-region depth (HRD) from [33] and spatial depth (SD) for functional data proposed in [10], [9] and [50].

Consistency of the sample estimators of these depths and other desired statistical properties have been proven by [16] for ID, [71] for PBD, [19] for BD and HRD, and [9] and [48] for SD. See [20] for a comprehensive review.

Some multivariate depths are degenerate in the functional case, that is, the resulting $D(x,P)$ is the same for any x in the set ς (see [9] and [20]). This problem is known to burden BD and HRD when the observed curves are very wiggly and hence cross each other frequently. In practice, the degeneracy issue may be avoided by using the depths of the smoothed functions, instead of the raw "fuzzy" curves. Such pre-processing is best left to the specific needs in the application context. To alleviate this problem, modified versions MBD and MHRD were proposed. However, MBD and MHRD are not good in detecting a spiky outlier curve (see Section 4.3 for illustration).

Besides being useful for the purposes described in the next sections, functional data depth can be applied (a) to define a robust estimate of a typical (or "center") pattern, using the median or a trimmed average ([16]); (b) to build a confidence band around the estimated "center" curve ([8], [7]); (c) to define surrogate responses in a classification problem ([12], [44]).

4.2.2 ID and WID

Given a univariate depth $D\left(x(t),F_t\right)$ of a functional value $x(t)$ relative to its marginal distribution F_t, one can form an *integrated depth* (ID) as $ID=\int_a^b D\left(x(t),F_t\right)\,dt$. The univariate depth can take any form, such as Tukey depth or simplicial depth. For arithmetic simplicity [16] proposed the median

depth $D\left(x(t),F_t\right) = 1 - \left|1/2 - F_t\left(x(t)\right)\right|$, where F_t also denotes the marginal cumulative distribution function, which results in

$$ID(x,F) = b - a - \int_a^b \left|\frac{1}{2} - F_t\left(x(t)\right)\right| dt.$$

In the sample version the population distribution function F_t is replaced by the empirical distribution function of the observed data points $x_1(t),\ldots,x_n(t)$ at t. The function *depth.FM* from the R package ***fda.usc*** will be used to calculate this integrated depth in later sections.

In our practical situation of interest, the population usually is a mixture involving multiple stereotypical functions. We propose a measure of functional depth, weighted integrated depth (WID), that is appropriate for this situation. A useful population model is a mixture of Gaussian processes (GPs), defined structurally as

$$\begin{aligned} X(t) &= \mu + Z(t), \qquad t \in [a,b], \\ \mu &\sim N(m_\mu, \varepsilon_\mu^2), \\ Z|\,c &\sim GP_c(m_c, k_{\theta_c}), \\ c &\sim (w_c)_{c=1,\ldots C}, \\ \kappa_{\theta_c}(s,t) = Cov\left(Z(s),Z(t)|\,c\right) &= l_c^2 \exp\left(-\frac{(s-t)^2}{\tau_c^2}\right) + \varepsilon_c^2 \delta_{s,t}, \end{aligned} \tag{4.2}$$

where the scalar μ reflects the overall location of the curves, and the functions m_c, for $c = 1,\ldots,C$, are the different stereotypical curves after centering. In the model (4.2) the functions m_c are the mean functions of Gaussian processes, which are further specified by their covariance functions k_{θ_c}. Given a latent variable c with distribution (w_c), the process Z is distributed as a Gaussian process with parameters (m_c, k_{θ_c}). The covariance functions are specified by vectors of three positive numbers $\theta_c = (l_c, \tau_c, \varepsilon_c)$, and are the sum of the Gaussian kernel, with spread and smoothness determined by l_c and τ_c, and a white noise function with variance ε_c^2, reflecting observational noise. (The latter process could be restricted to the finite grid at which the processes are observed.) The notation $\delta_{s,t}$ is used for the Kronecker delta, i.e. $\delta_{s,t} = 1$ if $s = t$ and $\delta_{s,t} = 0$ otherwise. The Gaussian kernel is chosen for illustration, and because the resulting Gaussian process can adapt to any smoothness by proper choice of τ_c. The distribution of Z is a mixture $F = \sum_{c=1}^C w_c GP_c(m_c, k_{\theta_c})$ of Gaussian distributions.

We define the WID of a *centered* function x (i.e. a function x with $\int_a^b x(t)\,dt/(b-a) = 0$) as

$$WID(x,F) = \sum_{c=1}^{C} w_c P_{Z_c \sim GP_c}\left(O(x, m_c, k_{\theta_c}) \le O(Z_c, m_c, k_{\theta_c})\right), \tag{4.3}$$

where $O(x, m_c, k_{\theta_c})$ is a weighted square distance of x relative to m_c:

$$O(x, m_c, k_{\theta_c}) = \int_a^b \frac{\big(x(t) - m_c(t)\big)^2}{\kappa_{\theta_c}(t,t)}\, dt. \tag{4.4}$$

The weights in (4.4) are chosen so that the distances $O(x, m_c, k_{\theta_c})$ are on the same scale across the component Gaussian processes. The square distance (4.4) will be larger if x is further from the mean curve of the c^{th} cluster, and hence the corresponding probability in the definition of $WID(x,F)$ will be larger if x is closer to m_c. All probabilities in (4.3) will be small if x is far from every m_c. Thus $WID(x,F)$ measures data depth of x relative to the mixture distribution.

In practice F will be unknown and replaced by an estimator $F_n = \sum_{c=1}^{C} \hat{w}_c GP_{c,n}$, based on a random sample $\{x_i\}_{i=1}^n$ from the distribution of X in (4.2). This involves estimators $\{\hat{w}_c\}_{c=1}^C$, $\{\hat{m}_c\}_{c=1}^C$ and $\{\hat{\theta}_c\}_{c=1}^C$ of the parameters w_c, m_c and θ_c. A rough, but computationally efficient, method of estimation could be to cluster first the curves $x_1, \ldots, x_n$, after centering, next let the weights $\hat{w}_c$ be proportional to the sizes of the clusters, and estimate the remaining parameters by fitting the Gaussian process model to the curves in the corresponding cluster. As a final note, a pruning procedure may be needed to remove very small clusters. Centering of a curve x_i can be done by subtracting the trapezium-rule approximation $v_i = 0.5 \sum_{j=1}^{d-1} (t_{j+1} - t_j)\big(x_i(t_j) + x_i(t_{j+1})\big)/(b-a)$ to its weighted integral. The numbers $\{v_i\}_{i=1}^n$ may be ranked separately using a univariate depth measure. In practice we may observe curves of similar patterns but with varying vertical locations, and decomposing depth in a measure for curve pattern and vertical height will provide more insight.

The probabilities in (4.3) may be approximated by simulating large samples $\{x_b^*\}_{b=1}^B$ from the estimated Gaussian process $GP_{c,n}$ and computing

$$B^{-1} \sum_{b=1}^{B} \mathbf{1}_{\{O(x, \hat{m}_c, k_{\hat{\theta}_c}) \le O(x_b^*, \hat{m}_c, k_{\hat{\theta}_c})\}}.$$

If the chosen estimators $\hat{w}_c$, $\hat{m}_c$ and $\hat{\theta}_c$ converge to their population counterparts, then the corresponding estimator of $WID(x,F)$ will also be asymptotically consistent.

4.2.3 PBD, RPD and MPD

Projection-based depth (PBD) defines the depth of a vector as the smallest univariate depth of its one-dimensional projections. It is developed for multivariate data in [71]. The outlyingness of $x \in \mathbb{R}^d$ relative to a distribution F on $\mathbb{R}^d$ is defined as

$$O(x,F) = \sup_{\|u\|=1} \frac{|u'x - \mu(F_u)|}{\sigma(F_u)},$$

where F_u is the univariate distribution of $u'X$ if $X \sim F$, for a given $u \in \mathbb{R}^d$, and $(\mu(F_u), \sigma(F_u))$ are any pair of location and scale functionals of F_u, e.g. (mean, standard deviation) or (median, median

absolute deviation). (If $\sigma(F_u) = 0$, the quotient is read as 0.) The population depth of x is then defined to be $1/\big(1+O(x,F)\big)$. For a sample version, the distributions F_u are replaced by their empirical counterparts.

Multivariate projection-based depth was generalized to functional data by [12]. The idea is to replace the vectors u by functions and $x'u$ by the integrals $\int_a^b x(t)u(t)\,dt$. Because there are too many possible choices for function u, [12] proposed a random projection depth (RPD) based on functions $u_1,\ldots,u_B$ generated from a Gaussian distribution and standardized to norm 1, given by

$$RPD(x,F) = \frac{1}{1+O(x,P)}, \tag{4.5}$$

where

$$O(x,F) = \frac{1}{B}\sum_{u\in\{u_1,\ldots,u_B\}}\left|\frac{\int_a^b u(t)x(t)\,dt - \mu(F_u)}{\sigma(F_u)}\right|, \tag{4.6}$$

and where F_u denotes the distribution of $\int_a^b u(t)X(t)\,dt$ for $X\sim F$, for a given u, with location and scale $\mu(F_u)$ and $\sigma(F_u)$. (The functions u could be normed, but this will be irrelevant if the μ and σ are scale invariant.) The properties of this RPD will not only depend on the choice of (μ,σ), but also on B and the assumed stochastic process for u. In practice the functions will only be evaluated on a grid and the integrals will be approximated by sums. The function *depth.RP* from the R package ***fda.usc*** will be used to calculate RPD.

We investigate an alternative, which is relevant under the population model given by the mixture of Gaussian processes as in (4.2). Rather than the random directions we use the mean functions m_c of the components, leading to

$$MPD(x,P) = \left(1+\max_c w_c\left|\frac{\int_a^b m_c(t)x(t)\,dt - \mu(F_{m_c})}{\sigma(F_{m_c})}\right|\right)^{-1}. \tag{4.7}$$

Here F_{m_c} is the distribution of $\int_a^b m_c(t)X(t)\,dt$ if $X\sim GP_c$. We choose (μ,σ) to be the median and median absolute deviation to avoid the impact of outliers.

In the sample version, we first estimate w_c and m_c from data $\{x_i(t)\colon t=t_1,\cdots,t_d\}_{i=1}^n$, by the same procedure as described for for WID. Furthermore, we evaluate $\mu(F_{m_c})$ and $\sigma(F_{m_c})$ based on large samples from GP_c.

4.2.4 BD and MBD

For an arbitrary subset $I\subset\{1,\ldots,n\}$, we define the 'band' formed by the set of sample functions $x_I=\{x_i\colon i\in I\}$ as

$$B(x_I) = \{(t,y)\colon t\in V,\ \min_{k\in I}(x_k(t)) \leqslant y \leqslant \max_{k\in I}(x_k(t))\}.$$

Next for $G(x) = \Big\{ \big(t, x(t)\big) : t \in [a,b] \Big\}$ the graph of an arbitrary function x, define the sample band depth (BD) of x relative to $x_1, \ldots, x_n$ as

$$BD_{n,J}(x) = \sum_{j=2}^{J} \left[\frac{1}{\binom{n}{j}} \sum_{I \subset \{1,\ldots,n\};\, |I|=j} \mathbf{1}_{\{G(x) \subseteq B(x_I)\}} \right]. \tag{4.8}$$

Here J is a fixed integer with $2 \leqslant J \leqslant n$, which is commonly recommended to be chosen equal to $J = 3$, in view of the computational burden.

A modified version (MBD) is obtained by not requiring that the full graph be contained in the band, but instead measure the length of the interval on which x falls within the band, i.e. the length of the set

$$A(x_I) = \{t \in V \colon \min_{k \in I}(x_k(t)) \leqslant x(t) \leqslant \max_{k \in I}(x_k(t))\}.$$

Next we replace the indicator function $\mathbf{1}_{(\cdot)}$ by the Lebesgue measure of $A(x_I)$, and define

$$MBD_{n,J}(x) = \sum_{j=2}^{J} \left[\frac{1}{\binom{n}{j}} \sum_{I \subset \{1,\ldots,n\};\, |I|=j} \frac{\lambda(A(x_I))}{\lambda(V)} \right]. \tag{4.9}$$

The functions *BD* and *MBD* from the R package ***roahd*** will be used to calculate BD and MBD in later analyses.

4.2.5 HRD and MHRD

The half-region depth measures the proportion of sample curves that are above or below a given function. Define the hypograph (hyp) and epigraph (epi) of a function x with support V as

$$hyp(x) = \{(t,y) \in V \times \mathbb{R} \colon y \leqslant x(t)\},$$
$$epi(x) = \{(t,y) \in V \times \mathbb{R} \colon y > x(t)\}.$$

Then for $G(x_i)$ the graph of the sample function x_i, the sample half-region depth (HRD) of x is defined as

$$HRD_n(x) = \min\{K_1(x), K_2(x)\}, \tag{4.10}$$

where

$$K_1(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{G(x_i) \subset hyp(x)\}},$$
$$K_2(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{G(x_i) \subset epi(x)\}},$$

are the proportions of sample values with graph fully below or above x.

This depth measure can be modified by not requiring that a sample function is everywhere below of above x, but by measuring the length of the sets where this is the case, in the same spirit as MBD relaxes BD. we define MHRD as

$$MHRD_n(x) = \min\{K_1^*(x), K_2^*(x)\}, \tag{4.11}$$

where

$$K_1^*(x) = \frac{1}{n\lambda(V)} \sum_{i=1}^{n} \lambda\{t \in V\colon x(t) \leqslant x_i(t)\},$$
$$K_2^*(x) = \frac{1}{n\lambda(V)} \sum_{i=1}^{n} \lambda\{t \in V\colon x(t) > x_i(t)\}.$$

The functions *HRD* and *MHRD* from the R package ***roahd*** will be used to calculate HRD and MHRD in later analyses.

4.3 Topic I - Visualization of Outliers

Given a sample of curves, a future curve can be viewed as an outlier if (a) it is at an abnormal location, e.g. the entire curve lays above or below the majority of the sample curves, (b) its pattern is not similar to any of the sample curves, or (c) a combination of (a) and (b). Additionally, a deviation in pattern in the sense of (b) may be global or a local. In this section, we use a simulated dataset to examine the sensitivity of the different FDDs with respect to outlier functions of different designs, and illustrate the potential advantage to monitor outliers via the depth values of curves after centering. A bivariate plot is presented as a useful tool to separate outlyingness due to (a) or (b).

We consider a population that is a mixture of two stochastic components (M1) and (M2), specified below, with weights p_{M1} and $p_{M2} = 1 - p_{M1}$:

- (M1) $X(t) = \mu + 4t + \varepsilon(t)$, where $\mu \sim N(0, 0.9^2)$ and ε is a centered Gaussian process with covariance kernel $\kappa(s,t) = 0.1\exp\left(-(t-s)^2/0.07\right)$, for $0 \leq s,t \leq 1$.
- (M2) $X(t) = \mu + 0.9\sin(3\pi t) + 4t + \varepsilon(t)$, where $\mu \sim N(0, 0.4^2)$ and ε is a centered Gaussian process with covariance kernel $\kappa(s,t) = 0.05\exp\left(-(t-s)^2/0.02\right)$, for $0 \leq s,t \leq 1$.

The curves are observed at 50 equally spaced points in $[0,1]$.

We design the following three outlier curves:

- (Outlier-1: Global pattern deviation): a sample curve from $X(t) = -1.5 + 7t + \varepsilon(t)$, where ε is the same as (M1).
- (Outlier-2: Sharp local spike): the sum of a sample curve from (M1) and the function that is 4.5 when $t \in (0.7, 0.75)$ and zero otherwise.

- (Outlier-3: Modest local spike): the sum of a sample curve from (M1) and the function that is 2.5 when $t \in (0.7, 0.75)$ and zero otherwise. In addition, the curve location is chosen to be relatively low so that the modest spike is among the sample curves.

We simulated 250 curves from (M1), and 250 curves from (M2) and added the three outlier curves, giving a dataset of 503 curves in total. The 503 curves are shown in the top left panel of Figure 4.1. The top right panel gives the same curves but centered vertically by subtracting the constant

$$FL(x) = \frac{1}{d}\sum_{j=1}^{d} x(t_j) \tag{4.12}$$

from each curve x. These are estimates for the 'functional locations' $\int_0^1 x(t)\,dt$ of the curves.

We calculated the depth of all 503 individual curves within the total sample using every of the methods ID, WID, BD, MBD, HRD, MHRD, RPD and MPD. To facilitate visual comparison, we rescaled the depth values to the interval $[0,1]$, by substracting per method the minimum depth over the 503 curves and dividing by the range of the depth values. These rescaled values are plotted in the bottom left panel of Figure 4.1, while the bottom right panel shows the same but now for the centered curves. The depth values of WID are absent in the left plot, since this method is designed for centered curves, while BD and HRD are absent in the right plot in view of the degeneracy of these methods for the original curves (the curves cross each other too often).

We may expect that the FL-centered curves are more suitable for discovering functional patterns. Besides, we found that clustering the centered curves lead to more accurate recovery of the two clusters in the data (needed for WID and MPD calculation).

Details of the calculations are as follows. For WID and MPD we clustered the curves by the K-means method, with $K = 2$, the correct number of clusters in this case. In practice, we would determine a suitable value of K by a preliminary analysis. Given the clustering the center curve per cluster was estimated by pointwise $\alpha-$trimmed means with $\alpha = 0.05$, to avoid the influence of the extreme values. For WID, we set the value of $\kappa_{\theta_c}(t,t) + \varepsilon_\mu^2$ in (4.4) for every t equal to the pointwise variance of the functions in a cluster averaged over the 50 grid points. We used here the fact that in the present data the variance is constant in t. In practice we would make full inference on the parameters of the Gaussian processes.

Ideally the depth values of the three outlying curves would be smallest among all depth values. In the left bottom panel of Figure 4.1, we see that only BD is sensitive enough to detect all the three outliers in this way. This comes at a high computational burden as the number of the curves increases. The HRD method seems to be sensitive to a sharp local spike, while the other FDDs all perform poorly overall.

Centering the curves appears to enhance the sensitivity to outliers for all FDDs. All FDDs except MHRD and RPD can detect the global pattern deviation. MBD and MPD are not sensitive to the

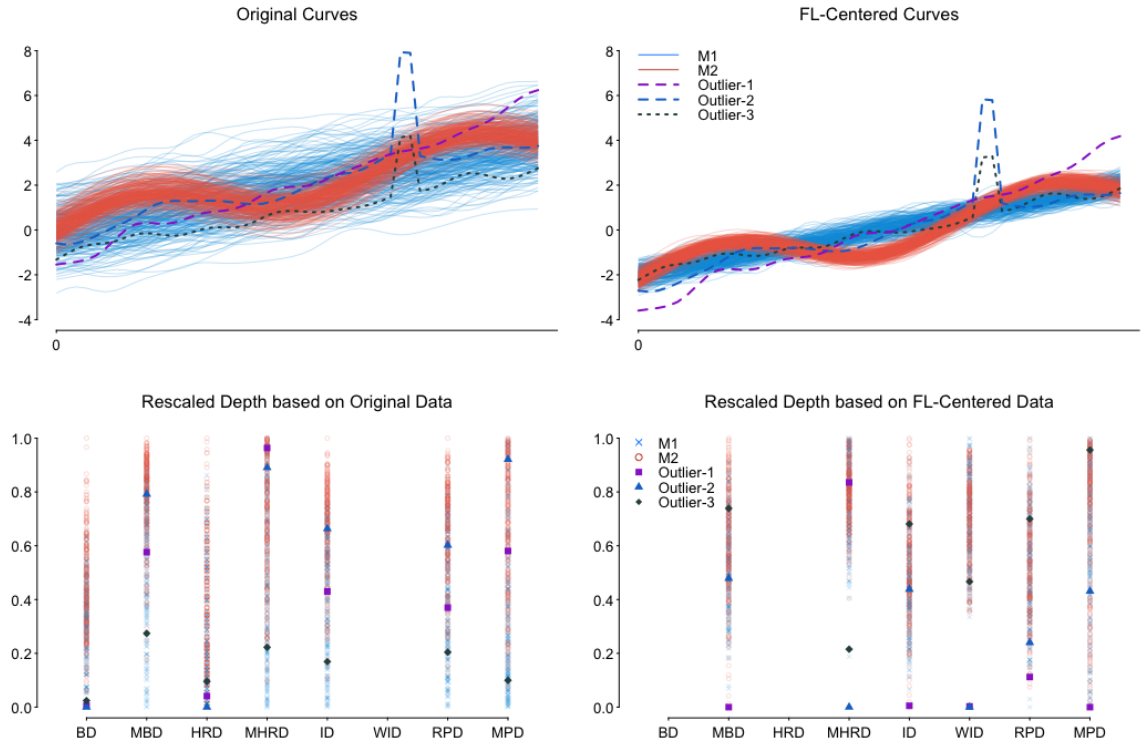

Fig. 4.1 Top: the complete set of 503 simulated curves (top left) and the corresponding centered curves (top right): 250 sampled from (M1), 250 from (M2) and 3 outlier curves. Bottom: scaled FDDs of eight methods based on the original curves (bottom left) and on the centered curves (bottom right). Outlier curves are marked with different colors, line and point types.

local spike, as can be expected from their design through integrating a deviation metric over the full support, while WID has a higher sensitivity by integrating a squared deviation metric. MHRD is good at detecting a local spike, even if this is of modest strength. RPD fails in assigning any of the three outliers the lowest depth value.

The advantage of using the centered curves is clear in the preceding. However, centering does neglect the vertical location of the curve, which itself can be a source of the outlyingness. To remedy this one can calculate the univariate depths of the locations $FL(x)$ of the curves. In Figure 4.2 these are plotted against the corresponding FDD values based on the centered curves. If a future curve shows up at the left side of this bivariate plot, then it is likely a shape outlier, while a location outlier will show up at the bottom of the point cloud. Points in the bottom left corner are outlying relative to both aspects.

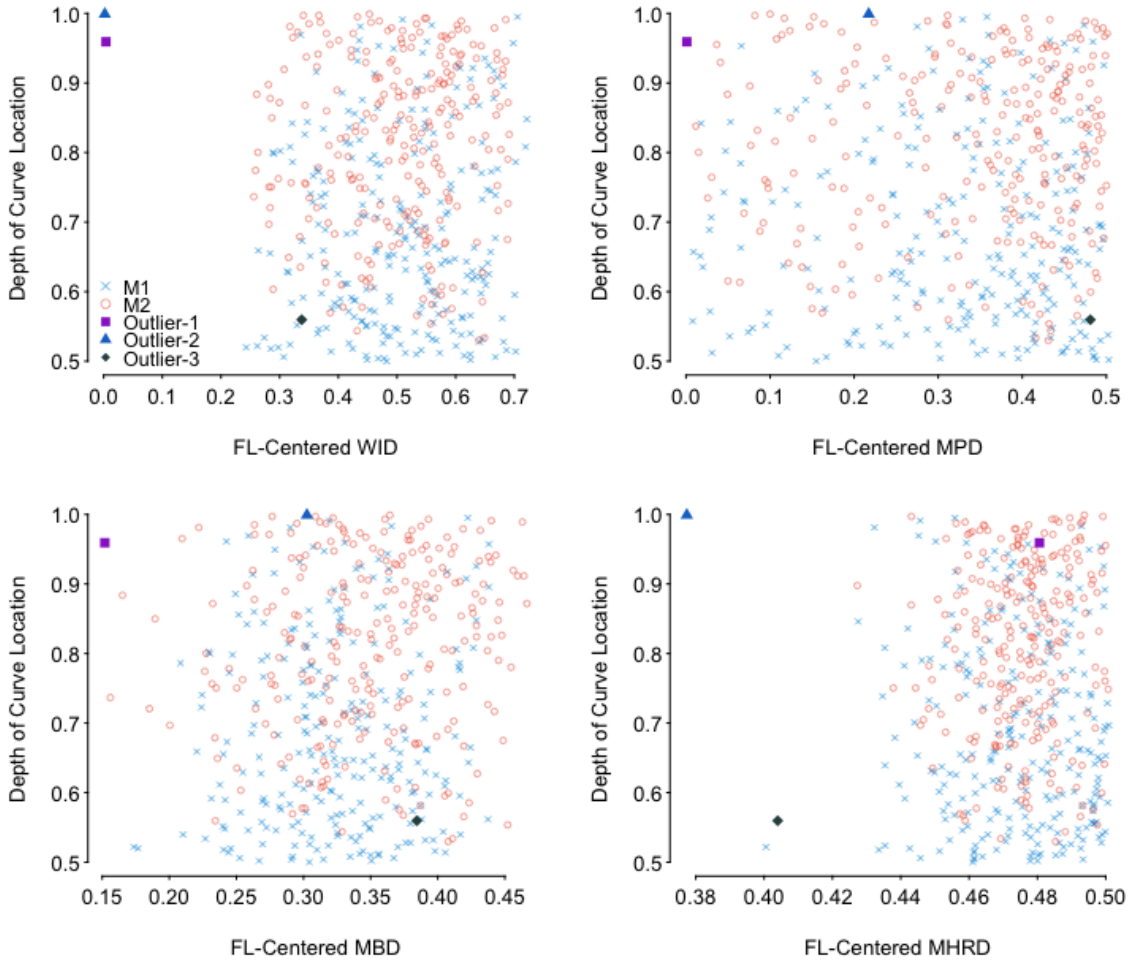

Fig. 4.2 Plot of univariate depth of curve locations versus FDD value of the centered curves. The four panels show the four types of FDD: WID, MPD, MBD and MHRD. In every case the univariate depth is the median depth $1 - \left|1/2 - F_n\left(FL(x)\right)\right|$, where F_n is the empirical cumulative distribution function of the sample of curve locations $LF(x_i)$.

4.4 Topic II - Prediction Sets

In this subsection, we discuss two approaches to build a prediction set for functional data using depth values. In our practice, constructing a prediction set is needed for monitoring a future abnormal curve or assessing comparability of two groups of curves. In the latter application, one may be requested to compare a test group with only 3–5 individuals to a large reference group. A rigorous statistical test might then lack power, but we could build a prediction set based on the reference data, and conclude that the groups are "comparable" if the samples from the test group are within this prediction set.

Practitioners often build a prediction set for functional data by combining pointwise prediction intervals into a prediction band for a curve. A future curve is considered abnormal if any point of its graph falls outside the band. Although this has an intuitive visual interpretation, a drawback is that the coverage of the band may be much smaller than the coverage of the individual intervals, and this may be hard to correct. In addition, this approach neglects curves with abnormal patterns, as long as their values are within the band.

An alternative is to build a prediction set using functional depth values. A first possibility is to use the depth values directly, and define a prediction set as the set of curves with depth higher than some threshold. This is reasonable, since FDD ranks curves in a center-outwards manner, where a higher depth value indicates that a curve is closer to the "center" (and hence more "typical"). Given a FDD, it suffices to determine the threshold to ensure a desired coverage. The sensitivity of such a prediction set to different types of abnormality depends on the chosen FDD and how it is calculated (based on the original curves or the centered curves). For instance, in Section 4.3 Outlier-2 and 3 were seen to be within the 95%, or even 90%, prediction sets for most choice of data-depth (see Figure 4.1).

Another approach is to project the curves to a set of basis functions, and to build a prediction set based on the multivariate depth measure of the projection coefficients. We shall use the (functional) eigenbasis of the reference sample, and likelihood depth based on a Bayesian density estimator for the distribution of the coefficient vectors, as described in Section 3.6, to build a prediction set. Since an outlier curve may be badly approximated by the eigenfunctions, but still have similar projection coefficients, we also perform a preliminary test on the distance between the curve and its projection. If this distance is larger for the new curve than for any of the reference curves, or exceeds a large quantile of the latter distances, than the new curve is classified as an outlier. By using a very large quantile, this test will be stringent, and the preliminary check will hardly change the coverage of the prediction set.

The eigenfunctions $\xi_1, \xi_2, \ldots$, or 'functional principal components', of a sample $x_1, \ldots, x_n$ of curves sequentially maximize the sum of squares $\sum_i v_{il}^2$ of the loadings $v_{il} = \int_a^b \left(x_i(t) - \bar{x}(t)\right) \xi_l(t)\, dt$ of the centered x_i on ξ_l, under the constraints that $\int_a^b \xi_l(t)^2\, dt = 1$ and $\int_a^b \xi_l(t)\xi_j(t)\, dt = 0$, for $l > j$.

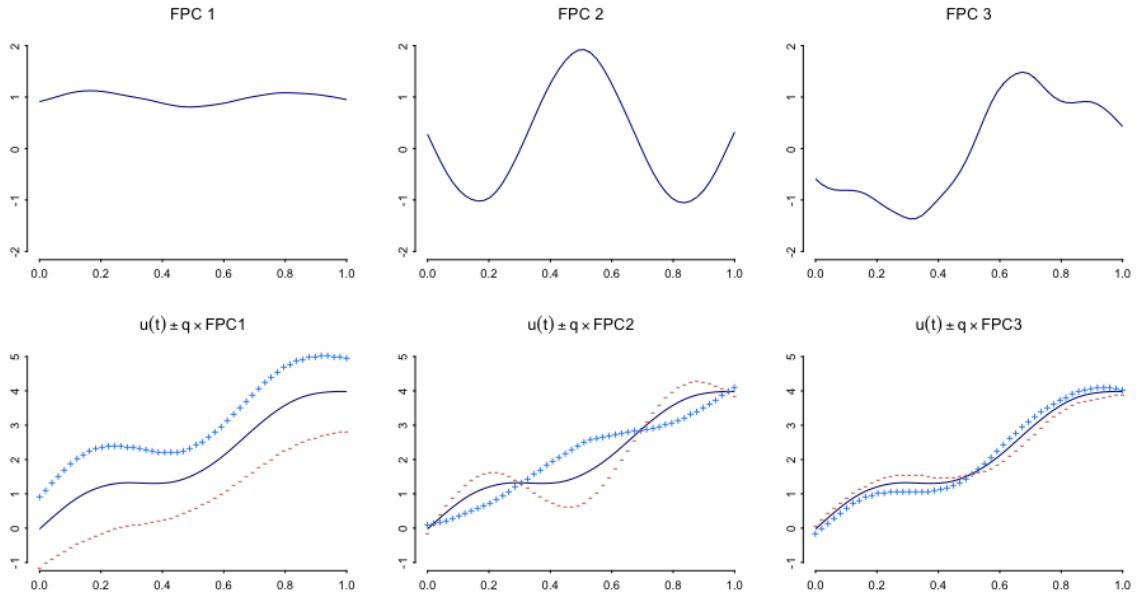

Fig. 4.3 Top row: first three eigen-functions ξ_1, ξ_2, ξ_3 corresponding to the centered set of 500 reference curves. Bottom row: average curve $u(t) = \bar{x}(t)$ (solid line), and average curve $+q_{upp} \times \xi_l(t)$ (line made of '+') and $-q_{low} \times \xi_i(t)$ (line made of '-'), where q_{upp} is the 0.95 quantile of the coefficients $\{v_{il}\}_{i=1}^n$ and q_{low} is the 0.05 quantile, for $l \in \{1,2,3\}$.

Given the first L eigen-functions, a sample curve x_i can be approximated by

$$\hat{x}_i(t) = \bar{x}(t) + \sum_{l=1}^{L} v_{il}\xi_l(t) \tag{4.13}$$

A new function can be similarly projected onto the span of $\xi_1, \ldots, \xi_L$. In practice the functions will be discretized on a grid and the integrals will be approximated by sums. The number L of basis functions will typically be chosen much smaller than the number of grid points. Details on functional principal component analysis can be found in Chapter 8 of [45]. We use the R package $\{fda.usc, \text{version } 2.0.1\}$ for FPCA in this study.

We implemented this method on the dataset generated in Section 4.3, consisting of a reference sample of 500 curves and three outlying curves, visualized in the top-left panel of Figure 4.1. We decided to retain the first three basis functions, which explained 94% of the total variation in the 500 reference curves. They are shown in the top panels of Figure 4.3, and superimposed as deviations on the mean function $u = \bar{x}$ in the bottom panels of the same figure.

The coefficients of the three outlying curves x_{n+1} were calculated as $v_{n+1,l} = \int_a^b (x_{n+1}(t) - \bar{x}(t))\xi_l(t)\,dt$, and their projections similarly to (4.13) (shown in Figure 4.4). We calculated the square distances between all x_i and their projections $\hat{x}_i$ as $\Delta_i = \int_a^b \left(x_i(t) - \hat{x}_i(t)\right)^2 dt$. For the reference sample

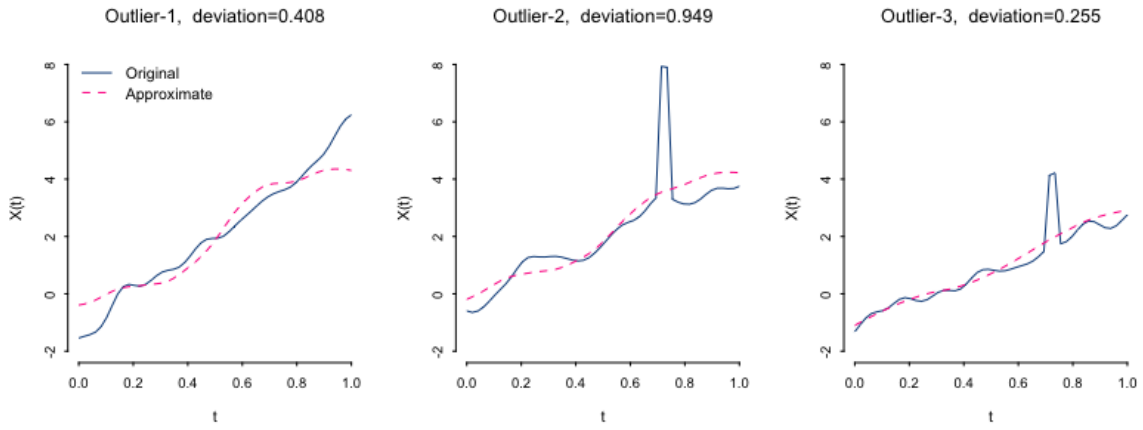

Fig. 4.4 The three outlier curves (solid lines) and their approximation based on $\{\xi_l\}_{l=1}^3$ (visualized in Figure 4.3). Deviations Δ between the observed and the approximated are shown in the title.

these ranged from 0.002–0.184, while for the three outlying curves these values were 0.408, 0.949 and 0.255, all exceeding 0.184.

We computed the likelihood data depth based on fitting a Bayesian nonparametric density estimator to the vectors $\{(v_{i1}, v_{i2}, v_{i3})\}_{i=1}^n$ of coefficients relative to the first three eigenfunctions. This density estimator assumes a Dirichlet process mixture of Gaussians as a prior. We used the SUGS algorithm, described in Section 3.6, to approximate the posterior mean density, with the prior hyper-parameters set to: $\kappa_0 = 0.1$, $\nu_0 = 7$, and μ_0 and Λ_0 set to be the sample mean and the inverse of the sample covariance matrix of $\{(v_{i1}, v_{i2}, v_{i3})\}_{i=1}^n$. We used 50 random permutations of the input sequence for SUGS, and selected the sequence with the highest predictive likelihood to perform the final fit. Given a sequence of input data, the predictive likelihood is the log-likelihood of the last 20% of the data based on the density estimator trained on the first 80% of the data via SUGS. The prediction set was then set equal to functions with a likelihood depth greater than the threshold set to achieve Bayesian $1-\delta-$expectation coverage, as in (3.12). This threshold was 1.8×10^{-5} for $\delta = 0.01$ (99% coverage), and 6.5×10^{-5} for $\delta = 0.05$. The likelihood depths of the three outlying curves were 2.9×10^{-11} for Outlier-1 (global deviation), 2.0×10^{-6} for Outlier-2 (sharp local spike), and 6.8×10^{-4} for Outlier 3 (modest local spike). Hence only the Outlier-3 sits in the prediction set.

We confirmed the coverage of the prediction set in a simulation study, using the mixture of Gaussian curves (M1) and (M2) with weights $p_{M_1} = 0.5$, described in Section 4.3, as population model. We generated a validation set of 5000 curves from this population, and checked inclusion in the prediction sets computed from 100 different samples of 500 curves. For each of the 100 samples we computed the first three eigenfunctions, the likelihood depth of the 500 coefficient vectors, and the lower thresholds of the prediction sets for 90%, 95% and 99% coverage. We found that proportions of 0.9019, 0.9486 and 0.9889 of the 5000 validation curves belonged to the prediction sets, all of which are very close to their nominal values.

In all calculationss curves were discretized to a grid of 50 equally spaced nodes in $[0,1]$. The first three eigenfunctions explained between 93-95% of the total variation across the 100 simulated datasets.

4.5 Topic III - Two Functional Sample Testing

Assume given a reference group $\{x_i\}_{i=1}^{n_x}$ of n_x curves and a test group $\{y_j\}_{j=1}^{n_y}$ of n_y curves, asssumed to be $i.i.d.$ samples from distributions with mean curves μ_X and μ_Y, respectively. An ubiquitous question in practice is whether the two distributions are equal or not. Here the curves may be preprocessed, for instance centered, if it is desired to focus on a difference of particular patterns between the two groups.

We discuss three approaches to conduct two-sample testing with functional data: a rank test from [32], a permutation test, and a bootstrap test. We illustrate the tests on growth curves of groups of boys and girls, originally from [45], used in [32], and compare the results to [32]. The motivation for the new tests is that the rank test proposed in [32] requires to split the reference sample in a seemingly arbitrary way.

4.5.1 Rank Test

The rank test proposed by [32] starts by randomly dividing the reference group into two parts, a 'baseline' sample $Z = \{z_k\}_{k=1}^{n_0} \subset \{x_i\}_{i=1}^{n_x}$ and a 'reference' sample consisting of the remaining $n_x - n_0$ curves x_i. It is advised to choose n_0 not too small, for instance $n_0 = \max(n_x/2, n_y)$.

For a given functional depth measure $D(x,P)$ and P_{n_0} the empirical distribution of Z, a score of a function x is defined as $\sum_k \mathbf{1}_{\{D_n(z_k,P_{n_0}) \leq D_n(x,P_{n_0})\}}$. The scores of all functions in the pooled sample of $n_x - n_0$ remaining 'non-baseline' functions x_i and the n_y functions y_j are computed, ordered, and $R(x_i)$ and $R(y_j)$ are set to be their ranks in this pooled sample. If there are ties among the scores, the rank is defined as the midpoint of the unadjusted ranks, as usual. The proposed test statistic W is the sum of the ranks $\{R(y_j)\}_{j=1}^{n_y}$ of the test group.

The scores are a measure of centrality relative to the distribution P_{n_0}, which should be close to the distribution of the x_i. Thus we may expect that the scores in the test group will be smaller, and hence their ranks in the pooled sample to be smaller as well. Thus the null hypothesis that there is no difference between the reference and test group is rejected for small values of W.

Under the null hypothesis that there is no difference, the ranks in the pooled sample should be a uniform permutation of the numbers $\{1,2,\cdots,n_x+n_y-n_0\}$, adjusted for ties if there are any. Thus the test statistic W has the same null distribution as the sum of a random sample of size n_y without replacement drawn from $\{1,\cdots,n_x+n_y-n_0\}$, adjusted for ties if there are any.

4.5.2 Permutation Test

Given any two-sample test statistic, a permutation test compares the value of the statistic on the two observed samples to the values on all pairs of samples obtained by splitting the pooled sample $x_1,\ldots,x_{n_x},y_1,\ldots,y_{n_y}$ arbitrarily in groups of sizes n_x and n_y. If the sample sizes are large, one often uses a randomly chosen set of splits rather than all splits, together with the original split. Because under the null hypothesis all splits are equally likely to give the original grouping, the P-value of the permutation test can be computed as the proportion of splits giving a larger value of the test statistic than the observed value.

If the test statistic is a distance $\Delta(\hat{\mu}_X,\hat{\mu}_Y)$ between estimated mean functions in the two samples, then this P-value becomes $\sum_{b=1}^{B}\mathbf{1}_{\{\Delta(\hat{\mu}_X^{(b)*},\hat{\mu}_Y^{(b)*})\geqslant\Delta(\hat{\mu}_X,\hat{\mu}_Y)\}}/B$, where b ranges over all splits. The null hypothesis is rejected when this is smaller than the significance level. The number of possible splits is $\binom{n_x+n_y}{n_x}$. If this is too large, then B will be chosen a preset value and a random selection of splits will be used.

An alternative test statistic is the centrality of the mean difference curve $\hat{\mu}_X-\hat{\mu}_Y$ within an estimate of the distribution of this difference under the null hypothesis. For $\{\hat{\mu}_X^{(b)*}-\hat{\mu}_Y^{(b)*}\}_{b=1}^{B}$ the mean difference curves after B reassignments of the group labels, the latter distribution can be estimated as the empirical distribution F_B of these difference curves. The test statistic on the original sample then becomes $D(\hat{\mu}_X-\hat{\mu}_Y,F_{B+1})$ and the P-value is evaluated as $\sum_{b=1}^{B}\mathbf{1}_{\{D(\hat{\mu}_X^{(b)*}-\hat{\mu}_Y^{(b)*},F_B)\leqslant D(\hat{\mu}_X-\hat{\mu}_Y,F_B)\}}/B$.

4.5.3 Bootstrap Test

The bootstrap procedure is the same as the permutation procedure, except that the new samples are now formed by drawing n_x and n_y curves with replacement from the pooled sample of curves.

When the sample sizes n_x and n_y of the two groups are highly unbalanced, some caution is necesssary. For instance, the chance of redrawing all curves from the testing group is $(\frac{n_x}{n_x+n_y})^{n_x+n_y}$, which can be larger than desired if $n_y/(n_x+n_y)$ is extremely small.

4.5.4 Example Analysis

The example dataset, shown in Figure 4.5, consists of the growth curves of 39 boys and 54 girls, In [32] the P-value of the rank testing procedure was found to be 0.0001 with BD, 0.1199 with MBD, and 0.1636 with ID. We applied the permutation and bootstrap test with $B=5000$ and test statistics (1) **L1-Delta** $\Delta(x,y)=\int_a^b\left|x(t)-y(t)\right|dt$, (2) **L2-Delta** $\Delta(x,y)=\int_a^b\left(x(t)-y(t)\right)^2dt$, (3) MBD, or (4) ID. We found that BD was too computationally intensive to be calculated at such large sample size.

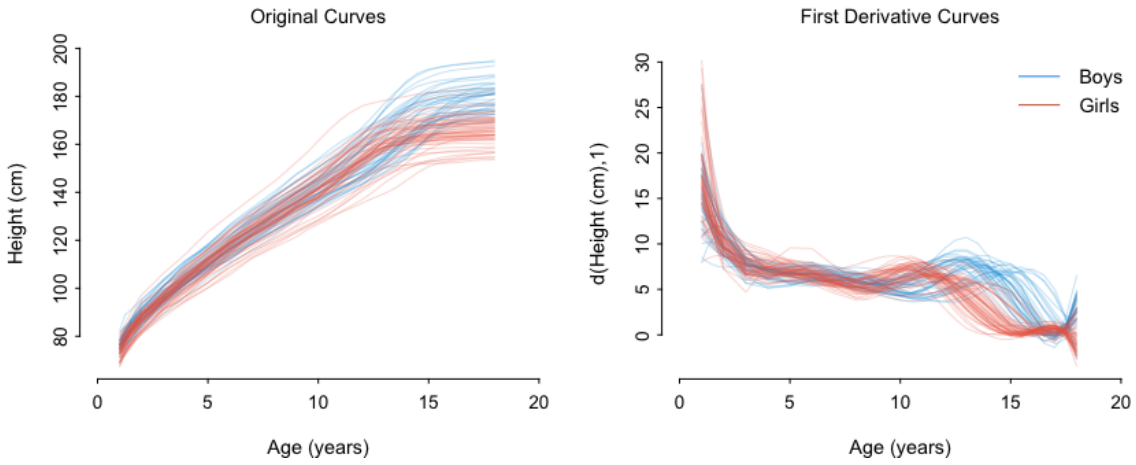

Fig. 4.5 Growth curves (left subplot) of boys (in blue) and girls (in red), and their 1st derivative curves (right subplot).

Table 4.1 P-values of the different test procedures with the growth curve dataset

Setting	**Rank Test**[1]	**Permutation**	**Bootstrap**
L1-Delta	-	0.0010	0.0018
L2-Delta	-	<0.0001	<0.0001
MBD	0.1199	0.0836	0.0744
ID	0.1636	0.1098	0.0984

The results are summarized in Table 4.1. For illustration the redrawn curves $\{\hat{\mu}_X^{(b)*} - \hat{\mu}_Y^{(b)*}\}_{b=1}^{B+1}$ are shown in Figure 4.6.

Figure 4.5 shows that there is a clear difference between the curves of the two groups, with a change in pattern starting around age 12. The pointwise difference curve $\hat{\mu}_{boys} - \hat{\mu}_{girls}$ is also clearly different from the difference curves simulated by the permutation or bootstrap procedure, shown in Figure 4.6. Permutation and Bootstrap tests with L1 and L2-Delta all return very small p-values. However, MBD and ID are not sensitive enough to flag this obvious deviation, due to their global definition, while the deviation is local. Actually, if we perform the two-sample testing based on the first derivative of the original curves (shown in the right subplot of Figure 4.5), all instances of the permutation and the bootstrap tests in Table 4.1 return p-values <0.0001. This is explained by the fact that the first derivative curves of girls starts to show different pattern already around age 9 comparing to boys, while the full range of observation is from age 1 to 18.

As a visual instrument, one may build a prediction band based on the data from the girls, and see starting from what age the growth curves of the boy increase outside the band (see Figure 4.7 for

[1] results are cited from [32]

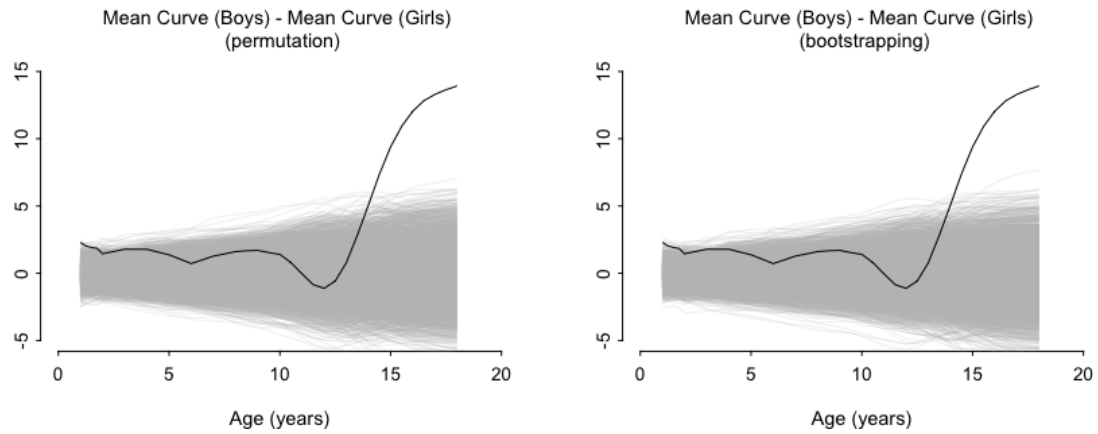

Fig. 4.6 Pointwise difference curve $\hat{\mu}_{boys} - \hat{\mu}_{girls}$ between original mean curves of boys and girls (black line), and 5000 simulated difference mean curve $\hat{\mu}^{(b)*}_{boys} - \hat{\mu}^{(b)*}_{girls}$ with the permutation procedure (grey lines in the left subplot) or the bootstrap procedure (grey lines in the right subplot).

illustration). We constructed such a prediction band by performing FPCA on the observed growth curves $\mathbf{X} = \{x_i(t)\}_{i=1}^{54}$, $t \in [1,18]$ for girls, retaining the first 7 eigen-functions, which explained 99.5% of the total variation. We simulated 1000 vectors v^*_{bl} from the multivariate normal distribution with mean and covariance matrix being the sample mean and covariance matrix of the loading vectors of the 54 girl's curves, and then calculated 1000 curves via $x^*_b(t) = u(t) + \sum_l v^*_{bl}\xi_l(t)$, where u is the mean of the girl's curves and ξ_l are the eigenfunctions. Determine a 95% prediction set S^*_{95} based on $\{v^*_b\}_{b=1}^{1000}$ and their likelihood depths, and denote $B_{95} = \{b\colon\ v^*_b \in S^*_{95}\}$. The representative band is $\{(t,y)\colon t \in [1,18],\ \min_{b\in B_{95}}(x^*_b(t)) \leqslant y \leqslant \max_{b\in B_{95}}(x^*_b(t))\}$.

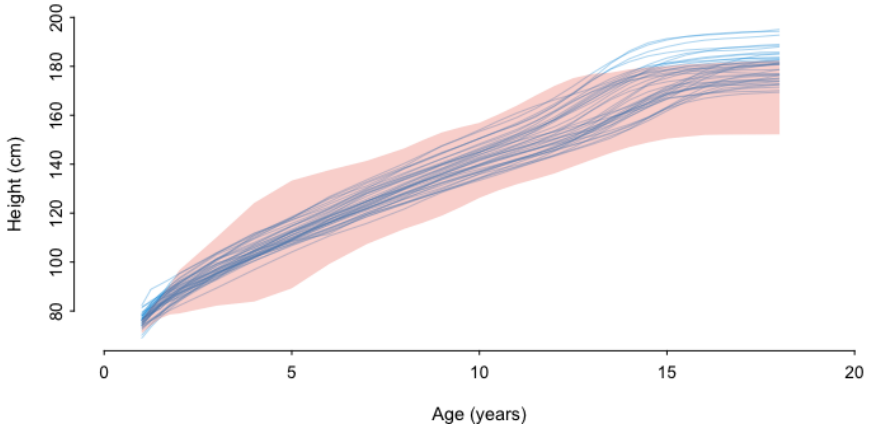

Fig. 4.7 Growth curves of the boys (in blue) and the prediction band (in red) of the girls, the band intended to cover at least 95% of the girl population.