Pointwise testing with functional data using the Westfall–Young randomization method

BY DENNIS D. COX
Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A.
dcox@stat.rice.edu

AND JONG SOO LEE
Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.
jslee@stat.cmu.edu

SUMMARY
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 & 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.

Some key words: 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, store the coefficients of the expansion and then function values can be computed as needed. Assume that 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 & 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 becomes larger.

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 (Zerbe & Murphy, 1986; Fan & Lin, 1998; Shen & Faraway, 2004). Just as with ordinary analysis of variance and multivariate analysis of variance, after we decide that the means from 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 & Young (1993, Chapter 2). 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, such as Bonferroni correction, 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 comparisons, 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 related work that can possibly be used to extend our results. 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 familywise 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 that are suitable for Gaussian processes. This method does not require a permutation pivotality condition which is needed for our method.

2. Example: The Two-Sample Case

Suppose we have two independent samples $y_{ij}(x)$, with $i = 1, 2$ and $j = 1, \ldots, n_i$, where $i$ denotes the population from which the observation is drawn and $j$ indexes the observation within the sample. Assume that the observations within the samples are independent and identically distributed. We assume that 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 $x \in D$, against the general alternative $H_1 : \mu_1(x) \neq \mu_2(x)$, for at least one $x \in D$. Consider the pointwise testing problems $H_0(x) : \mu_1(x) = \mu_2(x)$ against $H_1(x) : \mu_1(x) \neq \mu_2(x)$. This is a simple univariate two-sample problem for each $x$. As a result 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 $x_k$, $k = 1, \ldots, u$. To apply the method analyzed here, first a $p$-value $p(x_k)$ is computed using the univariate two sample $t$-test of the null hypothesis $H_0(x_k)$ at each grid point $x_k$, and then a multiple comparison
procedure is applied that controls the familywise error rate. To define the familywise error rate, consider \( C_m = \{x_k : H_0(x_k) \text{ is true}, \ k = 1, \ldots, m \} \), which is the set of grid points for which \( H_0(x_k) \) is true. Then the familywise error rate is \( \Pr \{ \text{reject } H_0(x_j) \text{ for any } x_j \in C_m \} \), which depends on the set of grid points. Our goal is to make the familywise error rate less than or equal to \( \alpha \), where \( \alpha \) is given, no matter what is the set \( C_m \) of true null hypotheses.

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, where, for the familywise error rate to be less than or equal to \( \alpha \), 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 \( \alpha \). 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 that 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, \( p_1 \leq \cdots \leq p_m \). 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 \( H_0(x^\prime) \) are highly positively correlated when \( x \) and \( x^\prime \) are close. Thus, we need an approach that 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 non-normal 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 common 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 \( \{y_{ij}(x_k) : i = 1, 2, j = 1, \ldots, n_i, k = 1, \ldots, m \} \). Then for any subset \( C_m \) for which the null hypotheses are true, the joint distribution \( \mathcal{L}[y_{ij}(x_k) : i = 1, 2, j = 1, \ldots, n_i, x_k \in C_m] \) 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 Holm’s method: \( p_1 \leq \cdots \leq p_m \). Let \( \pi \) be the permutation that maps the original grid sequence on to the sorted \( p \)-value sequence, such that \( p_{(j)} = p(x_{\pi(j)}) \). 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 by 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 \( N \), to obtain 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 \( \ell = 1, \ldots, N \). The corrected \( p \)-value
corresponding to \( p_j \) is the proportion of the \( q_{(j),l}^* \) less than or equal to \( p_j \); that is,

\[
r_j = N^{-1} \sum_{l=1}^{N} I\{q_{(j),l}^* \leq p_j\},
\]

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 \( j^+ \) be given by \( p(x_j^+) = \min\{p(x_j) : x_j \in C_m\} \). It has been shown that the method controls the familywise error rate to be less than or equal to \( \alpha \), since \( \Pr(r_j \leq \alpha) \leq \alpha \) (Westfall & Young, 1993, Chapter 2). 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 \( p \)-value for the continuum of pointwise null hypotheses \( H_0(x), x \in D \). The correspondingly corrected \( p \)-values under Bonferroni’s or Holm’s method will converge to unity 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 that the Westfall–Young randomization method is appropriate for functional data.

3. Theoretical results

3.1. Main result

We now assume a very general set-up. Denote the functional data by \( y_k(x), x \in D \), where \( k \) may be a vector index; for example, \( k = (i, j) \), with \( i = 1, 2 \), and \( j = 1, \ldots, n_i \), for a two-sample set-up. 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 \( H_0(x) \) concerning \( \mathcal{L}(Y(x)) \). There will also be a corresponding alternative hypothesis \( H_1(x) \), but it will generally not be mentioned. We suppose further that there is a test that can be summarized as a \( p \)-value, and \( p(x) \) will denote the \( p \)-value function from the original data. The following assumptions will be made.

**Assumption 1.** As a function of \( x \), \( p(x) \) is continuous. Similarly, any \( p \)-value function computed from permuted data is continuous.

**Assumption 2.** The domain \( D \) is compact.

**Assumption 3.** The grid points for evaluation form an increasing sequence \( D_m, m = 1, 2, \ldots \) of finite subsets of \( D \) such that \( \bigcup_{m=1}^{\infty} D_m \) is dense in \( D \).

**Assumption 4.** The following permutation pivotality condition holds. Let \( C = \{x \in D : H_0(x) \text{ is true}\} \), and assume that \( \mathcal{G} \) 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.

Let \( p^*(x) \) be the \( p \)-value function computed from the permuted data obtained by selecting an element of \( \mathcal{G} \) at random. Let \( \Pr^* \) be the probability measure associated with the random \( p^* \)'s. More precisely, \( \Pr^*(A) \) 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
\[ D(\theta) = \{ x \in D : p(x) \geq \theta \}, \quad D_m(\theta) = \{ x \in D_m : p(x) \geq \theta \}, \]
and let
\[ r(\theta) = \text{pr}^* \left\{ \inf_{x \in D(\theta)} p^*(x) \leq \theta \right\}, \quad r_m(\theta) = \text{pr}^* \left\{ \inf_{x \in D_m(\theta)} p^*(x) \leq \theta \right\}. \]

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 that 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
\[ \text{pr}[r(p_0) \leq \alpha] \leq \alpha. \]

**Proof.** Define \( p_0^* = \inf_{x \in C} p^*(x) \), and let \( F^* \) be the permutation cumulative distribution function of \( p_0^* \). Since \( C \subset D(p_0) \), \( r(p_0) \geq \text{pr}^* \{ \inf_{x \in C} p^*(x) \leq p_0 \} = \text{pr}^* (p_0^* \leq p_0) = F^*(p_0) \), which implies that
\[ \text{pr}[r(p_0) \leq \alpha] \leq \text{pr}[F^*(p_0) \leq \alpha]. \]

By permutation pivotality,
\[ \text{pr}[F^*(p_0) \leq \alpha] = \text{pr}[F^*(p_0^*) \leq \alpha]. \]

Since \( \text{pr}^* \) 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 \( \text{Un}[0, 1] \) random variable, so that \( \text{pr}^*[F^*(p_0^*) \leq \alpha] \leq \alpha \) (Casella & Berger, 2002). Therefore,
\[ \text{pr}[r(p_0) \leq \alpha] \leq E[\text{pr}^*[F^*(p_0^*) \leq \alpha]] \leq \alpha, \]
as was claimed. \( \square \)

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 familywise error rate will be less than or equal to \( \alpha \).

Now we consider the behaviour of \( r_m(\theta) \) as \( m \to \infty \). Since \( D_1(\theta) \subset \cdots \subset D_m(\theta) \subset D(\theta) \), we have that \( r_m(\theta) \) is non-decreasing and bounded by \( r(\theta) \), so that there is a limit,
\[ r_\infty(\theta) = \lim_{m \to \infty} r_m(\theta) \leq r(\theta). \]

Ideally we would have \( r_\infty(\theta) = r(\theta) \), but in fact \( r_\infty(\theta) < r(\theta) \) is possible. Since \( \bigcup_m D_m \) is dense in \( D \) and all \( p \)-value functions are assumed to be continuous, it is...
easy to see that \( \lim_{m \to \infty} \text{pr}^* [\inf_{x \in D_m(\theta)} p^*(x) < \theta] = \text{pr}^* [\inf_{x \in D(\theta)} p^*(x) < \theta] \). However, 
\( \lim_{m \to \infty} \text{pr}^* [\inf_{x \in D_m(\theta)} p^*(x) = \theta] \) may be strictly less than \( \text{pr}^* [\inf_{x \in D(\theta)} p^*(x) = \theta] \). 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

\[
\tilde{r}_m(\theta, \epsilon) = \text{pr}^* \left\{ \inf_{x \in D_m(\theta-\epsilon)} p^*(x) = \theta + \epsilon \right\},
\]

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

\[
r_\infty(\theta) \leq r(\theta) \leq \lim_{\epsilon \downarrow 0} r_\infty(\theta, \epsilon)\]  

**Proof.** For each fixed \( \theta \) and \( \epsilon \), \( r_m(\theta, \epsilon) \) is a non-decreasing 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 non-decreasing function of \( \epsilon \) for fixed \( m \) and \( \theta \), and so \( r_\infty(\theta, \epsilon) \) is a non-decreasing function of \( \epsilon \) for fixed \( \theta \). In particular, \( \lim_{\epsilon \downarrow 0} r_\infty(\theta, \epsilon) \) exists. The theorem will follow once we show that, for every \( \epsilon > 0 \),

\[
r(\theta) \leq r_\infty(\theta, \epsilon).
\]

By continuity of \( p(x) \), for \( x \in D(\theta) \) there is a neighbourhood of \( x \) such that \( p(x') \geq \theta - \epsilon \) for all \( x' \) in the neighbourhood. 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 \( \inf_{x \in D(\theta)} p^*(x) \leq \theta \). 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 a neighbourhood such that \( p^*(x) \leq \theta + \epsilon \) on that neighbourhood. 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, \( \inf_{x \in D_m(\theta - \epsilon)} p^*(x) \leq \theta + \epsilon \). This inclusion argument establishes (3). \( \square \)

Theorem 2 suggests one strategy for trying to approximate the continuum-corrected \( p \)-values. If one is interested in a familywise error rate of 0.05, then consider a very fine grid of points \( D_m \) and compute, say, \( \text{pr}^* [\inf_{x \in D_m(0.04)} p^*(x) \leq 0.06] \). This difficulty has arisen because we have insisted on using conservative \( p \)-values, which is a consequence of using a permutation-based test. In practice, we advocate simply computing the discrete corrected \( p \)-values using the method of Westfall and Young, and when one observes a number of contiguous grid points where the corrected \( p \)-values are below 0.05, then there is statistical significance throughout the entire region. This is justified by the results presented in this section and will be demonstrated in \( \S \) 4.

### 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 \( \alpha / m \), where \( m \) is the total number of null hypotheses. Holm’s method is a stepwise modification which still controls the familywise error rate. The Westfall–Young method is a stepwise method based on randomization.

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 that are less than or equal to the observed \( p \)-value. One would reject the corresponding null hypothesis if this corrected \( p \)-value were less than or equal to \( \alpha \).

Define continuum and discrete one-step corrected \( p \)-values as before, \( q(\theta) = \text{pr}^* [\inf_{x \in D} p^*(x) \leq \theta] \), \( q_m(\theta) = \text{pr}^* [\inf_{x \in D_m} p^*(x) \leq \theta] \). Clearly \( q(\theta) \geq r(\theta) \), so the result of
Fig. 1. Plots of all four choices of $\mu_2$ from Table 1: (a) beta spike, (b) beta, (c) constant, (d) linear. The dashed line is a horizontal line at 0.

Table 1. A list of four mean functions $\mu_2(x)$, for $0 \leq x \leq 1$. The function $\text{be}_{a,b}(x)$ is the density of the $\text{Be}(a, b)$ distribution. The amplitudes have been adjusted to allow for sufficient power to detect some difference

<table>
<thead>
<tr>
<th>Name</th>
<th>$\mu_2(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta spike</td>
<td>$0 \cdot 00001 \text{be}_{1000,1000}(x)$</td>
</tr>
<tr>
<td>Beta</td>
<td>$0 \cdot 0001 \text{be}_{5,5}(x)$</td>
</tr>
<tr>
<td>Constant</td>
<td>$0 \cdot 00015$</td>
</tr>
<tr>
<td>Linear</td>
<td>$0 \cdot 0004(x - 0.5)$</td>
</tr>
</tbody>
</table>

Theorem 1 holds, $\Pr(q(p_0) \leq \alpha) \leq \alpha$, 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 familywise error rate will be less than or equal to $\alpha$. 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 $\theta$, $q_m(\theta)$ is non-decreasing in $m$, and is bounded above by $q(\theta)$, so that 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 $q_\infty(\theta) < q(\theta)$; for example, 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,
4. Simulation Study

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. Each example consists of two independent samples of functional data simulated from a Gaussian process with the same covariance structure. In each example, the first sample has zero-mean and the second has a non-zero mean function.

To obtain zero-mean Gaussian functional data, we generate 1400 independent and identically distributed random normal variates with mean zero 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 fed into a spline smoother with smoothing parameter $\lambda = 0.95$. We removed 200 points at each end to obtain 1000 equally spaced $x$ values in $[0, 1]$. This approach yields smooth curves that do not have boundary artifacts from the spline smoothing.

In each example, we generate two samples of functional data with 250 observations each. The first sample has mean function $\mu_1(x) \equiv 0$, where $0 \leq x \leq 1$. The second sample has non-zero

Fig. 2. Simulation results. The uncorrected $p$-values for (a) beta spike, (b) beta, (c) constant, (d) linear.
mean function $\mu_2(x)$. The formula for $\mu_2$ for each example is presented in Table 1 and their graphs are presented in Fig. 1. In each example, we first compute uncorrected $p$-values using a two-sample $t$-test at each of the 1000 grid points; see Fig. 2. Then Holm’s method and the method of Westfall and Young are applied with two levels of discretization: $m = 1000$, corresponding to the original grid at which the functional data were computed, and a subgrid of $m = 50$ equally spaced $x$-values. The corrected $p$-values are presented in Figs. 3 and 4. We also include the results of the one-step method in Fig. 4 for both the $m = 50$ and $m = 1000$ levels of discretization. For Holm’s method, some results are statistically significant, at the $\alpha = 0.05$ level, for the first three examples at the coarser $m = 50$ grid, but these become non-significant as the grid is refined to $m = 1000$. For the method of Westfall and Young, the corrected $p$-values for $m = 50$ are very similar to but slightly smaller than those for $m = 1000$, 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 similar to 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.

For the individual examples, the Westfall and Young method does a good job of picking out the region of non-zero 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 $\mu_2$. In particular,
for the constant and linear choices of $\mu_2$, the region of significant difference does not accurately indicate the nature of the deviation in the two functions.

5. Application

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 seven time periods when different factors were used to correct for the optical and electronic transfer function. For this objective, we will apply the above multiple comparison methodology to measurements on a fluorescence standard, coumarin, made at different time points. The correction factors were computed using other standard 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 nm excitation, which is near the peak response for coumarin, at emission wavelengths between 445 and 550 nm in increments of 1 nm. There are 433 curves in total, and the sample sizes for each of the seven time periods are
80, 35, 68, 88, 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 Fig. 5. The raw and Westfall–Young corrected $p$-values seem to be similar 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. The analysis of variance $F$-statistic is the square of the ratio of these latter two quantities. For much of the domain of emission wavelengths, the root mean square due to treatment is about 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 Fig. 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 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. DISCUSSION

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 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 modelling of covariance structure 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 § 5 is that, when comparing several populations, we can use the methodology to determine where there is a difference in the means, but cannot determine which populations are different. Another issue is the relevance of the familywise error rate with the advent of false discovery rate methods that have attracted much attention (Benjamini & Hochberg, 1995). However, 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.

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 § 4, may be improved.

ACKNOWLEDGEMENT

This research was supported by the U.S. National Science Foundation and the U.S. National Cancer Institute. We would like to thank Professor D. M. Titterington and the referees for insightful and helpful comments that led to a substantial improvement of the paper.

REFERENCES


[Received January 2007. Revised January 2008]