Pointwise Testing with Functional Data Using the Westfall-Young Randomization Method

DENNIS D. COX
Rice University, Department of Statistics,
6100 Main St. MS-138, Houston, TX 77005, USA,
e-mail: dcox@stat.rice.edu.
and
JONG SOO LEE
Carnegie Mellon University, Department of Statistics
Pittsburgh, PA 15213, USA,
e-mail: jslee@stat.cmu.edu

Abstract

We consider hypothesis testing with smooth functional data by performing pointwise tests and applying a multiple comparisons procedure. Methods based on general inequalities (such as Bonferroni's method) do not perform well because of the high correlation between observations at nearby points. We consider the multiple comparison procedure proposed by Westfall and Young (1993) and show that it approximates a multiple comparison correction for a continuum of comparisons as the grid for pointwise comparisons becomes finer. Simulations and an application verify that this result applies in practical settings.

Keywords: Functional data analysis; Hypothesis testing; Multiple comparison procedure; Permutation method.

1 Introduction

In the setting of functional data analysis, an idealized observation is a smooth function $y(x)$ with x in some domain D. In practice, we would typically not deal directly with the $y(x)$ for all $x \in D$ but rather evaluate the function at some finite set of points $D_m = \{x_1, \ldots, x_m\} \subset D$ and produce a data vector $y = (y_1, \ldots, y_m)'$ where $y_j = y(x_j)$. The choice of the set of evaluation points D_m is somewhat arbitrary. Alternatively, one may represent the observed functions as a linear combination of basis functions, simply store the coefficients of the expansion, and then function values can be computed as needed. Assume we have a sample of independent functional observations $y_1(x), \ldots, y_n(x)$. The dimension m of the stored vectors will often be much larger than the sample size n. Thus, most of the inferential methods from multivariate analysis (Rencher, 2002) cannot be used directly since they require inversion of the sample covariance matrix. Moreover, multivariate methods do not take into account the functional nature of the data. The methods of functional data analysis (Ramsay and Silverman, 2005) are designed to deal with these issues.

Here we consider one important matter when the inferences are based on the functions evaluated on a grid of points. Since the grid of evaluation points D_m is arbitrary, a minimal requirement is that any inferential method be somewhat independent of the choice of D_m. Furthermore, it would be highly desirable that any such method converge to an appropriate continuum method as the number of grid points increases and D_m becomes dense in D.

We refer to this as the grid refinement invariance property. If the inferences are based on coefficients of a basis function expansion, then they should have a similar property as the order of the basis expansion become large.

Suppose we wish to test for differences in the means of curves from several populations, i.e. functional analysis of variance. There are methods to detect the overall difference in the mean curves (Fan and Lin, 1998; Shen and Faraway, 2004; and Zerbe and Murphy, 1986). Just as with ordinary analysis of variance and multivariate analysis of variance, after we decide that the means from the two or more samples are significantly different by an overall testing method, we want to identify more specifically where the differences are. In the functional analysis of variance setting, a natural goal is to determine the specific region of D where the differences occur.

For this problem, we consider a testing procedure based on a randomization method due to Westfall and Young (1993). This could be used as a follow-up to an overall test, like one of the methods referenced above, or simply by itself. The Westfall-Young randomization method is a stepwise procedure which first conducts an appropriate univariate test at each $x \in D_m$ and then applies a correction for the multiple comparisons. Many multiple comparison correction methods (e.g. Bonferroni) will be highly dependent on the number of comparisons m, which in our setting is the number of grid points, and any statistical significance will disappear as $m \to \infty$. This is very undesirable with functional data since one wants a large number of evaluation points for high accuracy, and the number of grid points is arbitrary. We will show that if one uses the method of Westfall and Young for the multiple comparison, then as $m \to \infty$ and D_m becomes dense in D, the region of statistically significant differences will converge to a reasonable limit which is almost the region that would be found if one corrected for the

continuum of comparisons at all $x \in D$.

There exists some recent work that are related and can possibly be used to extend our the results presented here. Efron (2007) shows how one can use permutations and other transformations of the data to estimate and adjust for the correlation between test statistics. It may be possible to use this approach to satisfy other error criteria than the family wise error rate, which is the error criterion used here. A methodology which is similar in spirit to the procedure investigated here is found in Taylor, et. al. (2007) but is based on extensions to Bonferroni's inequality which are suitable for Gaussian processes. This method doesn't require a permutation pivotality condition which is needed for the method analyzed here.

We describe the procedure in Section 2 in a simple setting. The main results are in Section 3. We present some simulation results in Section 4, and an application in Section 5. We conclude by discussing some open problems in Section 6.

2 Example: The Two Sample Case

In this section, we consider the problem of testing equality of means in the two sample setting. Suppose we have two independent samples $y_{ij}(x)$, $i = 1, 2$, $j = 1, \ldots, n_i$, where i denotes the population from whence the observation is drawn and j indexes the observation within the sample. Assume the observations within the samples are independent and identically distributed. We assume the observations are realizations of Gaussian processes with continuous mean functions $\mu_i(x) = E\{y_{ij}(x)\}$, $i = 1, 2$, and a common covariance function. We wish to test the null hypothesis $H_0 : \mu_1(x) = \mu_2(x)$, for all $x \in D$ versus the general alternative $H_1 : \mu_1(x) \neq \mu_2(x)$, for at least

one $x \in D$. Consider the pointwise testing problems $H_0(x) : \mu_1(x) = \mu_2(x)$ vs. $H_1(x) : \mu_1(x) \neq \mu_2(x)$. This is a simple univariate two-sample problem for each x. Because of the Gaussian and common covariance assumptions, the two sample t-test can be used to test each $H_0(x)$.

We assume that all curves have a common set of evaluation points x_k, $1 \leq k \leq m$. To apply the method analyzed here, one first computes a p-value $p(x_k)$ using the univariate two sample t-test of the null hypothesis $H_0(x_k)$ at each grid point x_k, and then applies a multiple comparison procedure that controls for the family-wise error rate. To define the family-wise error rate, consider $C_m = \{x_k : H_0(x_k) \text{ is true}, 1 \leq k \leq m\}$, which is the set of grid points for which $H_0(x_k)$ is true. Then the family-wise error rate is pr{reject $H_0(x_j)$ for any x_j in C_m}. Note that this depends on the set of grid points. Our goal is to make the family-wise error rate less than or equal to α, where α is given, no matter what is the set C_m of true null hypotheses.

We will show now that for some multiple comparison procedures the results can depend critically on the choice of the x_j grid in a very undesirable way. The simplest such procedure is the Bonferroni method. With this procedure, for the family-wise error rate to be less than or equal to α, we would reject any $H_0(x_j) : \mu_1(x_j) = \mu_2(x_j)$ for which $p(x_j) \leq \alpha/m$. Alternatively, we can define a p-value which is corrected for the multiple comparisons, namely $\min\{1, mp(x_j)\}$, and reject $H_0(x_j)$ if this corrected p-value is less than or equal to α. This procedure is clearly too conservative and depends heavily on the number of grid points m. In general, the p-value function $p(x)$ will be continuous and positive, so as we refine the grid and $m \to \infty$, we will eventually not reject any $H_0(x)$. An improvement on the Bonferroni procedure is Holm's method (Holm, 1979). This is a sequential step down method which proceeds as follows. First, the p-values, $p(x_j)$,

obtained from the univariate tests are ordered, $p_{(1)} \leq \ldots \leq p_{(m)}$. Then, find the smallest $j = j^*$ such that $p_{(j)} > \alpha/(m-j)$. Reject all $H_0(x_k)$ for which the corresponding $p(x_k) < p_{(j^*)}$. The rejection of the null hypothesis with the smallest p-value is equivalent to Bonferroni's method, so again, as we refine the grid we will eventually fail to reject any $H_0(x)$.

Bonferroni's and Holm's procedures are based on the subadditivity property of probability, which is accurate when the events of rejection are disjoint. However, with smooth functional data, the events of rejection of $H_0(x)$ and $H_0(x')$ are highly positively correlated when x and x' are close. Thus, we need an approach that somehow accounts for this correlation. To this end, we present the Westfall-Young randomization method, for which Westfall (2005) notes that it can "account for spatiotemporal correlations as well as nonnormal distributional characteristics." The analysis presented in the next section verifies this claim for functional data.

To use the Westfall-Young method, we need an additional property known as the Permutation Pivotality Condition. A general statement is given in the next section, but with our equal covariance Gaussian model, it is simple to obtain. Let $\mathcal{L}(W)$ denote the joint distribution of the random object W; if W is a stochastic process then $\mathcal{L}(W)$ is the collection of all finite dimensional distributions. In the two sample setting under discussion, suppose we randomly permute the population label $i \in \{1, 2\}$ in the data $\{y_{ij}(x_k) : 1 \leq i \leq 2,\ 1 \leq j \leq n_i,\ 1 \leq k \leq m\}$. Then no matter what is the subset C_m for which the null hypotheses are true, the joint distribution $\mathcal{L}\{y_{ij}(x_k) : 1 \leq i \leq 2,\ 1 \leq j \leq n_i,\ x_k \in C_m\}$ is left unchanged. It will be the distribution of independent and identically distributed multivariate normal observations with mean vector $\mu_1(x_k) = \mu_2(x_k)$, $x_k \in C_m$ and a covariance matrix determined from the assumed common covariance function.

We now describe the method of Westfall and Young starting from the sorted p-values as in the Holm's method: $p_{(1)} \leq \ldots \leq p_{(m)}$. Let π be the permutation that maps the original grid sequence onto the sorted p-value sequence, i.e. $p_{(j)} = p\{x_{\pi(j)}\}$. Now, randomly permute data between the two populations and call the resulting data set a randomized data set. We will denote quantities computed from such a randomized data set with a superscript $*$. Let $p^*_{(j)}$, $j = 1, \ldots, m$ be the p-values computed from the randomized data set put in the same order as the sorted p-values for the original data set, $p^*_{(j)} = p^*\{x_{\pi(j)}\}$. We repeat the randomization many times, say N times, to get an array of such randomized p-values, $\{p^*_{(j),\ell} : j = 1, \ldots, m \text{ and } \ell = 1, \ldots, N\}$. Next, compute $q^*_{(j),\ell} = \min\{p^*_{(s),\ell} : s \geq j\}$ for all $\ell = 1, \ldots, N$. Now, the corrected p-value corresponding to $p_{(j)}$ is the proportion of the $q^*_{(j),\ell}$ less than or equal to $p_{(j)}$, that is

$$r_{(j)} = N^{-1} \sum_{l=1}^{N} I\{q^*_{(j),l} \leq p_{(j)}\} \tag{1}$$

where $I(A)$ denotes the indicator random variable for the event A. Find the smallest j, call it j^+, such that $r_{(j)} > \alpha$. Reject $H_0(x_j)$ for all j satisfying $j < j^+$, and accept all other $H_0(x_j)$.

Let $j^\dagger$ be given by $p(x_{j^\dagger}) = \min\{p(x_j) : x_j \in C_m\}$. It has been shown that the method described controls the family-wise error rate to be less than or equal to α, since $\text{pr}\{r_{(j^*)} \leq \alpha\} \leq \alpha$ (Westfall and Young, 1993). We will show in the next section that as the set of evaluation points becomes dense in the domain, the set $\{r_{(j)} : j = 1, \ldots, m\}$ of corrected p-values approaches a limit which is almost the corrected p-values for the continuum of pointwise null hypotheses $H_0(x)$, $x \in D$. Note that the correspondingly corrected p-values under Bonferroni or Holm's method will converge to 1 under the same limiting conditions.

This idea of pointwise testing with multiple corrections has already been introduced and used in applications (Hoyte, et. al., 2001; Cox, et. al., 2003). The novelty of this work lies in the development of limiting properties which shows the Westfall-Young randomization method is appropriate for functional data.

3 Theoretical Results

3.1 Main Result

We now assume a very general setup. Denote the functional data by $y_k(x)$, $x \in D$, where k may be a vector index (e.g., $k = (i, j)$ with $i = 1, 2$ and $j = 1, \ldots, n_i$ for a two sample setup). Let $Y(x)$ denote the vector-valued function obtained by concatenating the values of $y_k(x)$ in some convenient way. We suppose that at each x, there is a null hypothesis $H_0(x)$ about $\mathcal{L}\{Y(x)\}$. There will also be a corresponding alternative hypothesis $H_1(x)$, of course, but it will not generally be mentioned. We suppose further that there is a test which can be summarized as a p-value $p(x)$. The following assumptions will be made:

1. We assume that for each x, $p(x)$ is a valid p-value, i.e., if $H_0(x)$ is true, then the distribution of $p(x)$ is stochastically greater than uniform.

2. As a function of x, $p(x)$ is continuous. Similarly, any p-value function computed from permuted data is continuous.

3. The domain D is compact.

4. The grid points for evaluation form increasing sequence D_m, $m = 1, \ldots,$ of finite subsets of D such that $\bigcup_{m=1}^{\infty} D_m$ is dense in D.

5. The following permutation pivotality condition holds. Let

$$C = \{x \in D : H_0(x) \text{ is true}\}.$$

Assume $\mathcal{G}$ is a group of transformations that acts on vectors of the same dimension as $Y(x)$ through permutations on the indices k. We suppose that for all $g \in \mathcal{G}$,

$$\mathcal{L}\{gY(x) : x \in C\} = \mathcal{L}\{Y(x) : x \in C\}.$$

Here, the equality means equality of all finite dimensional distributions. Since C is not known, this condition must hold for all possible C as well.

In the example presented in the previous section, we assumed the data were realizations of a Gaussian process with a common covariance, and used a two sample t-test to obtain the pointwise p-values. These were convenient assumptions which allow one to verify the conditions above, which are clearly more general than the setup of the previous section.

Let $p^*(x)$ be the p-value function computed from the permuted data obtained by selecting an element of $\mathcal{G}$ at random. Always $p(x)$ will denote the p-value function from the original data. Let pr^* be the probability measure associated with the random p^*'s. More precisely, $\text{pr}^*(A)$ is the probability measure conditional on the original data $\{Y(x) : x \in D\}$ which measures the proportion of the randomized p-values $p^*(x)$ satisfying the criteria in A.

Define

$$D(\theta) = \{x \in D : p(x) \geq \theta\} \quad \text{and} \quad D_m(\theta) = \{x \in D_m : p(x) \geq \theta\}.$$

Let

$$r(\theta) = \mathrm{pr}^* \left\{ \inf_{x \in D(\theta)} p^*(x) \leq \theta \right\}$$

and

$$r_m(\theta) = \mathrm{pr}^* \left\{ \inf_{x \in D_m(\theta)} p^*(x) \leq \theta \right\}$$

Note that $r_{(j)}$ defined in (1) satisfies $r_{(j)} = r_m\{p_{(j)}\}$, and we have already called $r_{(j)}$ a (discrete) corrected p-value. Call $r\{p(x)\}$ the *continuum corrected p-value*.

We now give a theorem that shows the continuum corrected p-value is a valid p-value for the continuum of multiple comparisons.

Theorem 1. *Let*

$$p_0 = \inf_{x \in C} p(x)$$

be the smallest p-value for which H_0 is true. Then

$$\mathrm{pr}\{r(p_0) \leq \alpha\} \leq \alpha.$$

Proof. Define $p_0^* = \inf_{x \in C} p^*(x)$, and let F^* be the permutation cumulative distribution function of p_0^*. Since $C \subset D(p_0)$,

$$\begin{aligned} r(p_0) &\geq \mathrm{pr}^* \left\{ \inf_{x \in C} p^*(x) \leq p_0 \right\} \\ &= \mathrm{pr}^*(p_0^* \leq p_0) \\ &= F^*(p_0) \end{aligned}$$

which implies

$$\mathrm{pr}\{r(p_0) \leq \alpha\} \leq \mathrm{pr}\{F^*(p_0) \leq \alpha\}.$$

By permutation pivotality,

$$\mathrm{pr}\{F^*(p_0) \leq \alpha\} = \mathrm{pr}\{F^*(p_0^*) \leq \alpha\}.$$

Since pr^* is the conditional permutation distribution given the original data

$$\text{pr}\{F^*(p_0^*) \leq \alpha\} = E[\text{pr}\{F^*(p_0^*) \leq \alpha | Y(x), x \in D\}] = E[\text{pr}^*\{F^*(p_0^*) \leq \alpha\}],$$

and since F^* is the (conditional) cumulative distribution function of the random variable p_0^* (given the data), it follows that $F^*(p_0^*)$ is stochastically larger than a uniform $[0, 1]$ random variable, so that $\text{pr}^*\{F^*(p_0^*) \leq \alpha\} \leq \alpha$, (Casella and Berger, 2002). Therefore,

$$\text{pr}\{r(p_0) \leq \alpha\} \leq E[\text{pr}^*\{F^*(p_0^*) \leq \alpha\}] \leq \alpha,$$

as was claimed. □

To see how the previous result can be used, consider

$$R(\alpha) = \{x : p(x) < \inf[p(x') : r\{p(x')\} \geq \alpha]\}.$$

Then by the previous theorem,

$$\text{pr}\{R(\alpha) \cap C \neq \emptyset\} \leq \alpha.$$

Thus, if we reject $H_0(x)$ for all $x \in R(\alpha)$, the family-wise error rate will be less than or equal to α.

Now we consider the behavior of $r_m(\theta)$ as $m \to \infty$. Since $D_1(\theta) \subset \ldots \subset D_m(\theta) \subset D(\theta)$, we have that $r_m(\theta)$ is nondecreasing and bounded by $r(\theta)$, so there is a limit

$$r_\infty(\theta) = \lim_{m \to \infty} r_m(\theta) \leq r(\theta). \tag{2}$$

Ideally we would have $r_\infty(\theta) = r(\theta)$, but in fact $r_\infty(\theta) < r(\theta)$ is possible. Because $\cup_m D_m$ is dense in D and all p-value functions are assumed to be continuous, it is easy to see that $\lim_m \text{pr}^*\{\inf_{x \in D_m(\theta)} p^*(x) < \theta\} =$

$\mathrm{pr}^*\{\inf_{x\in D(\theta)} p^*(x) < \theta\}$. However, $\lim_m \mathrm{pr}^*\{\inf_{x\in D_m(\theta)} p^*(x) = \theta\}$ may be strictly less than $\mathrm{pr}^*\{\inf_{x\in D(\theta)} p^*(x) = \theta\}$. The following result shows that with a minor modification, we can achieve a bracketing of the continuum corrected p-value through limits of the discrete corrected p-values.

Theorem 2. *Let*

$$r_m(\theta,\epsilon) = \mathrm{pr}^*\left\{\inf_{x\in D_m(\theta-\epsilon)} p^*(x) \le \theta+\epsilon\right\}$$

Then for every $\epsilon > 0$, $r_m(\theta,\epsilon)$ converges to a limit $r_\infty(\theta,\epsilon)$ as $m\to\infty$, and

$$r_\infty(\theta) \le r(\theta) \le \lim_{\epsilon\downarrow 0} r_\infty(\theta,\epsilon).$$

Proof. For each fixed θ and ϵ, $r_m(\theta,\epsilon)$ is a nondecreasing sequence in m and converges to some limit $r_\infty(\theta,\epsilon)$ as in the derivation of (2). Note that $r_m(\theta,\epsilon)$ is a nondecreasing function of ϵ for fixed m and θ, and so $r_\infty(\theta,\epsilon)$ is a nondecreasing function of ϵ for fixed θ. In particular, $\lim_{\epsilon\downarrow 0} r_\infty(\theta,\epsilon)$ exists. The theorem will follow once we show that for every $\epsilon > 0$,

$$r(\theta) \le r_\infty(\theta,\epsilon). \tag{3}$$

By continuity of $p(x)$, for $x \in D(\theta)$ there is a neighborhood of x such that $p(x') \ge \theta - \epsilon$ for all x' in the neighborhood. Since $\bigcup_m D_m$ is dense in D, it follows that $D(\theta)$ is contained in the closure of $\bigcup_m D_m(\theta-\epsilon)$.

Suppose $p^*(x)$ satisfies $\inf_{x\in D(\theta)} p^*(x) \le \theta$. Then since each $p^*(x)$ is continuous the infimum is achieved on the compact set $D(\theta) = p^{-1}([\theta,1])$, and at any such point there is an neighborhood such that $p^*(x) \le \theta+\epsilon$ on that neighborhood. By the aforementioned inclusion of $D(\theta)$ in the closure of $\bigcup_m D_m(\theta-\epsilon)$ and the fact that the sequence D_m is increasing, we conclude that for all m sufficiently large $\inf_{x\in D_m(\theta-\epsilon)} p^*(x) \le \theta+\epsilon$. This inclusion argument establishes (3). □

Theorem 2 suggests one strategy for trying to approximate the continuum corrected p-values. If one is interested in a family-wise error rate of 0.05, then consider a very fine grid of points D_m and compute, say, $\mathrm{pr}^*\{\inf_{x \in D_m(0.04)} p^*(x) \leq 0.06\}$. This difficulty has arisen because we have insisted on using conservative p-values. For any finite D_m, the Westfall-Young randomization method still gives valid multiple comparison corrections on that set, and extending the inference to all of D in some approximate way requires trust that the the behavior of the observed functions is accurately captured by values on D_m. We have shown that in fact the corrected p-values do converge to a meaningful limit which is almost the continuum corrected p-value, unlike the Bonferroni method where the discrete corrected p-values all converge to 1. In practice, we advocate simply computing the discrete corrected p-values using the method of Westfall and Young, and taking the leap of faith that when one observes a number of contiguous grid points where there the corrected p-values are below 0.05, then there is statistical significance throughout the entire region.

3.2 One Step Methods

Here we consider a simpler procedure than the one analyzed above. The classical Bonferroni method is a one-step method: simply reject all null hypotheses for which the p-value is less than or equal to α/m where m is the total number of null hypotheses. This will give a family-wise error rate which is less than or equal to α. Holm's method is a stepwise modification which still controls the family-wise error rate but has greater power. However, Holm's method is slightly more difficult to apply since it requires sorting the p-values and comparing them with different critical values.

A one-step method based on randomization is the following. For each randomization, compute the minimal p-value among all the null hypotheses. For each p-value from the real data set, a corrected p-value would be the proportion of randomized minimal p-values which are less than or equal to the observed p-value. One would reject the corresponding null hypothesis if this corrected p-values is less than or equal to α.

Define continuum and discrete one-step corrected p-values as before:

$$\begin{aligned} q(\theta) &= \mathrm{pr}^* \left\{ \inf_{x \in D} p^*(x) \leq \theta \right\}, \\ q_m(\theta) &= \mathrm{pr}^* \left\{ \inf_{x \in D_m} p^*(x) \leq \theta \right\}. \end{aligned}$$

Clearly $q(\theta) \geq r(\theta)$, so the result of Theorem 1 holds:

$$\mathrm{pr}\left\{ q(p_0) \leq \alpha \right\} \leq \alpha,$$

where p_0 is the minimal p-value over the unknown set of all true null hypotheses. Thus, if we reject $H_0(x)$ only for x such that $q\{p(x)\} \leq \alpha$, then the family-wise error rate will be less than or equal to α. However, we will only reject a subset of the null hypotheses that are rejected using the stepwise procedure of the previous section.

Next we turn to the limit of $q_m(\theta)$ as $m \to \infty$. For each θ, $q_m(\theta)$ is nondecreasing in m, and are bounded above by $q(\theta)$, so the limit exists and

$$q_\infty(\theta) = \lim_{m \to \infty} q_m(\theta) \leq q(\theta).$$

As in the previous section, we may have $q_\infty(\theta) < q(\theta)$, e.g. a particular permutation may have $\inf_{x \in D} p^*(x) = \theta$ and the infimum not achieved for any point $x \in \bigcup_m D_m$ so that for for this permutation, $\inf_{x \in D_m} p^*(x) > \theta$ for all m. Thus, we must still be satisfied with a result like

$$q_\infty(\theta) \leq q(\theta) \leq \lim_{\epsilon \downarrow 0} \lim_{m \to \infty} \mathrm{pr}^* \left\{ \inf_{x \in D_m} p^*(x) \leq \theta + \epsilon \right\}.$$

4 Simulation Study

Now we present a small simulation study of the method of Westfall and Young applied to functional data. We wrote MATLAB procedures to perform the calculations, but we could have as well used the `mt.minP` function in `multtest` library of R. Each example consists of two independent samples of functional data simulated from a Gaussian process with the same covariance. In each example, the first sample has zero mean and the second has a nonzero mean function.

To obtain zero mean Gaussian functional data, we generate 1,400 independent and identically distributed random normal variates with zero mean and standard deviation $\sigma = 0.01$. These white noise values are paired with x grid points equally spaced between -0.2 and 1.2 and input to a spline smoother with smoothing parameter $\lambda = 0.95$. We removed 200 points at each end to obtain 1,000 equally spaced x values in $[0, 1]$. This approach yields smooth curves that do not have boundary artifacts from the spline smoothing. Note however that the variance function is still somewhat larger near the boundary than in the middle.

In each example, we generate two samples of of functional data with 250 observations each. The first sample has zero mean function $\mu_1(x) \equiv 0$, where $0 \leq x \leq 1$. The second sample has nonzero mean function $\mu_2(x)$. The formulae for μ_2 for each example is presented in Table 1 and their graphs are presented in Figure 1. In each example, we first compute uncorrected p-values (see Figure 2) using a two sample t-test at each of the 1,000 grid points. Both Holm's method and the method of Westfall and Young are applied with two levels of discretization: $m = 1,000$ (the original grid where the functional data were computed) and a subgrid of $m = 50$ equally spaced

Table 1: A list of four mean functions μ_2. The function $\text{beta}_{a,b}(x)$ is the density of the Beta(a,b) distribution. The amplitudes have been adjusted to allow for sufficient power to detect some difference.

Name	$\mu_2(x), \quad 0 \leq x \leq 1$
beta spike	$0.00001\text{beta}_{1000,1000}(x)$
beta	$0.0001\text{beta}_{5,5}(x)$
constant	0.00015
linear	$0.0004(x - 0.5)$

x values. The corrected p-values are presented in Figures 3 and 4. We also include the results of the one-step method in the second figure for both the $m = 50$ and $m = 1,000$ levels of discretization. For Holm's method, we see that there are some statistically significant (at the $\alpha = 0.05$ level) results for the first three examples at the coarser $m = 50$ grid, but these become nonsignificant as the grid is refined to $m = 1,000$. For the method of Westfall and Young, note that the corrected p-values for $m = 50$ are very close to but slightly smaller than those for $m = 1,000$, as would be expected from the analysis presented above. For $m = 50$, Westfall and Young's method is generally more powerful than Holm's method, and retains this power when the grid is refined. The one-step method is generally quite close the the full version of the method of Westfall and Young. For the constant alternative, the corrected p-value function for the one-step method differs to some extent but mainly in the region where results are not significant. In all cases $10,000$ permutations were used to compute the corrected p-values for the method of Westfall and Young and the one-step method.

Looking at the individual examples, we see that the Westfall and Young

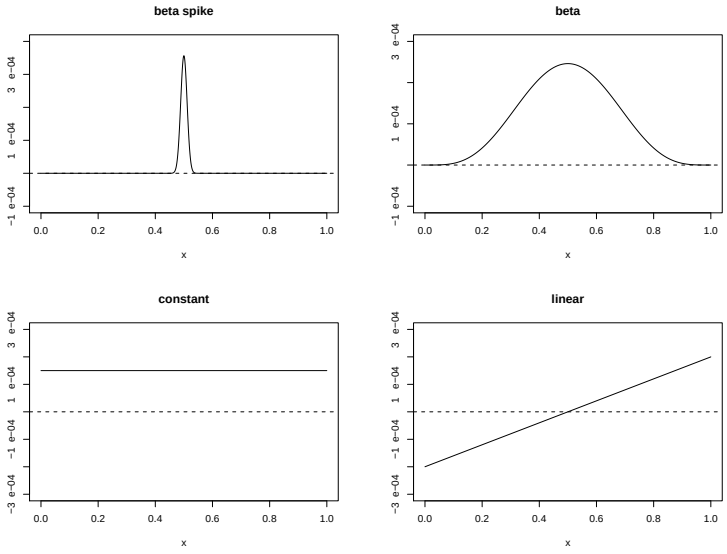

Figure 1: Plot of all four choices of μ_2 from Table 1. The dashed line is a horizontal line at 0.

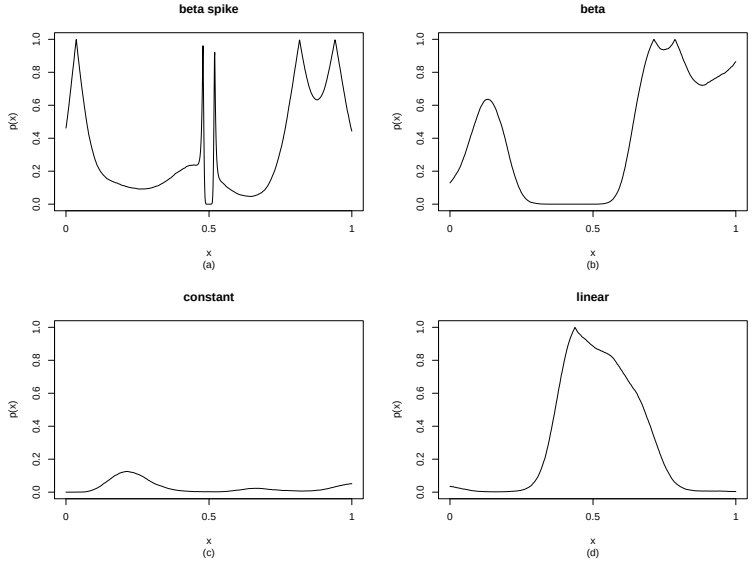

Figure 2: Simulation results: The uncorrected p-values.

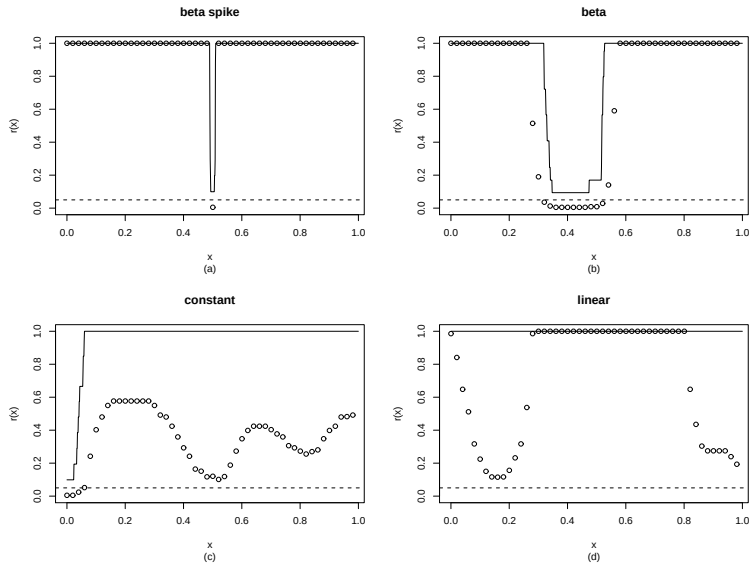

Figure 3: Simulation results: The corrected p-values from Holm's procedure. The larger circles correspond to the $m = 50$ and the solid curve to $m = 1,000$. The horizontal slashed lines show the level 0.05.

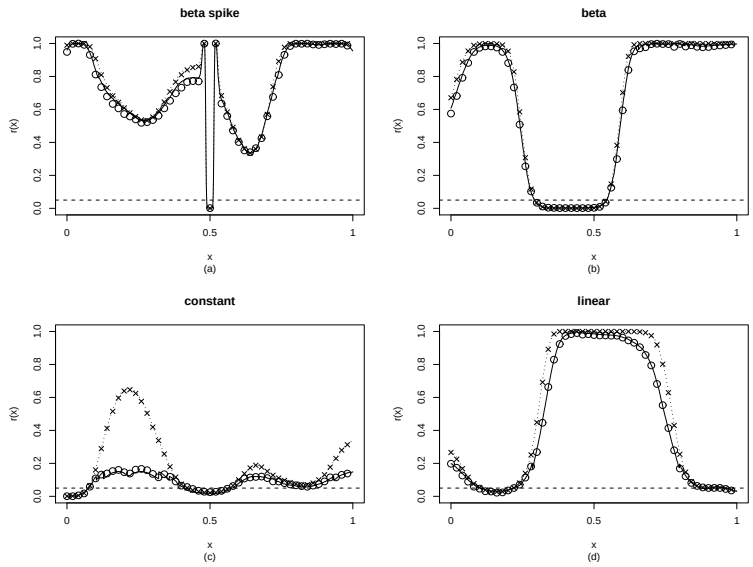

Figure 4: Simulation results: The corrected p-values from the Westfall-Young randomization method and the one-step method. The circles show the corrected p-values for $m = 50$ evaluation points using the full step-down method and the X's show the corrected p-values for $m = 50$ evaluation points using the one-step method. The curves show the corrected p-values for $m = 1,000$ evaluation points. Note that the curves almost interpolate the plotting points for the corresponding one-step or full step-down method with $m = 50$. The horizontal slashed lines show the level 0.05.

method does a good job of picking out the region of nonzero values for the beta spike. Its performance on the other examples is not as good, although this is dependent on the particular amplitudes chosen for each μ_2. In particular, for the constant and linear choices of μ_2 we see that the region of significant difference does not accurately indicate the nature of the deviation in the two functions.

5 Application

Here we apply the method analyzed above to real data. The application concerns a medical device designed to detect pre-cancerous cervical lesions by illuminating the cervix with monochromatic light at various excitation wavelengths and measuring the fluorescence at a number of emission wavelengths. Further details may be found in Cox, et. al. (2003). We consider measurements made by a single device for several different time periods. Specifically, there are seven time periods when different factors were used to correct for the optical and electronic transfer function. For this objective, we will apply the multiple comparison methodology above to measurements on a fluorescence standard (coumarin) made at different time points. The correction factors were computed using other standards measurements. Our null hypothesis is that the measurements made during the different periods are indistinguishable. Thus, permuting the measurement amongst the different periods would make no difference under the null hypothesis. We use the measured intensities at 390 nanometers excitation (which is near the peak response for coumarin) at emission wavelengths between 445 and 550 nanometers in increments of 1 nanometer. There are 433 curves in total, and the sample sizes for each of the seven time periods are 80, 35, 68, 88,

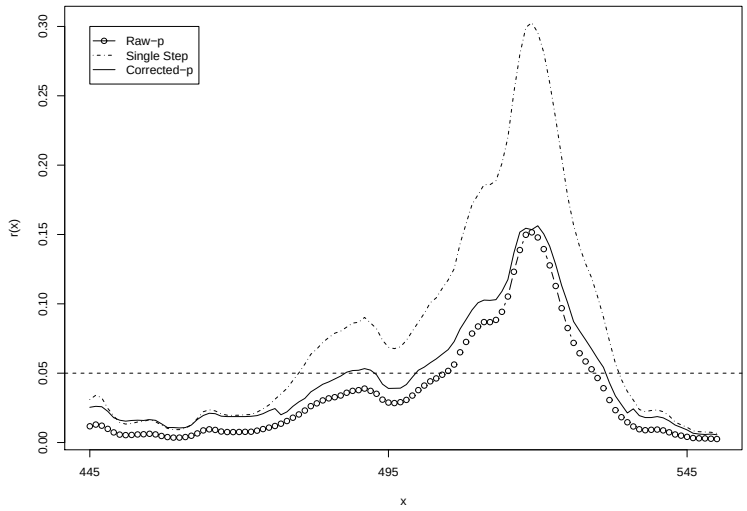

Figure 5: Plots of raw and corrected p-values. A dotted horizontal line at 0.05 is drawn for a reference.

32, 39, and 91. The analysis of variance statistic was computed pointwise and converted to a p-value using the F-distribution. We then applied the Westfall-Young randomization method to correct the p-values, including the one-step and the full step down corrections. The results are shown in Figure 5. We note that raw and Westfall-Young corrected p-values seems to be close to each other, while the single step corrected p-values are quite conservative in this case.

We examine the results in more detail to assess the nature of the differences. Figure 6 shows a plot of the grand mean, and the root mean squares due to treatment and error. Of course, the analysis of variance F-statistic is the square of the ratio of these latter two quantities. For much the domain of emission wavelengths, the root mean square due to treatment is about

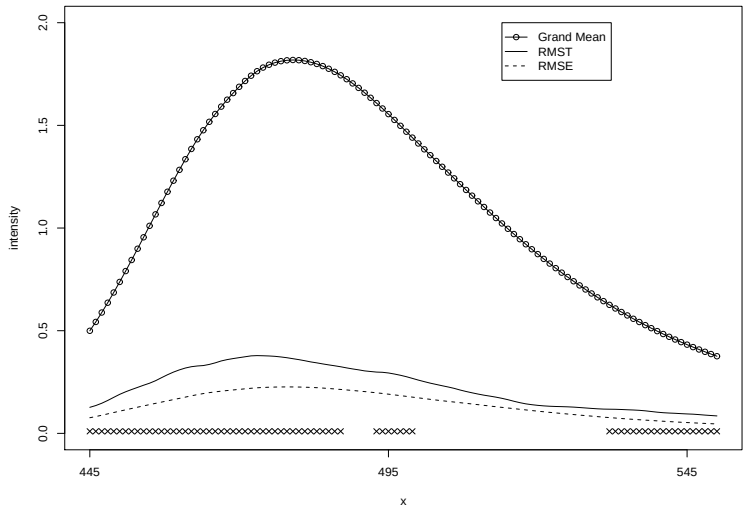

Figure 6: Plot of grand mean, along with root mean square due to treatment (RMST) and root mean squared error (RMSE). The region of significance given by the Westfall-Young randomization method are indicated by the crosses on the horizontal line at 0.

20% of the overall signal, while the "noise" as measured by the root mean square due to error is somewhat smaller. Such a large systematic difference has practical significance for the intended application of this technology.

One difficulty is that the approach does not lend itself to pairwise comparisons at each point. However, we can informally assess which periods are different. In Figure 7 we plot the difference between the grand mean and the mean within each period, normalized by the average of the grand mean over the 106 emission wavelengths. The time periods are numbered in order. It is clear that for most of the domain, the means of periods 2 and 4 are somewhat lower than the others, and there is not an obvious pattern

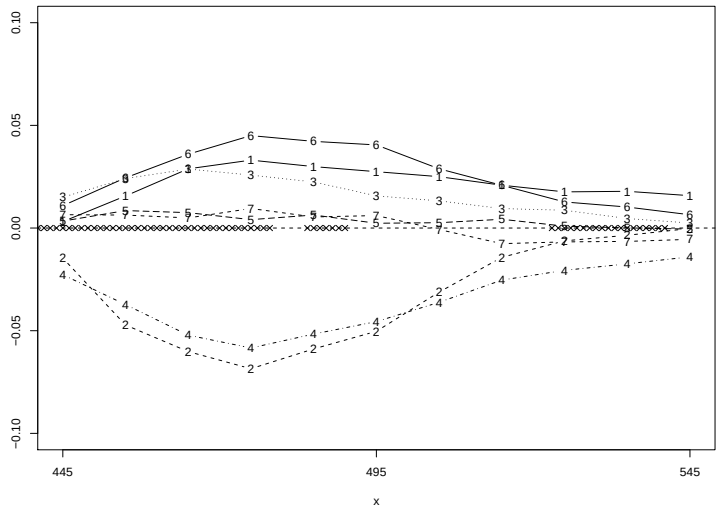

Figure 7: Plot of scaled deviation from grand mean. All curves are scaled by the average of the grand mean over the domain. The region of significance given by the Westfall-Young randomization method are indicated by the crosses on the horizontal lines at 0.

with respect to time order. While we cannot rule out the possibility that the differences are due to some other time trend rather than deficiencies in the correction factors, the lack of such a pattern does not support this.

6 Conclusions

We have shown that the Westfall and Young randomization method applied to functional data works well even as the grid of evaluation points is refined so that the number of pointwise tests becomes large. The method is probably not as powerful as an overall test for most alternatives. However, the results from such pointwise testing are more directly interpretable than the

results of an overall test which finds some difference in the means (or other parameters) but may give little indication of where the difference is. Application of the method of Westfall and Young only requires implementation of pointwise univariate tests and does not need any modeling of covariance or the joint distributions, as is required for an overall test.

There may be some criticisms raised with the present work. One that has already become evident in the application in Section 5 is that when comparing several populations, we can use the methodology to determine where there is a difference in the means, but not which populations are different. Another issue is the relevance of the family-wise error rate with the advent of false-discovery rate methods which have attracted much attention (Benjamini and Hochberg, 1995), especially when the family-wise error rate tends to be somewhat more conservative than the false discovery rate. We argue that there still is a place for multiple comparison procedures controlling for family-wise error rate. Westfall (2005) shows that a method that controls for the false discovery rate can be too liberal and thus its inference may be misleading in some applications. Furthermore, we could not find a method comparable to that of Westfall and Young which applies to functional data controlling for the false discovery rate, although the methods presented in Efron (2007) may prove useful for solving this problem.

One of the surprising things that emerged from our simulation examples was that the one-step method performed almost as well as the full step down method of Westfall and Young in many cases. However, in the real data example, the one step method was somewhat more conservative. We note that calculation of the full version is more complicated than for the one-step method but requires little additional computer time since most of the computation is in getting the p-values.

The widespread applicability of the method of Westfall and Young suggests other approaches, such as including some principal component scores as well as function values on a grid of evaluation points. With the additional variables, the power at alternatives (such as the "constant" example of Section 4) may be improved. This can be another interesting future avenue of research.

Acknowledgements: This research was supported by the National Science Foundation and by the National Cancer Institute. We would like to thank the editor and the referees for helpful comments.

References

[1] Casella, G. and Berger, R. (2002), Statistical Inference (Duxbury, Pacific Grove, CA, 2nd ed.).

[2] Cox D.D., Chang S.K., Dawood M.Y., Staerkel G., Utzinger U., Richards-Kortum R., and Follen M. (2003) Detecting a signal from the menstrual cycle in fluorescence spectroscopy of the cervix, *Applied Spectroscopy*, **57**, 67-72.

[3] Efron, B. (2007) Correlation and large-scale simultaneous significance testing, *J. Amer. Statist. Assoc.*, **102**, 93-103.

[4] Fan, J. and Lin, S. (1998), Test of significance when data are curves, *J. Amer. Statist. Assoc.*, **93**, 1007-1021.

[5] Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. Royal Statist. Society, Sers. B*, **57**, 289-300.

[6] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* **6**, 65-70.

[7] Hoyte L., Schierlitz L., Zou K., Flesh G., Fielding J.R. (2001) Two- and 3-dimensional MRI comparison of levator ani structure, volume, and integrity in women with stress incontinence and prolapse, *American Journal of Obstetrics and Gynecology*, **185**, 11-19.

[8] Ramsay, J.O. and Silverman, B.W. (2005), Functional Data Analysis (Springer, New York, 2nd ed.).

[9] Rencher, A.C. (2002), Methods of Multivariate Analysis (Wiley, New York, 2nd ed.).

[10] Shen, Q. and Faraway, J. (2004), An F test for linear models with functional responses, *Statistica Sinica*, **14**, 1239-1257.

[11] Taylor, J.E., Worsley, K.J., Gosselin, F. (2007) Maxima of discretely sampled random fields with an application to 'bubbles,' *Biometrika*, **94**, 1-18.

[12] Westfall, P. (2005). Comment on Benjamini and Yekutieli, 'False discovery rate adjusted confidence intervals for selected parameters,' *J. Amer. Statist. Assoc.*, **100**, 85-89.

[13] Westfall, P. and Young S. (1993), Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment (Wiley, New York).

[14] Zerbe, G.O. and Murphy, J.R. (1986), On multiple comparisons in randomization analysis of growth and response curves, *Biometrics*, **42**, 795-804.