Bootstrap-Based Hypothesis Testing in Functional Data for Assessing Equality of Mean Functions

Liam Bullen[a], Camila de Souza[b]

[a]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada

[b]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada

Abstract

This paper develops a method using bootstrap-based hypothesis testing to evaluate the equality of two mean curves from data sets in functional data. We test multiple statistics and evaluate their performance at identifying equality in mean curves using the power function of each statistic, estimated via the empirical rejection probability. As the mean curves diverge, the best performing statistics are those whose power functions approach one at the fastest rate. The statistics tested are the L_1 and L_2 norms, the Hellinger distance, the affinity, and the Kullback-Leibler divergence. The testing is performed with both simulated data as well as real world examples. The simulated data is generated by adding random error to data points from a chosen function, then using a cubic B-spline basis with a continuous roughness penalty to construct functions from the data points.

Keywords: functional data analysis; bootstrap; hypothesis testing

1 Introduction

Hypothesis testing is vital to be able to infer characteristics of a population based on a sample of observations. Functional data analysis is a field very quickly growing in popularity as a tool for analyzing data sets containing curves. Traditional means of hypothesis testing are not available because of the nature of the data points as functions. Instead, alternative methods must be developed to be able to compare the equality between two sets of functional data. Recently it has been shown that bootstrap-based hypothesis testing is effective in functional data sets for asymptotic approximation(Paparoditis, 2016; Zhang, 2010).

The primary novelty introduced in this paper is the evaluation of a larger number of test statistics than previously considered in the literature for bootstrap-based inference methods. Traditionally past studies have limited themselves to use of the L_1 or L_2 distances as statistics, however we also test the Hellinger distance, the affinity and the Kullback-Leibler Divergence. The Hellinger distance is used to quantify the similarity for density functions. As we work with functions when performing analysis in functional data, the Hellinger distance could outperform statistics such as L_1 and L_2 norms, which were made to compare discrete data points. The Hellinger distance also has some properties that may give advantages. Unlike the L_1 and L_2 norms, the Hellinger is invariant to continuous monotonic functions (Su, 2008). Similarly, the Kullback-Leibler divergence, which is also used to compare probability distributions, has good asymptotic properties when used in density estimation (Dias, 2007).

The test statistics are evaluated based on their power, that is, their ability to reject the null hypothesis. We first simulate samples of functional data and smooth the curves via smoothing splines. We then perform a bootstrap-based hypothesis test to assess the equality of the mean curves using all test statistics, and find the empirical rejection probabilities of tests to validate strongest statistic. The findings are also applied to a real data set.

This paper begins with a review of the relevant literature used to build our methodology and experiment. The methodology of the experiment is then discussed in depth, detailing how we acquire the results. A description of how the data is obtained and the setting of the experiments is outlined. Finally, we show the results of the experiments performed, and how each test statistics performed, as well as a discussion based off of the findings.

2 Literature Review

Hypothesis testing in functional data sets has been a growing topic in the literature as functional data analysis continues to become more relevant as an area of research in statistics, see S. Ramsay (2005), Ferraty (2007). Functional Data Analysis involves all data where the data points take the form of a curve, for example the growth curves of children (R. Silverman, 2002). The earliest work on this topic was done by D. Ramsay (1982) where he analyzed the types of functional data that occur when collecting data. Later a more thorough text was published by Ramsay and R. Silverman (2002) that explores case studies and describes the theory behind Functional Data as well as how to work with this data.

Early efforts at hypothesis testing for functional data such as Horvath (2012) considered testing with functional principal components for the equality of two mean functions. Other techniques were also explored by Cuevas (2003) using ANOVA. More recently, however, the literature has increasingly turned to bootstrap-based procedures for hypothesis testing in FDA, to improve the asymptotic approximations of the tests. Zhang (2010) performed a bootstrap-based test on the equality of two means that generated successful results. Crainiceanu (2012) performed multiple types of bootstrap methods with parametric and non-parametric estimations evaluated with different levels of smoothing to compare correlated functional data sets. Similarly, Paparoditis (2016) also tested the efficacy of a bootstrap technique for testing equality of means, with good results shown for the power of the L_2-norm used in the study. These papers all showed satisfactory results for the bootstrapping methods, with high power obtained for the hypothesis tests. However, all of these studies focus on the L_1-norm or L_2-norm as the test statistics for the equality between the mean curves.

In this paper we evaluate the power of the bootstrap-based hypothesis testing of the equality of mean functions considering multiple test statistics. The L_1-norm is one of the statistics considered. This is a common statistic used in bootstrapping (Allen, 1997), and was the statistic used by Paparoditis (2016) for their hypothesis testing. The integrated square difference, or the L_2-norm is another popular statistic for testing in Functional Data (Crainiceanu, 2012; Zhang, 2010). Another statistic considered is the Kullback-Leibler divergence, which was found to perform well asymptotically, and can be used to construct an asymptotically normal test statistic for density estimation (Dias, 2007). We also use the Hellinger distance as a statistic. Su (2008) performed non parametric tests for density functions using the Hellinger statistic, with high power obtained. The final test statistic used is the affinity between the mean functions, which is based on the square of the Hellinger Distance. The power of the bootstrap tests are evaluated using the empirical rejection probability (Hall, 1991).

The statistics are evaluated using simulated data via spline-smoothing techniques. Similar techniques have been used in the other papers discussed such as Zhang (2010), as well as papers testing for equality of functional data with different methods (Lima, 2018). Spline-smoothing as a tool for non-parametric regression curve fitting was explored in Silverman (1985). He found that it was both a flexible and effective method for estimation. The specific spline smoothing technique used in this paper is a B-spline basis with smoothing splines, or a continuous roughness penalty. A study performed by Aguilera (2013) compared different B-spline basis smoothing approaches to functional data. The results showed that the roughness penalty is very effective at reducing the mean squared error of the sample curves.

3 Methodology

3.1 Smoothing Splines

The following steps are used to fit smooth curves to the data points generated via simulation. In this paper we use cubic b-splines.

1. Define the knot sequence $\xi_1 \leq ... \leq \xi_d \leq \xi_{d+1} < ... < \xi_{d+K+1} \leq \xi_{d+K+2} ... \leq \xi_{2d+K+2}$ where sets $\xi_{d+2} = \tau_1, ..., \xi_{d+K+1} = \tau_K$ and $\xi_{d+1} = a, \xi_{d+K+2} = b$ are referred to as inner and boundary knots. The additional knots are chosen arbitrarily, and are here set equal to boundary knots

2. Obtain the $m \times (K + d + 1)$ matrix of B-spline basis function evaluation where each element $B_k(x_i)$ for $k = 1, ..., K$ and $i = 1, ..., m$ is defined by recursive formula:
   $B_k^3 = \frac{x-\xi_k}{\xi_{k+d}-\xi_k} B_k^{d+1}(x) - \frac{\xi_{k+d+1}-x}{\xi_{k+d+1}-\xi_{k+1}} B_{k+1}^{d-1}(x), k = 1, ..., K + d + 1$

   Where $B_k^0(x) = \begin{cases} 1 & \xi_k \leq x < \xi_{k+1} \\ 0 & else \end{cases}$

   Fortunately, B-spline evaluations can be easily obtained using the R package *fda()*

Find estimates of coefficients $\hat{\beta}$ where $\hat{\beta} = argmin_\beta \sum_{i=1}^n (f_\beta(x_i) - y_i)^2 + \lambda \int_a^b (\partial^2 f / \partial x^2)^2 dx$

Construct smooth curve $f(x) = \sum_{k=1}^{K+d+1} \hat{\beta}_k B_k(x)$

3.2 Constructing the Test Statistics

Let W_2^2[a,b] be the set of functions in [a,b] where the first derivative of the function is absolutely continuous and the second derivative is squared integrable. This allows us to compute the roughness penalty when smoothing the data. Additionally, consider the transformation

$$t_f = \frac{f^2}{\int f^2} \tag{1}$$

for any square integrable function f. $t_f \geq 0$ and $\int t_f = 1$, therefore t_f is a density function.

Let the functions μ_x and $\mu_y \in W_2^2$[a,b] be the mean curves of a set of functions. We can find the estimates of these curves, $\hat{\mu}_x$ and $\hat{\mu}_y$ by taking the pointwise mean of the smooth curves. The test statistics are then defined based on the transformation of these curve estimates, $t_{\hat{\mu}_x}$ and $t_{\hat{\mu}_y}$.

The L_1 Distance is the most common way to measure the distance between 2 functions, and is defined by:

$$L_1 = \int |t_{\hat{\mu}_x} - t_{\hat{\mu}_y}|. \tag{2}$$

The integrated square difference, which is equivalent to the L_2 distance, is another common measure of distance between functions, defined by:

$$ISD = \int (t_{\hat{\mu}_x} - t_{\hat{\mu}_y})^2. \tag{3}$$

The Hellinger distance is a statistic often used to quantify the distance between two probability distributions. As our transformation t_f creates a density function, we can use this distance as a statistic for the equality of curves, defined as:

$$H = \left(\int \left(\sqrt{t_{\hat{\mu}_x}} - \sqrt{t_{\hat{\mu}_y}} \right)^2 \right)^{\frac{1}{2}}. \tag{4}$$

The Hellinger distance can also be used to define another of our test statistics, the affinity, based of the square of the Hellinger. From Equation (3.4), we can square to create:

$$H^2 = \int \left(\sqrt{t_{\hat{\mu}_x}} - \sqrt{t_{\hat{\mu}_y}} \right)^2 = 2(1 - \rho) \tag{5}$$

where

$$\rho = \int (\sqrt{t_{\hat{\mu}_x} t_{\hat{\mu}_y}}). \tag{6}$$

Finally, the Kullback-Leibler divergence is given by:

$$KL = \int (\log t_{\hat{\mu}_x} - \log t_{\hat{\mu}_y}) t_{\hat{\mu}_x}. \tag{7}$$

All integrals above are approximated by a finite sum. For example when computing the L_1 statistic we will use $\hat{L}_i = \Delta s[\sum_{i=1}^{m} |t_{\hat{\mu}_x}(s_i) - t_{\hat{\mu}_y}(s_i)|]$. Where Δs is the length of the equally spaced intervals between observations.

3.3 Generating Simulated Data

Here we show the necessary steps to generating the sample of curves to be used in the bootstrap procedure.

1. A function $f \in W_2^2$[a,b] is chosen and m points of the function are calculated.

2. A random error with distribution $N(0, \sigma^2)$ is added to each point to create observations $y_i = f(t_i) + \epsilon_i, i = 1, 2, ..., m$, to be used as data in experiment.

3. Repeat step 2 n times to create set of n functional data, where y_{ij} represents the ith observation of the jth datum. We also create a covariance structure within each curve by allowing $y_{ij} = f(t_i) + a_j + \epsilon_{ij}$, where $a_1, ..., a_n$ are iid $N(0, \sigma_a^2)$. This will not interfere with independence between curves, as this structure is only inside each curve.

4. Smooth the data as described in section 4.1, creating a sample of n smooth curves $\hat{f}_1, ..., \hat{f}_n$.

   A new sample of curves can be created by choosing a different function $g \neq f$ for step 1.

3.4 Bootstrap-Based Test Procedure

The bootstrap re-sampling method is used to test the hypothesis of equality between the two mean curves, or $H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x \neq \mu_y$. The bootstrap procedure obtains samples by re-sampling with replacement from the original set of curves. Let $\{X_1(t), X_2(t), ..., X_n(t)\}$ and $\{Y_1(t), Y_2(t), ..., Y_n(t)\}$ be two independent curve samples and let θ be value of test statistic, the bootstrap procedure is as follows:

1. Calculate value θ from the two original samples.

2. Join original samples together to form new set of samples of size $n_1 + n_2$.

3. Create bootstrap samples by sampling $n_1 + n_2$ curves with replacement from new sample. The first n_1 curves are denoted as $X_1^*(t), X_2^*(t), ..., X_n^*(t)$, next n_2 curves as $Y_1^*(t), Y_2^*(t), ..., Y_n^*(t)$.

4. Calculate value θ^* from the bootstrap samples.

5. Repeat steps 3 and 4 B times, obtain B θ^* values.

6. For a significance level α, reject null hypothesis $H_0 : \mu_x(t) = \mu_y(t)$ if

$$p\text{-value} = \frac{\{\#\theta^* \geq \theta\} + 1}{B + 1} \leq \alpha,$$

, or in the case of the affinity test statistic reject if

$$p\text{-value} = \frac{\{\#\theta^* \leq \theta\} + 1}{B + 1} \leq \alpha.$$

Note that the reason that there are two p-values is due to the fact that the smaller values of the affinity are evidence that the null hypothesis is false. This is in contrast to the other test statistics being measured, where larger values are evidence against the null hypothesis. We also add 1 to both the numerator and denominator of the p-value to increase stability, in case the null hypothesis is rejected 0 time(Hall, 1991).

3.5 Evaluating Power of the Tests

The power of a test is the probability that H_0 will be rejected. When the null hypothesis is true, the power should be as close to the significance level α as possible. When the null hypothesis is false, the power is going to be the power of the test. The empirical rejection probability will be used to approximate the results of the power function. For each test statistic the empirical rejection probability is calculated as follows:

1. Generate two samples of curves $\{X_1(t), X_2(t), ..., X_n(t)\}$ and $\{Y_1(t), Y_2(t), ..., Y_n(t)\}$ using the data simulator described in Section 4.2.
2. Smooth the curves and find mean curves $\hat{\mu}_x$ and $\hat{\mu}_y$, and apply transformations to acquire $t_{\hat{\mu}_x}$ and $t_{\hat{\mu}_y}$ so that test statistics can be computed.
3. Test the null hypothesis $H_0 : \mu_x = \mu_y$ using each test statistic and bootstrap procedure outlined in Section 4.3.
4. Repeat previous steps R times, define the empirical rejection probability as number of times H_0 is rejected divided by R.

4 Dataset description

The datasets that are used in this paper are primarily generated via the simulation techniques described in Section 3.3. The code required to simulate this data can be found in the repository Liam Bullen on GitHub.

Additionally, the method is tested used on growth data from boys and girls in the UK. This dataset can be found under the name growth on the FDA package in RStudio. The dataset is a collection of growth curves containing the heights of 39 boys and 54 girls. The data was collected at unequally spaced intervals from the ages of 1-18.

5 Experimental Setting

There are 4 scenarios tested in this paper:

1. Balanced samples with small sample size, $n_1 = n_2 = 50$
2. Unbalanced samples with varying sample size, $n_1 = 50, n_2 = 100$
3. Balanced samples with large sample size, $n_1 = n_2 = 200$
4. Scenario 3 with random effects added, $g_{ij}(t) = f(t) + a_j + \epsilon_{ij}$

There are two functions that are used during the simulation study. When creating the smooth curves from the data, 100 knots were used for both functions. For the roughness penalty a λ of $1e^-5$ was used for function 1 and $1e - 4$ for function 2. A sample dataset of the curves generated by these functions, along with the mean curve, can be seen below. Figure 5.1 is displaying the function $2 - 5t + 5\exp[-100(t - 0.5)^2]$, while Figure 5.2 shows $1.5 \times 0.4(1.5t)^{-0.6} \exp(-(1.5t)^{0.6})$. Scenarios 1 and 2 are tested on the first function, while scenarios 3 and 4 are tested on both functions 1 and 2. The method is then used on the growth dataset.

The growth curves are also evaluated. Figure 5.3 shows all of the growth curves, with the black curves representing boys and the red curves representing girls.In Figure 5.4 we see each growth curve plotted in grey, while the mean curve for both the boys and the girls are highlighted in black and red respectively. We can see the curves are extremely similar until a divergence at the age of around 13.

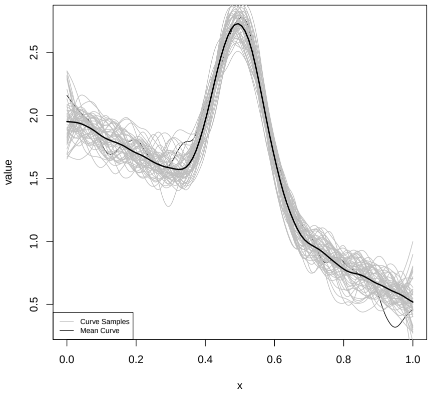


Figure 5.1: Function 1

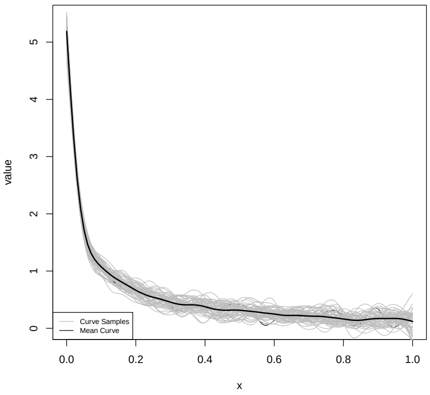


Figure 5.2: Function 2

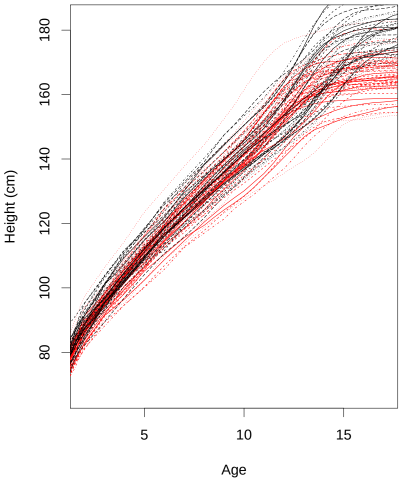


Figure 5.3: Child Growth Curves

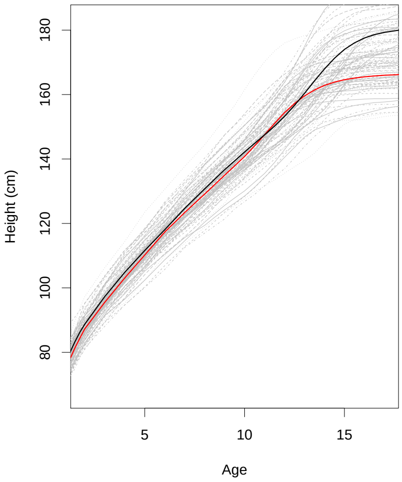


Figure 5.4: Mean Growth Curves

6 Results and discussion

6.1 Simulation Study

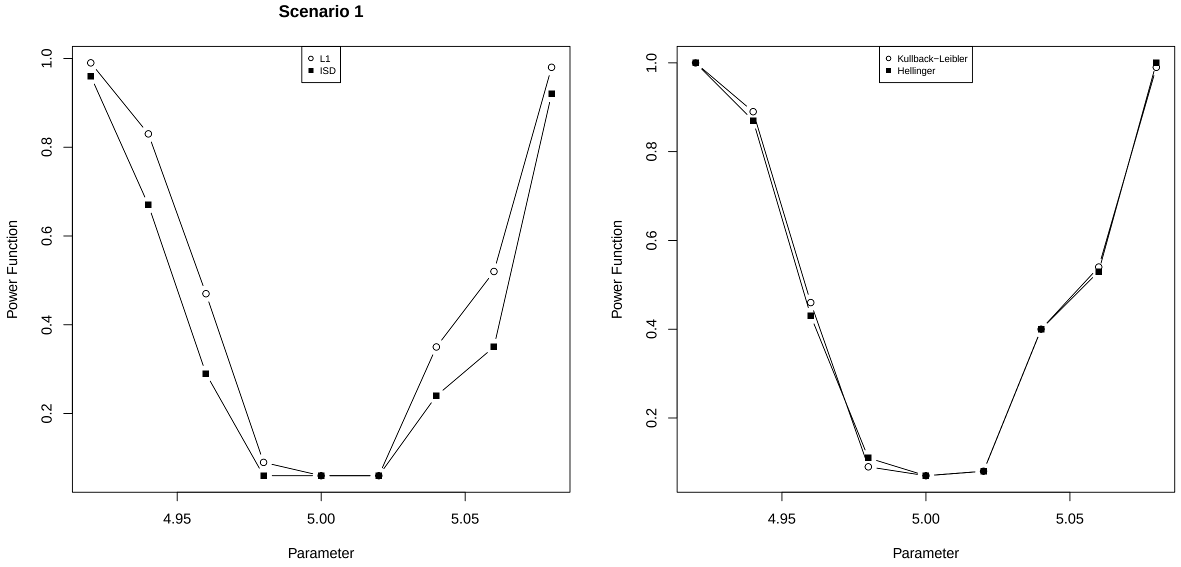
6.1.1 Scenario 1

We begin with the results of the first scenario, where the samples are balanced with a sample size of 50. The results of the experiment can be seen in table 6.1 below. When the value pf β was equal to 5, the empirical rejection probability was seen to be around 0.05, as we would like to see. It is important to note that due to the small samples involved in the curves, there is more noise to the data. The best statistics under this scenario were the Kullback-Leibler divergence, the Hellinger distance and the affinity, which all performed similarly well. The L_1 distance and ISD, however, lagged behind these statistics with less rejection coverage at each value of β. The Hellinger distance and affinity also returned identical rejection probabilities for all values of β, this is true for all scenarios for both functions, due to the fact that the affinity is derived from the Hellinger distance.

	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
4.92	0.97	0.86	1.00	1.00	1.00
4.94	0.83	0.67	0.89	0.87	0.87
4.96	0.38	0.26	0.36	0.32	0.32
4.98	0.09	0.06	0.09	0.11	0.11
5.00	0.06	0.06	0.07	0.05	0.05
5.02	0.06	0.06	0.08	0.08	0.08
5.04	0.35	0.24	0.40	0.40	0.40
5.06	0.52	0.35	0.54	0.53	0.53
5.08	0.98	0.92	1.00	0.99	0.99

Table 6.1: Scenario 1 results, $\hat{\mu}_x = 2 - 5t + 5\exp[-100(t-0.5)^2]$, $\hat{\mu}_y = 2 - \beta t + \beta\exp[-100(t-0.5)^2$

The results are also displayed in figure 6,1 below, displaying the rejection probability as the parameter varies for each test statistic. Again we can see the best performing statistics are the KL divergence, Hellinger and Affinity as they all approach an empirical rejection probability of 1 the quickest, at an extremely similar rate. We can also see that the ISD is lagging behind the L_1 in terms of power function. This is a significant finding as previously the L_1 distance and ISD have been the only statistics used in hypothesis testing in functional data. Interestingly there is a slight asymmetry in the results, as all functions were able to approach 1 faster when the value of β was less than the true mean curve than when the value was greater.



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

Figure 6.1: Estimated Power Functions: Scenario 1

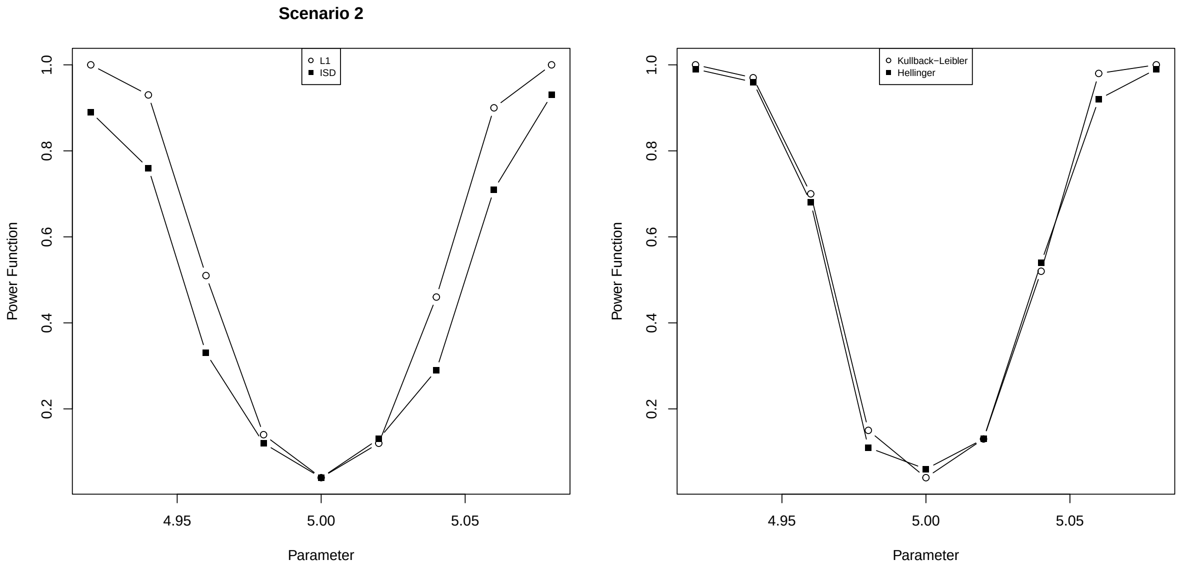
6.1.2 Scenario 2

In the next scenario we examine the datasets have an unbalanced sample size. This can frequently be seen in real world analysis due to datasets having different collection methods, sample sizes, or partial corruption or loss of data. The sample sizes used in this scenario are 50 for the first group, and 100 for the second. The results of the experiment can be seen in table 4.2 below. Again the test statistics all were all able to reject the null hypothesis of equality with high power when the functions were not equal. The statistics had a slightly higher empirical rejection probability than the first scenario, likely due to the increased total sample size from the first, as the second curve had double the sample size. Again we see the same statistics performing best, as the Kullback-Leibler divergence had the converged the quickest closely followed by the Hellinger distance.

	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
4.92	1.00	0.89	1.00	0.99	0.99
4.94	0.93	0.76	0.97	0.96	0.96
4.96	0.51	0.33	0.70	0.68	0.68
4.98	0.14	0.12	0.15	0.11	0.11
5.00	0.04	0.04	0.04	0.06	0.06
5.02	0.12	0.13	0.13	0.13	0.13
5.04	0.46	0.29	0.52	0.54	0.54
5.06	0.90	0.71	0.98	0.92	0.92
5.08	1.00	0.93	1.00	0.99	0.99

Table 6.2: Scenario 2 results, $\hat{\mu}_x = 2 - 5t + 5\exp[-100(t-0.5)^2]$, $\hat{\mu}_y = 2 - \beta t + \beta\exp[-100(t-0.5)^2$

The estimations of the power functions can again be seen in figure 6.2 for each parameter. The results are very similar to scenario 1, with the $L-1$ and ISD power functions approaching one significantly slower than the Kullback-Leibler divergence and Hellinger Distance. The power functions also had more symmetry than the results of Scenario 1, perhaps as a result of the increased sample size resulting in more stable estimates for rejection probability.



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

Figure 6.2: Estimated Power Functions: Scenario 2

6.1.3 Scenario 3

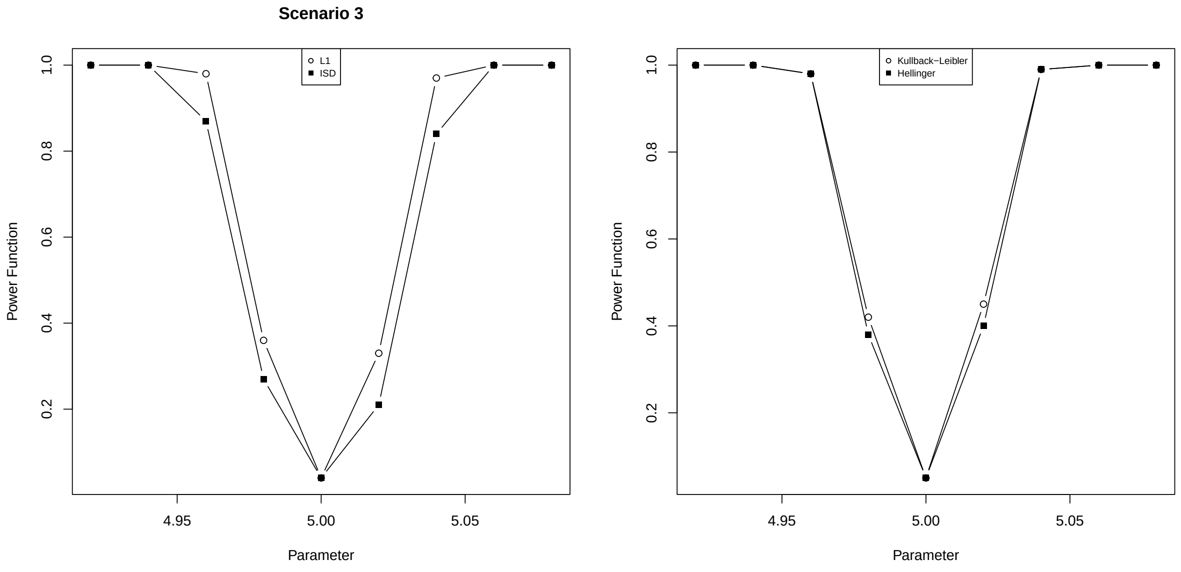
In this scenario we have a large sample size for both of the curve datasets, with $n_1 = n_2 = 250$. For this scenario we test both of the functions mentioned in the experimental setting. The results for the first function,

$y = 2 - 5t + 5\exp[-100(t-0.5)^2]$, are summarized in table 6.3. There was a significant improvement in the speed with which the empirical rejection probability approached one for all statistics. This is to be expected when we raise the sample size by a large degree, as it will lead to increased stability in the estimates.

	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
4.92	1.00	1.00	1.00	1.00	1.00
4.94	1.00	1.00	1.00	1.00	1.00
4.96	0.98	0.87	0.98	0.98	0.98
4.98	0.36	0.27	0.42	0.38	0.38
5.00	0.04	0.04	0.05	0.05	0.05
5.02	0.33	0.21	0.46	0.40	0.40
5.04	0.97	0.84	0.99	0.99	0.99
5.06	1.00	1.00	1.00	1.00	1.00
5.08	1.00	1.00	1.00	1.00	1.00

Table 6.3: Scenario 3 results, $\hat{\mu}_x = 2 - 5t + 5\exp[-100(t-0.5)^2]$, $\hat{\mu}_y = 2 - \beta t + \beta\exp[-100(t-0.5)^2$

The increased performance of the power functions are shown very clearly in figure 6.3, as the slope of the functions is much steeper than in the previous scenarios. As is the case with the earlier scenarios, the Kullback-Leibler divergence was the statistics with the best performance by empirical rejection probability for function 1.



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

Figure 6.3: Estimated Power Functions: Scenario 3, Function 1

When testing function 2, $y = 1.5 \times 0.4(1.5t)^{-0.6}\exp(-(1.5t)^{0.6})$, the statistics had significantly different results than function 1. The power obtained for all statistics was satisfactory, as empirical rejection probability again approached 1 quickly as the mean curves were no longer equal. However, the results for which statistics performed better were reversed. In table 6.4, we can see that the ISD had the best performance for this function, as well as the L_1 norm. The Kullback-Leibler divergence outperformed the Hellinger distance slightly again, but they both had lower power than the L_1 norm and ISD.

Figure 6.4 shows the power functions for the second function, where clearly the ISD and L_1 are converging to 1 faster than the Kullback-Leibler and Hellinger. The estimated power functions are again quite symmetric in this scenario.

	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
1.10	1.00	1.00	1.00	0.99	0.99
1.20	0.91	0.95	0.82	0.78	0.78
1.30	0.42	0.46	0.34	0.28	0.28
1.40	0.17	0.20	0.10	0.08	0.08
1.50	0.05	0.04	0.05	0.06	0.06
1.60	0.13	0.15	0.07	0.05	0.05
1.70	0.36	0.39	0.29	0.26	0.26
1.80	0.84	0.85	0.73	0.70	0.70
1.90	0.98	0.99	0.96	0.94	0.94

Table 6.4: Scenario 3 results, $\hat{\mu}_x = 1.5 \times 0.4(1.5t)^{-0.6} \exp(-(1.5t)^{0.6})$, $\hat{\mu}_y = \beta \times 0.4(\beta t)^{-0.6} \exp(-(\beta t)^{0.6})$

6.1.4 Scenario 4

The final scenario we examine in this paper is the large sample size of $n_1 = n_2 = 250$, with a random effect added to each curve. This will result in higher spread of sample curves, as the additional parameter of the random effect will add bias to the curve estimates. As tables 6.5 and 6.6 show, the increased sample size again caused the empirical rejection probability to be quite sensitive to the changing of the parameter in the mean curves for both the functions. There was a small decrease in coverage as compared to Scenario 3. This is likely due to the increased variance of the random effects adding more uncertainty, causing the results to become slightly less accurate.

However, looking at the power functions for function 1 and 2 in figures 6.5 and 6.6 respectively, we can again see that they are able to reach 1 quickly as the parameters change. The same test statistics perform well, with the Kullback-Leibler and Hellinger distance achieving the best results for function and the ISD and L_1 norm performing the best for function 2. It is likely that these trends will hold for all sample sizes, as there has been a consistent ordering in the performance of these statistics for both functions across scenarios.

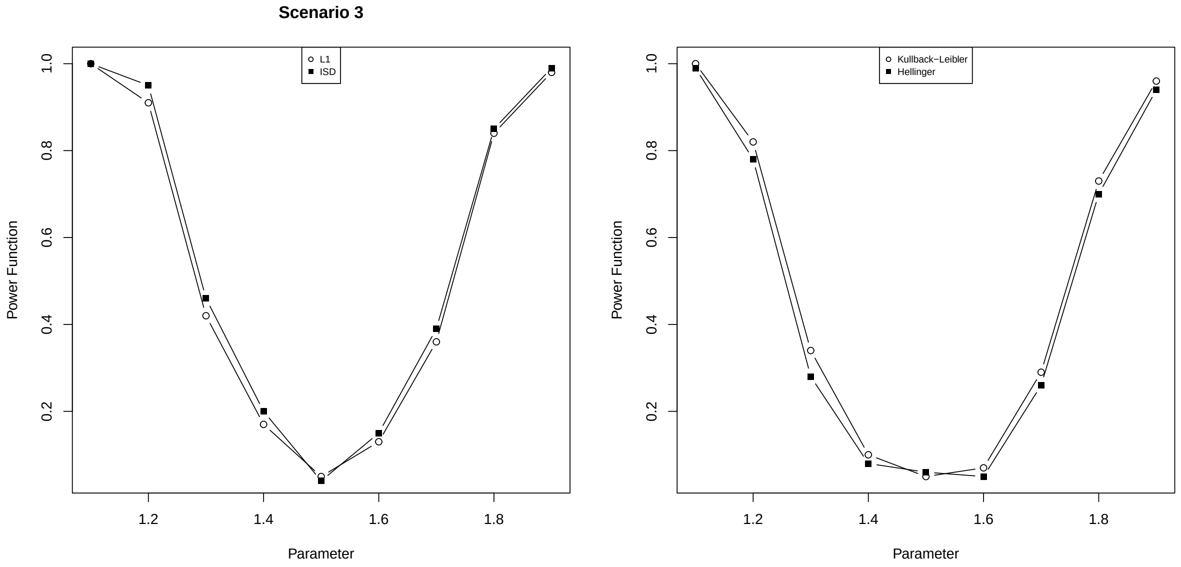
	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
4.92	1.00	1.00	1.00	1.00	1.00
4.94	1.00	0.99	1.00	1.00	1.00
4.96	0.87	0.69	0.91	0.88	0.88
4.98	0.20	0.18	0.28	0.26	0.26
5.00	0.05	0.06	0.07	0.07	0.07
5.02	0.16	0.14	0.24	0.24	0.24
5.04	0.82	0.64	0.89	0.88	0.88
5.06	1.00	0.96	1.00	1.00	1.00
5.08	1.00	1.00	1.00	1.00	1.00

Table 6.5: Scenario 4 results, $\hat{\mu}_x = 2 - 5t + 5\exp[-100(t-0.5)^2]$, $\hat{\mu}_y = 2 - \beta t + \beta \exp[-100(t-0.5)^2$

6.2 Application: Real World Data

The simulation studies obtained satisfactory results for all scenarios, successfully rejecting the null hypothesis at a high power when the mean curves were unequal. To show the performance of the method developed in this paper, we will now apply it to the growth dataset described in Section 5. As it is unclear which test statistic obtains the highest power on hypothesis tests depending on the function, we will use both the Kullback-Leibler divergence and the ISD when testing for equality, as they were the statistics that had the highest empirical rejection probability for functions 1 and 2 respectively.

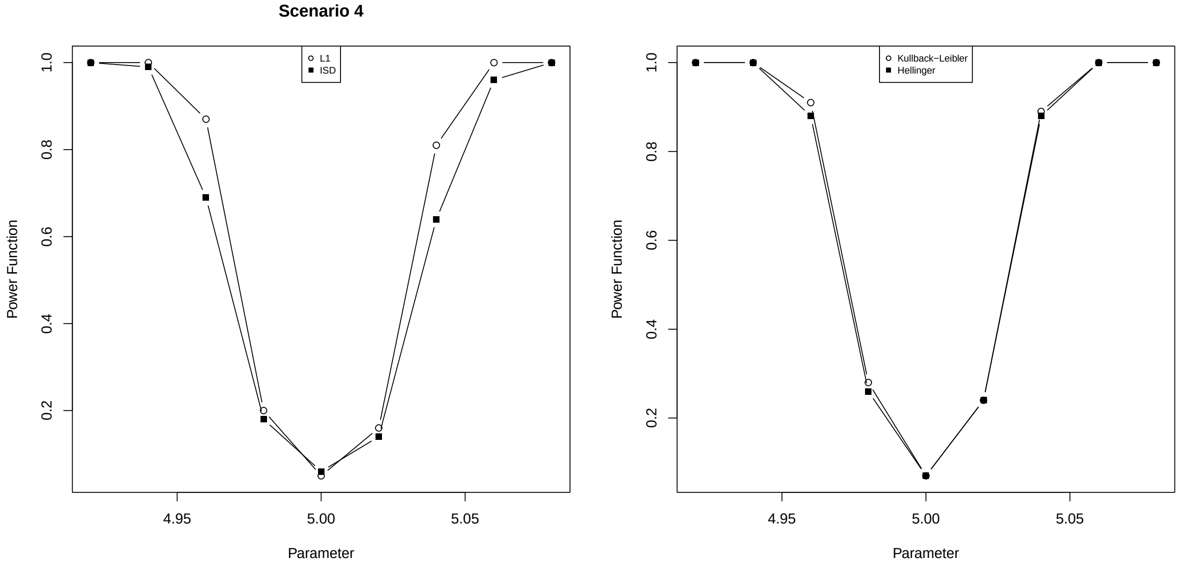
In Figure 6.7 the mean curves of the growth datasets are outlined, with the black curve representing the boys and the red curve representing the girls. We see that the curves are extremely similar until age 13, when the growth curve of the boys begins to outpace the girls. As the mean curves have a significant difference, we expect the bootstrap-based testing method developed is able to reject the null hypothesis of equality. Performing the test resulted in a p-value of 0.01 for both the Kullback-Leibler divergence and the ISD, successfully concluding that the mean curves of the two datasets are not equal.



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

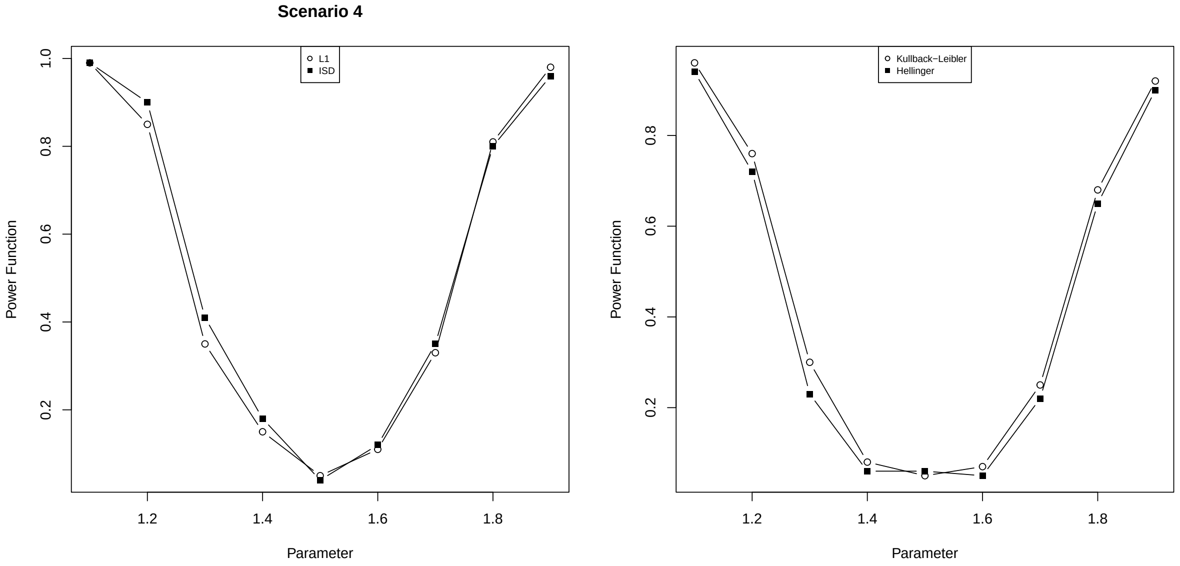
Figure 6.4: Estimated Power Functions: Scenario 3, Function 2



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

Figure 6.5: Estimated Power Functions: Scenario 4, Function 1



(a) L_1 and ISD

(b) Kullback-Leibler and Hellinger

Figure 6.6: Estimated Power Functions: Scenario 4, Function 2

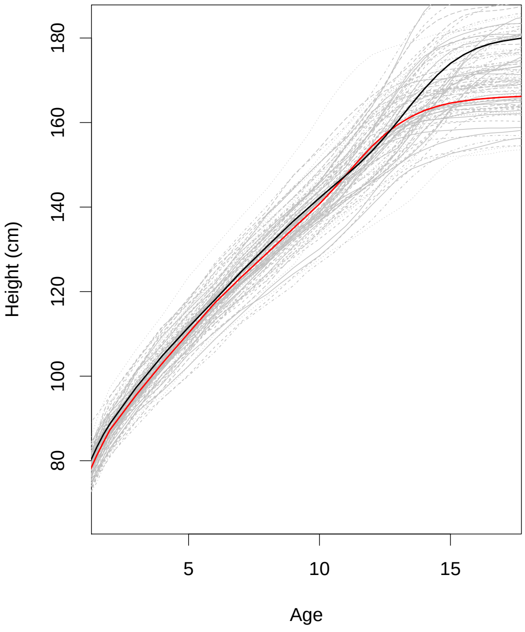


Figure 6.7: Growth Data Mean Curves

	Empirical Rejection Probability				
β	L_1	ISD	KL	Hellinger	Affinity
1.10	0.99	0.99	0.96	0.94	0.94
1.20	0.85	0.90	0.76	0.72	0.73
1.30	0.35	0.41	0.30	0.23	0.24
1.40	0.15	0.18	0.08	0.06	0.06
1.50	0.05	0.05	0.05	0.05	0.05
1.60	0.11	0.12	0.07	0.05	0.05
1.70	0.33	0.35	0.25	0.22	0.22
1.80	0.81	0.80	0.68	0.65	0.65
1.90	0.98	0.96	0.92	0.90	0.90

Table 6.6: Scenario 4 results, $\hat{\mu}_x = 1.5 \times 0.4(1.5t)^{-0.6} \exp(-(1.5t)^{0.6})$, $\hat{\mu}_y = \beta \times 0.4(\beta t)^{-0.6} \exp(-(\beta t)^{0.6})$

7 Conclusion and Future Work

From the results of our simulation studies, all statistics were able to successfully reject equality of mean functions when parameters were unequal. The power functions of every test statistic approached 1 quickly as the mean curves diverged more. The findings suggest that Kullback-Leibler and Hellinger can outperform the more common L_1 and ISD statistics depending on the form of the functions being tested. Namely, for Function 1 the Kullback-Leibler was the best performing statistic. This could be due to the exponential form in the function, as the Kullback-Leibler was able to The test statistics also performed as expected with a sample size of growth data that was smaller than simulated scenarios. Both the statistics measured, the L_1 norm and the Kullback-Leibler Divergence, were able to accurately reject the null hypothesis and conclude that the 2 growth curves were not equal. Despite the curves being equal up until the age of around 13, the statistics were sensitive enough to find the divergence after puberty in the growth curves.

We conclude that if working with functional data of similar forms to the functions used in this paper, the method developed in this paper with the optimal test statistics are able to accurately test the equality of 2 mean functions. However, if the functions are of a different form then more research is need to find the optimal test statistic. This method is still likely applicable to functions of different forms, as the power of all the test statistics were satisfactory in the simulated studies. The decision of which test statistic is best suited is left to researchers working with the functional data.

There were some constraints that limited the effectiveness of the conclusions in this paper. Due to limited computational power relatively low levels of bootstraps were necessary, as for the simulations B=75 was used. A higher number used could increase the confidence in these results, as there will be more variance due to the low level of bootstrapping. Additionally, only two functions were tested in this paper using this method. As the functions resulted in quite different power functions for the test statistics in the simulated studies, more research is likely needed. Different types of functions should be examined to see what determines which statistic has best performance. More scenarios could be tested in future research, as our scenarios were mainly focused on different sample sizes, as well as a random effect structure. For example, a scenario that generated outliers could be used to find robustness of statistics.

References

Aguilera, A.-M. (2013). Comparative study of different b-spline approaches for functional data. *Mathematical and Computer Modelling*, *58*(1), 1568–1579.

Allen. (1997). Hypothesis testing using an l1-distance bootstrap. *The American Statistician*, *51*(1), 145–150.

Crainiceanu, R., Staicu. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, *31*(26), 3223–3240.

Cuevas, F. (2003). An anova test for functional data. *Computational Statistics and Data Analysis*, *47*(1), 111–122.

Dias, G. (2007). Consistent estimator for basis selection based on a proxy of the kullback–leibler distance. *Journal of Econometrics*, *141*(1), 167–178.

Ferraty, V. (2007). *Non-parametric functional data analysis, theory and practice.* Springer-Verlag.

Hall, W. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, *47*(1), 757–762.

Horvath, K. (2012). *Inference with functional data for applications.* Springer-Verlag.

Lima, B., Cao. (2018). Robust simultaneous inference for the mean function of functional data. *Test*, *28*(2), 785–803.

Paparoditis, S. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, *103*(3), 727–733.

Ramsay, D. (1982). Some tools for functional data analysis. *Journal of Royal Statistical Society*, *53*(3), 539–561.

Ramsay, S. (2005). *LATEX: Functional data analysis*. Springer-Verlag.

Silverman. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of Royal Statistical Society*, *47*(1), 1–21.

Silverman, R. (2002). *Applied functional data analysis, methods and case studies*. Springer.

Su, W. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, *24*(4), 829–864.

Zhang, Z., Peng. (2010). Two sample test for functional data. *Communications in Statistics*, *39*(4), 559–578.