Communications for Statistical Applications and Methods
2015, Vol. 22, No. 2, 121–135

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.2.121
Print ISSN 2287-7843 / Online ISSN 2383-4757

A Two Sample Test for Functional Data

Jong Soo Lee[1,a], Dennis D. Cox[b], Michele Follen[c]

[a]Department of Mathematical Sciences, University of Massachusetts Lowell, USA
[b]Department of Statistics, Rice University, USA; [c]Department of Obstetrics and Gynecology, Brookdale University Hospital and Medical Center, USA

Abstract

We consider testing equality of mean functions from two samples of functional data. A novel test based on the adaptive Neyman methodology applied to the Hotelling's T-squared statistic is proposed. Under the enlarged null hypothesis that the distributions of the two populations are the same, randomization methods are proposed to find a null distribution which gives accurate significance levels. An extensive simulation study is presented which shows that the proposed test works very well in comparison with several other methods under a variety of alternatives and is one of the best methods for all alternatives, whereas the other methods all show weak power at some alternatives. An application to a real-world data set demonstrates the applicability of the method.

Keywords: Functional data analysis, mean functions comparison, two sample testing, permutation testing, fluorescence spectroscopy.

1. Introduction

We investigate the extensions to Functional Data Analysis (FDA) of the classical one and two sample tests for the population mean. In the FDA setting, an observation is idealized as a function $y(t)$ with t in some domain $\mathcal{T}$, where $y(t)$ may be viewed as a realization of an underlying stochastic process $Y(t)$. We assume each observed function $y : \mathcal{T} \longrightarrow \mathbb{R}$ is a continuous function of t, but in fact the methods investigated here could be applied to high dimensional data of any structure. As is typical in FDA, we also assume that the data have already been smoothed and registered to be put on a common domain (Ramsay and Silverman, 2005).

In the one sample setting, we assume we have a sample $y_1, \ldots, y_n$ which we model as i.i.d. realizations of a stochastic process $Y(t)$. We are interested in making inferences about the mean function

$$\mu(t) = E[Y(t)].$$

We assume that $E[Y^2(t)] < \infty$, so that first and second moments exist. The one sample problem consists of testing the null hypothesis $H_0 : \mu \equiv \mu_0$, where μ_0 is a given function, which without loss of generality may be taken as $\mu_0 \equiv 0$. For the two sample problem, we have observations from two populations (denoted by y_{1i} where $1 \le i \le n_1$ and by y_{2i} where $1 \le i \le n_2$) and we wish to test $H_0 : \mu_1 \equiv \mu_2$ where μ_1 and μ_2 are the mean functions for their respective population. For both testing problems, the alternative is the general one.

The authors gratefully acknowledge the support from NIH-NCI Grant No. PO1-CA82710.

[1] Corresponding author: Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA 01854 USA. Email: jongsoo_lee@uml.edu

Published 31 March 2015 / journal homepage: http://csam.or.kr
© 2015 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

The preferred test in multivariate analysis is Hotelling's T-squared test, based on the test statistic $T^2 = \bar{\mathbf{y}}'\hat{\Sigma}^{-1}\bar{\mathbf{y}}$ for the one sample problem, and $T^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\hat{\Sigma}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$ in the two sample setting. For the one sample setting, $\bar{\mathbf{y}}$ is the sample mean vector and $\hat{\Sigma}$ is the sample covariance matrix (Rencher, 2002). The test statistic in the two sample setting is based on the sample means from both samples, and a "pooled" covariance estimate. These tests can be applied to the vector of values of the functional data: $\mathbf{y} = (y(t_1), \ldots, y(t_p))'$, where $p = \dim(\mathbf{y})$ is the number of points where the functions are evaluated. The problem that arises in FDA is that the dimension p of the data vectors is somewhat arbitrary, and one can easily take $p > n$. Clearly then, $\hat{\Sigma}$ will be singular and the test statistic is undefined. Even if one keeps $p < n$, p is still somewhat arbitrary, and $\hat{\Sigma}$ will typically be poorly conditioned for many values of $p < n$. Thus, the results may depend critically on the choice of p. One of the principles of FDA to which we try to adhere is the Grid Refinement Invariance Principle (GRIP): any procedure for functional data should be insensitive to the dimension of the representation, as long as the dimension is sufficiently large to give an accurate representation (Cox and Lee, 2008). A more formal statement is given below.

A general methodology for constructing tests that typically work well in nonparametric inference problems with single functions is the Adaptive Neyman methodology (Neyman, 1937; Inglot *et al.*, 1994). For most such inference problems, there is no optimal test, but the Adaptive Neyman tests have been shown to work well for many problems in that they have good power against a broad range of alternatives. In our simulation study presented below, we show that this appears to be the case for the Adaptive Neyman tests proposed here in that they are competitive for all alternatives investigated, whereas each of the other testing methods we investigate has very poor power for at least one alternative.

We now describe the general procedure for constructing Adaptive Neyman tests in the one sample setting. One begins with the principal component basis vectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots$ (with the corresponding eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots$) from $\hat{\Sigma}$ and computes the projections of the sample mean vector $\langle \bar{\mathbf{y}}, \hat{\mathbf{v}}_i \rangle$. By a Central Limit Theorem argument, these will be approximately independent under the null hypothesis, and by the properties of principal components, the smaller values of the index j correspond to more "important" projections. For a range of values of k, one computes

$$T_k^2 = \sum_{j=1}^{k} \hat{\lambda}_j^{-1} \left\langle \bar{\mathbf{y}}, \hat{\mathbf{v}}_j \right\rangle^2 .$$

One also needs normalizing constants c_k chosen to be approximately equal to $1/E[T_k^2]$. The final test statistic is chosen as

$$T^2 = \max_k c_k T_k^2 .$$

The term "adaptive" is used here to refer to this choice of k.

There are several problems to be dealt with. One is the choice of the range of k. Another is the determination of the normalizing constants c_k. Of course, one must also obtain a null distribution for the test statistic. The extension to the two sample setting also presents some issues. These are all dealt with at length below.

Fan and Lin (1998) also develop what they refer to as Adaptive Neyman tests for functional data. However, a critical issue is that they do not use principal components but rather preselected basis functions, either Fourier basis functions or wavelets. It is clear in the simulation study presented below that their proposed methodology performs poorly against four out of six alternatives we investigated.

Shen and Faraway (2004) propose another test which performs poorly on the alternatives where the Fan and Lin test do well. Recently, Lopes *et al.* (2012) proposed a version of Hotelling's T^2 test for functional data that involves randomly projecting the p dimensional data onto a smaller dimension, but this too performs poorly against four out of six alternatives. The only tests that performed consistently well were versions of the adaptive Neyman tests developed here. See Zhang (2011) for a survey of recent works on hypothesis testing with functional data.

2. Existing Methods

We restrict ourselves to hypothesis testing in the two sample setting. Suppose we have independent i.i.d. samples $y_{1i}(t)$, $i = 1, \ldots, n_1$ from Population 1, and $y_{2i}(t)$, $i = 1, \ldots, n_2$ from Population 2. Let $\mu_1(t) = E[y_{1i}(t)]$ and $\mu_2(t) = E[y_{2i}(t)]$. We assume finite second moments. We want to test the null hypothesis

$$H_0 : \mu_1(t) = \mu_2(t), \quad \text{for all } \ t \in \mathcal{T} \tag{2.1}$$

versus the general alternative $H_1 : \mu_1(t) \neq \mu_2(t)$ for some t.

It is difficult to derive a null distribution for any of the test statistics presented below. For our setting, if we extend the null to assume equal distributions (not just equal mean functions) for both populations, then we can interchange the observations between the 2 populations and the distribution remains the same, under the null. Thus, we randomly permute the population label (keeping the same sample sizes), recompute the test statistic, and the proportion of the values of the test statistics larger than the test statistic for the real data gives us an estimated p-value (assuming H_0 is rejected for larger values of the test statistic). We refer to this assumption of equal distributions under H_0 as the permutation pivotality condition for the two sample setting. In the one sample setting, if we assume that the distribution is symmetric about the constant 0 (*i.e.*, that $-Y(t)$ has the same distribution as a stochastic process at $Y(t)$), then we can apply random sign changes to obtain a null distribution by randomization. If we assume the $y(t)$ are realizations from a Gaussian process, then the permutation pivotality condition holds in the one sample setting, and it holds in the two sample setting if we assume equal covariance functions. Further discussion of permutation or randomization tests can be found in, for example, Good (2005). The permutation methodology is simple to understand and easy to implement. Also, assuming the permutation pivotality condition, this procedure has the advantage of being distribution free, and the test will be an exact level α test because of the properties of the permutation test and hence will give valid p-values.

We next briefly describe some of the existing testing procedures, all of which can be implemented with the permutation method.

2.1. The max-T statistic

Perhaps the easiest test statistic we can try is performing a pointwise two sample t-test, and obtain the maximum of the absolute value of t-statistics over all t, the "max-T" statistic:

$$\text{max-T} = \sup_{t \in \mathcal{T}} |T(t)|, \tag{2.2}$$

where $T(t)$ denotes the value of the two sample t-statistic based on $y_{11}(t), y_{12}(t), \ldots, y_{1n_1}(t)$ and $y_{21}(t)$, $y_{22}(t), \ldots, y_{2n_2}(t)$. One can derive an approximate distribution of the max-T statistic (Taylor *et al.*, 2007), but we utilize the permutation method described above to get a null distribution. Of course, we approximate the r.h.s. of (2.2) by taking the maximum over the grid of evaluation points.

2.2. Testing procedure of Fan and Lin

Fan and Lin (1998) proposed a test that uses truncation of the components of a test statistic to include only the useful ones. Their paper contains a number of ideas on choosing the number of components, but we only present a method that is relevant to our problem here. Assume that the domain $\mathcal{T}$ is a finite interval in the real line, and the evaluation points t_j are equally spaced. We Fourier transform the data $y_{ij}(t_k)$ with $i = 1, 2$ denoting the population, $j = 1, \ldots, n_i$, and $k = 1, \ldots, p$, so that the functions are transformed into approximately independent Gaussian random variables and the transform will compress signals into low frequencies, *provided that the data are approximately stationary*. This latter assumption will often not be the case with functional data.

Let $y^*_{ij}(\omega_k)$, with $j = 1, \ldots, n_i$ and $k = 1, \ldots, p$ be Fourier transformed data, for $i = 1, 2$. Because the Fourier transformed data have real and imaginary parts, Fan and Lin (1998) used the following scheme: let $y^*_{ij}(\omega_1)$ be the real part of the 0th Fourier frequency, $y^*_{ij}(\omega_2)$ be the real part of the 1st Fourier frequency, $y^*_{ij}(\omega_3)$ be the imaginary part of the 1st Fourier frequency, and so on until we obtain p such transformed data values. Then consider naïve estimates of the Fourier transformed mean curves, for $i = 1, 2$,

$$\bar{y}^*_i(\omega) = \frac{1}{n_i} \sum_{j=1}^{n_i} y^*_{ij}(\omega),$$

and the variance functions

$$\hat{\sigma}^2_i(\omega) = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} \left(y^*_{ij}(\omega) - \bar{y}^*_s(\omega) \right)^2 .$$

Let the standardized difference be

$$Z(\omega) = \frac{\bar{y}^*_1(\omega) - \bar{y}^*_2(\omega)}{\sqrt{\frac{\hat{\sigma}^2_1(\omega)}{n_1} + \frac{\hat{\sigma}^2_2(\omega)}{n_2}}}$$

and let $\mathbf{Z} = (Z(\omega_1), \ldots, Z(\omega_p))'$. When n_1 and n_2 are large, then $Z(\omega)$ is approximately distributed as $N(d(\omega), 1)$, where

$$d(\omega) \approx \frac{\mu^*_1(\omega) - \mu^*_2(\omega)}{\sqrt{\frac{\sigma^2_1(\omega)}{n_1} + \frac{\sigma^2_2(\omega)}{n_2}}}.$$

Here, $\mu^*_s(\omega)$ and $\sigma^2_s(\omega)$ denote the mean and variance of Fourier transformed data, respectively, for population $s = 1, 2$. Let $\mathbf{d} = (d(\omega_1), \ldots, d(\omega_p))'$. So under $H_0 : \mu_1 = \mu_2$, we have approximately

$$\mathbf{Z} \sim N\left(\mathbf{0}, \mathbf{I}_p\right).$$

Now, the maximum LRT statistic for the above problem is $\|\mathbf{Z}\|^2_2$, which tests for all components of $\mathbf{Z}$. But we want to test only some components of $\mathbf{Z}$, and we reject H_0 for large value of $\sum_{j=1}^{k} Z(\omega_j)^2$, if we know that only the first k components contain useful information. Fan (1996) and Fan and Lin (1998) proposed choosing k by

$$\hat{k} = \arg\max_k \left\{ k^{-\frac{1}{2}} \sum_{j=1}^{k} \left(Z\left(\omega_j\right)^2 - 1 \right) \right\}$$

then they obtain the adaptive Neyman test statistic

$$T^*_{AN} = \frac{1}{\sqrt{2\hat{k}}} \sum_{j=1}^{\hat{k}} \left(Z\left(\omega_j\right)^2 - 1 \right). \tag{2.3}$$

With some standardization, they were able to get a asymptotic distribution of T^*_{AN} and perform the test. We will not need this as we can apply the permutation methodology to obtain a null distribution.

2.3. Testing procedure of Shen and Faraway

Finally, we describe the testing procedure proposed by Shen and Faraway (2004). They considered a modified F test in the general functional regression setting, but the method can be specialized for our problem.

Their proposed test statistic for our setting is

$$\mathcal{F} = \frac{(\mathrm{rss}_0 - \mathrm{rss}_1)}{\mathrm{rss}_1/(n-2)}, \tag{2.4}$$

where

$$\mathrm{rss}_1 = \sum_{i=1}^{n_1} \int (y_{1i}(t) - \bar{y}_1(t))^2 \, dt + \sum_{i=1}^{n_2} \int (y_{2i}(t) - \bar{y}_2(t))^2 \, dt, \tag{2.5}$$

$$\mathrm{rss}_0 = \sum_{i=1}^{n_1} \int (y_{1i}(t) - \bar{y}(t))^2 \, dt + \sum_{i=1}^{n_2} \int (y_{2i}(t) - \bar{y}(t))^2 \, dt, \tag{2.6}$$

$$\bar{y}_s(t) = \frac{1}{n_s} \sum_{i=1}^{n_s} y_{si}(t),$$

$$\bar{y}(t) = \frac{n_1 \bar{y}_1(t) + n_2 \bar{y}_2(t)}{n_1 + n_2}.$$

Assume that the t_k are equally spaced. Then we can approximate the integrals in (2.5) and (2.6) by a rectangular quadrature rule; that is we have

$$\mathrm{rss}_1 = \frac{1}{p} \sum_{j=1}^{n_1} \sum_{k=1}^{p} \left(y_{1j}(t_k) - \bar{y}_1(t_k) \right)^2 + \frac{1}{p} \sum_{j=1}^{n_2} \sum_{k=1}^{p} \left(y_{2j}(t_k) - \bar{y}_2(t_k) \right)^2,$$

$$\mathrm{rss}_0 = \frac{1}{p} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{k=1}^{p} \left(y_{ij}(t_k) - \bar{y}(t_k) \right)^2$$

as approximations to (2.5) and (2.6).

The authors present a distributional properties of the $\mathcal{F}$ statistic and provide some results. But again, we will use the permutation methodology to derive a null distribution and obtain a p-value.

2.4. Random projection Hotelling's T^2 test

Lopes *et al.* (2012) proposed random projection of Hotelling's T^2 test, which in effect reduces a dimension from p to a smaller dimension. Recall the two-sample Hotelling's T^2 test from multivariate

analysis (Rencher, 2002), which has the test statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \widehat{\Sigma}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2),$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are vectorized sample mean functions ($p \times 1$), and $\widehat{\Sigma}$ is the pooled sample covariance matrix ($p \times p$). If $n < p$, then $\widehat{\Sigma}$ will be singular and hence not invertible. Lopes *et al.* deals with this problem by defining $\overline{T}_k^2$ to approximate T^2, where

$$\overline{T}_k^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbb{E}_{P_k} \left[P_k \left(P_k' \widehat{\Sigma} P_k \right)^{-1} P_k' \right] (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \tag{2.7}$$

with P_k a projection matrix ($p \times k$) and $\mathbb{E}_{P_k}$ an averaging operator with respect to P_k. The "random" part comes from the fact that P_k is drawn randomly from any possible projection matrix of the same dimension when computing $\mathbb{E}_{P_k}$ (at each step one randomly draws P_k to compute a quantity $P_k(P_k'\widehat{\Sigma}P_k)^{-1}P_k'$, and after N steps, one averages the resulting quantities - *i.e.*, compute the sample mean - to obtain $\mathbb{E}_{P_k}[P_k(P_k'\widehat{\Sigma}P_k)^{-1}P_k']$).

This is somewhat similar to our proposed method but does not have the data-driven adaptive feature of our method. We will compare all the methods in the simulations section.

3. Adaptively Truncated Hotelling's T^2 Test

3.1. Derivation of the test statistics

We now propose and describe test statistics, called the adaptively truncated Hotelling's T^2 test.

Suppose that we have vectors of functions evaluated on a grid $\mathbf{y}_{1i}$, $i = 1, \ldots, n_1$ from Population 1 and $\mathbf{y}_{2i}$, $i = 1, \ldots, n_2$ from Population 2, all with dimension p. Then, our null hypothesis becomes $H_0 : \mu_1 = \mu_2$, where μ_i now is an p-dimensional vector. If $p < n - 2$, we can use the two-sample Hotelling's T^2 test

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \widehat{\Sigma}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \tag{3.1}$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are vectorized sample mean functions, and

$$\widehat{\Sigma} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.2}$$

is the pooled sample covariance matrix.

This test has many attractive properties, as it is UMP invariant, admissible, and robust (Muirhead, 1982). But there are some assumptions that must be satisfied in order for this test to be valid. First, the data must be Gaussian with a common covariance structure, *i.e.* $\bar{\mathbf{y}}_1 \sim N(\mu_1, {n_1}^{-1}\Sigma_1)$ and $\bar{\mathbf{y}}_2 \sim N(\mu_2, {n_2}^{-1}\Sigma_2)$, with $\Sigma_1 = \Sigma_2 = \Sigma$. Also, in order for the estimated covariance $\widehat{\Sigma}$ to be non-singular, we must have that $n - 2 > p$, where $n = n_1 + n_2$.

Since we often have $n \ll p$ in the functional data setting, and the equal covariance assumption may not always hold, we try to modify the ordinary two-sample T^2 statistic (3.1) for the functional data setting. Because of some nice properties possessed by this test, we would like to develop a testing procedure that takes advantages of its good properties. First, we note that the pooled sample covariance matrix (3.2) is symmetric and non-negative definite. So we can write

$$\widehat{\Sigma} = \widehat{V}\widehat{D}\widehat{V}' = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j', \tag{3.3}$$

where the $\hat{\lambda}_j$ are the eigenvalues and the $\hat{\mathbf{v}}_j$ are the eigenvectors of $\widehat{\Sigma}$. We propose to keep only the first k components, for some $k \leq p$, and generally $k \ll p$. The adaptive Neyman methodology provides a rationale for selecting a value of k. Then, we substitute in $\widehat{\Sigma}$ above with k replacing p into the T^2 statistic (3.1) to obtain

$$\tilde{T}_k^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \left(\sum_{j=1}^{k} \hat{\lambda}_j^{-1} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j' \right) (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{k} \hat{\lambda}_j^{-1} \left((\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \hat{\mathbf{v}}_j \right)^2. \tag{3.4}$$

Note that $\tilde{T}_p^2$ (if $p < n$) is the Hotelling T^2 statistic, and if $k < p$, then not all components are included.

Now, the fundamental question is how many components to choose. Borrowing the idea of Fan and Lin (1998), we propose the following test statistic, which we call the *adaptively truncated Hotelling's T^2 statistic*: $\max_k c_k \tilde{T}_k^2$, where c_k is some suitable sequence of normalization constants to be determined.

In practice, we take the maximum over $k = 1, \ldots, p_0$ where $p_0 < p$, since many of the trailing eigenvalues $\hat{\lambda}_j$ are close to zero and so the $\tilde{T}_k^2$ will be very big for large k. Note that there is large relative numerical error in computing the small eigenvalues. Hence, we generally want the cutoff p_0 to be determined from the ordered sample eigenvalues $\hat{\lambda}_j$, and we use

$$p_0 = \max \left\{ j : \hat{\lambda}_j / \hat{\lambda}_1 > c \right\} \text{ for some threshold } c > 0, \text{ where } c = 10^{-16} \text{ for our simulation.} \tag{3.5}$$

Then, the technique amounts to finding the maximum of $c_1 \tilde{T}_1^2, \ldots, c_{p_0} \tilde{T}_{p_0}^2$.

Then the next question becomes how to find some suitable normalizations c_k. Again, different choices of c_k may lead to different determinations of k, and we present 2 possible ways of handling this problem. Our goal is that the $c_k \tilde{T}_k^2$ all be comparable in magnitude under H_0. One way of deriving the normalizing constant c_k is as follows. In the case $p < n - 2$, recall that

$$\frac{n - p - 1}{(n - 2)p} T^2 \sim F_{p, n-p-1}. \tag{3.6}$$

In our problem we want to transform the truncated statistic $\tilde{T}_k^2$. It is obvious that transformation (3.6) will not work in the FDA setting, because the covariance term in $\tilde{T}_k^2$ is of rank $k < p$. However, since we have a rank k covariance matrix in $\tilde{T}_k^2$, we may try multiplying $\tilde{T}_k^2$ by

$$c_k = \frac{n - k - 1}{(n - 2)k},$$

so that our test statistic will be

$$S^* = \max_k c_k \tilde{T}_k^2 = \max_k \frac{n - k - 1}{(n - 2)k} \tilde{T}_k^2. \tag{3.7}$$

Of course, the quantity $c_k \tilde{T}_k^2$ will not be distributed exactly as an F, but it will be an approximation adequate for our purpose. Note that under H_0 an F statistic will be about 1, so we achieve our goal of obtaining $c_k \tilde{T}_k^2$ comparable in magnitude under H_0.

Another way of obtaining the normalization makes use of the following fact: Let $\widehat{\Sigma} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j'$ be the sample pooled covariance (3.3) and let Σ be the true covariance matrix. In general, the sample eigenvalues $\hat{\lambda}_j$ and the sample eigenvectors $\hat{\mathbf{v}}_j$ of sample covariance matrix $\widehat{\Sigma}$ are consistent estimators

of true eigenvalues λ_j and true eigenvectors $\mathbf{v}_j$, respectively, for $j = 1, \ldots, p$ (see Hall and Hosseini-Nasab, 2006). Then under $H_0 : \mu_1 = \mu_2$, for large sample sizes n_1 and n_2, the statistic $\tilde{T}_k^2$ (3.4) is approximately distributed as a χ^2 distribution with k degrees of freedom (Anderson, 2003).

It follows that $\tilde{T}_k^2$ has mean k and variance $2k$. Hence, we can normalize $\tilde{T}_k^2$ by $(\tilde{T}_k^2 - k)/\sqrt{2k}$, so that $c_k = (1 - k/\tilde{T}_k^2)/\sqrt{2k}$.

Then our test statistic becomes

$$T^* = \max_k \frac{1}{\sqrt{2k}} \left(\tilde{T}_k^2 - k \right). \tag{3.8}$$

We called the test statistic S^* (3.7) the adaptively truncated Hotelling's T^2 statistic *with F-transform*, and we call T^* (3.8) the adaptively truncated Hotelling's T^2 statistic *with χ^2-transform*.

Having done the transformations, we still do not know the exact distribution of S^* nor that of T^*. In such a situation, we can resort to permutation methods. As mentioned previously, this procedure has many advantages. In particular, the procedure is distribution-free and is quite flexible to use. Nevertheless, it has a major disadvantage in that it is computationally expensive. To obtain the test statistic S^* or T^*, we must obtain eigenvalues and eigenvectors of $\widehat{\Sigma}$, which depends on the labeling from each randomization. We in fact calculate the Singular Value Decomposition (SVD) of the data matrix centered by subtracting the within group means at each randomization to obtain the eigen pairs, but as we will see in the coming sections, this can take a lot of computing time for a reasonable number of randomizations.

3.2. Wald and Wolfowitz modification of adaptively truncated Hotelling's T^2 statistic

The paper by Wald and Wolfowitz (1944) provides a way of tackling the computational problem. Rather than computing pooled sample covariance matrix $\widehat{\Sigma}$, we may compute the "ordinary" or "total" sample covariance matrix that is independent of labeling,

$$\widehat{\Sigma}_T = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'.$$

If we were to use $\widehat{\Sigma}_T$ in place of $\widehat{\Sigma}$ in computing (3.7) or (3.8), then we do not need to compute eigen pairs at each iteration (since $\widehat{\Sigma}_T$ is independent of labeling).

Wald and Wolfowitz (1944) showed that T^2 based on $\widehat{\Sigma}$ is a monotonic function of T^2 based on $\widehat{\Sigma}_T$. But again this is valid only for the case an invertible $\widehat{\Sigma}$. However, we implemented the Wald-Wolfowitz method based on the eigen-decomposition of $\widehat{\Sigma}_T$ and compared it with the full SVD method using $\widehat{\Sigma}$.

To summarize, we have developed four versions of adaptively truncated Hotelling's T^2 statistics for functional data and other high dimensional data where the sample covariance matrix is ill conditioned. They are listed in Table 1 (Tests 1 to 4). We will investigate and compare the properties of these and other tests.

4. Simulation

With the methodology developed in the previous section, we are ready to perform simulations and present some results. Most of the programming and analysis are done in MATLAB, with the exception of the smoothing used to generate the functional data, which is done in R. Because of our randomization approach, we are guaranteed all tests achieve the level of significance (we use $\alpha = 0.05$). Hence, we concentrate on simulations to assess power.

Table 1: A list of 8 testing procedures considered in simulations.

Test	Description
1	Full SVD version of adaptively truncated Hotelling's T^2 with F-transform (3.7)
2	Full SVD version of adaptively truncated Hotelling's T^2 with χ^2-transform (3.8)
3	Wald-Wolfowitz version of adaptively truncated Hotelling's T^2 with F-transform ((3.7) with $\widehat{\Sigma}_T$ in place of $\widehat{\Sigma}$)
4	Wald-Wolfowitz version of adaptively truncated Hotelling's T^2 with χ^2-transform ((3.8) with $\widehat{\Sigma}_T$ in place of $\widehat{\Sigma}$)
5	Max-T statistic (2.2)
6	Fan and Lin (2.3)
7	Shen and Faraway (2.4)
8	Random projection Hotelling's T^2 (2.7)

Table 2: A list of 6 alternatives for simulations.

Alternative	$\mu_2(t),\ 0 \le t \le 1$	Name
A	0.0001	constant
B	$0.0004(t - 0.5)$	linear
C	$0.0002\sin(2\pi t)$	sine
D	$0.0001\text{beta}_{5,5}(t)$	beta
E	$0.00001\text{beta}_{1000,1000}(t)$	beta spike
F	$0.001t\sin(e^{4t})$	chirp

We construct simulated functional data as follows. First we draw 1,000 i.i.d. random normal numbers with zero mean and a fixed variance, and then we smooth them with a smoothing spline with a predetermined smoothing parameter. Repeat this 250 times with the same variance and smoothing parameter. We use the t values $0, 0.001, 0.002, \ldots$. The resulting 250 curves form the sample from Population 1. Next, repeat the procedure again for 250 times, but now add a mean function which is evaluated at the same 1,000 points as above. Call the resulting curves the samples from Population 2.

Mathematically, we have that Population 1 consists of independent Gaussian process with mean zero function $\mu_1(t) \equiv 0$, where $0 \le t \le 1$. This will be the case throughout the simulation. In Population 2, we also have independent realizations of a Gaussian process with the same covariance structure as Population 1, but now the mean function $\mu_2(t)$ is not identically zero. We have that $n_1 = n_2 = 250$ and $p = 1{,}000$.

Recall that we have four versions of our proposed method. We then compare our proposed test statistics to the following procedures described in Section 2: the max-T statistic, the method of Fan and Lin (1998), the test proposed by Shen and Faraway (2004), and the random projection of Hotelling's T^2 by Lopes *et al.* (2012). For the random projection of Hotelling's T^2 test procedure, we have followed their methodology closely in the implementation (2.7), except that we have computed $N = 100$ projections for the averaging of the test statistics (instead of $N = 30$ recommended from their paper) and let the number of dimension to random projections be $k = \lfloor p_0/2 \rfloor$ (instead of recommended $k = \lfloor n/2 \rfloor$ from their paper), where p_0 is from (3.5) and $n = n_1 + n_2$, for performance and numerical stability reasons. We used the permutation methodology to give a null distribution and obtain a p-value for all proposed tests. For convenience, we have listed the eight tests in Table 1.

We tried six mean functions for Population 2, which are given in Table 2. Alternatives D and E use scaled versions of Beta densities. The alternatives are depicted in Figure 1.

We have the computation times listed for each test in Table 3. We get a clear picture of the disadvantage of the full SVD method, but the Wald-Wolfowitz modification actually is very computationally efficient. In fact, the Wald-Wolfowitz version of our test statistic is the fastest among all the testing procedures we considered, even though the tests 5, 6, 7 and 8 ran fairly fast as expected.

Now, if we look at the result in Table 3, we see that most methods do well in low frequency

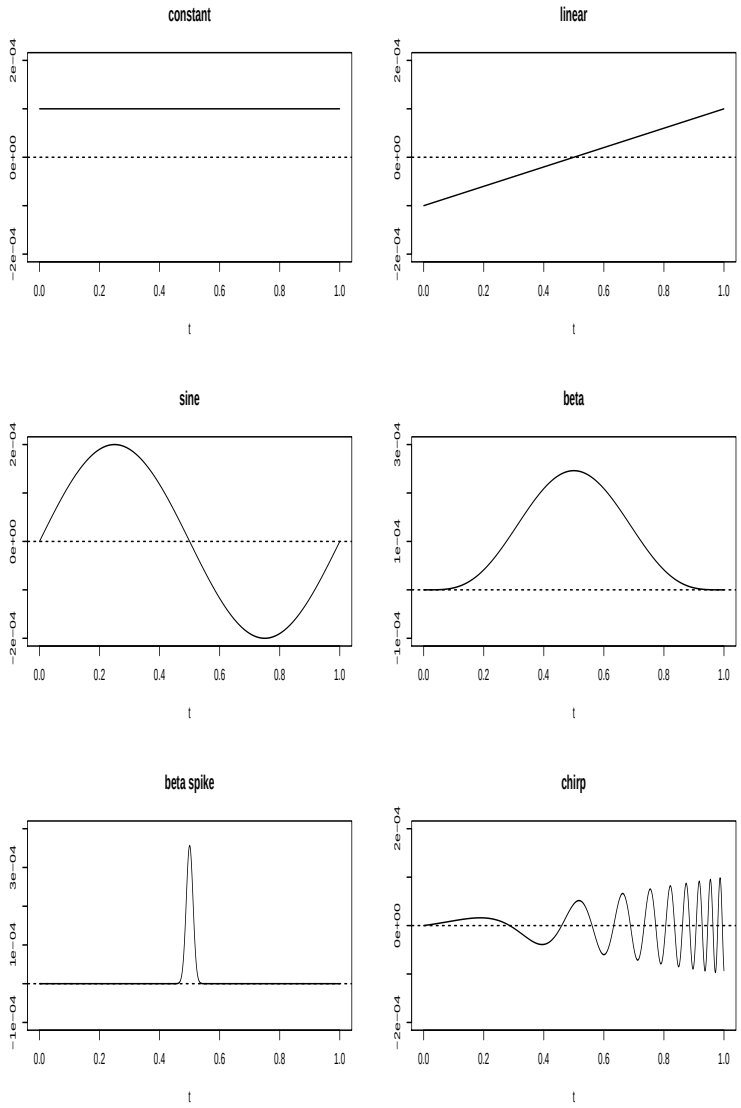

Figure 1: *Plot of all 6 alternatives from Table 2. A dashed line is a horizontal line at 0.*

alternatives (A through D), except the test 6 (Fan and Lin, 1998). On the other hand, if the alternative contains high frequency component (alternatives E and F), then most do a good job except the test 7 (Shen and Faraway, 2004).

Note that all the results in Table 3 are done for one monte carlo run to generate the two sample with 10,000 permutations to generate null distributions. To truly understand the performance of the tests, we consider a power study with extended number of monte carlo runs. However, since it takes too long to run the tests 1 and 2 (see Table 3), we decided to only include the tests 3 to 8 for this simulation. As before, we generate 500 independent realizations of Gaussian processes, and then add a $\mu_2(t)$ function for the alternative ($n_2 = 250$ for Population 2). We run the permutation tests and

Table 3: A table of p-values obtained from 10,000 permutation iterations of the 8 tests (rows 1 through 8) and 6 alternatives (columns A through F). The last column lists the CPU time for each test in seconds. The null hypothesis is $H_0 : \mu_1(t) = \mu_2(t)$ with $\mu_1(t) \equiv 0$.

Test	$\mu_2(t)$ (see Table 2)						Time
(see Table 1)	A	B	C	D	E	F	
1	0.0023	0.0509	0.0260	0.0019	0	0	4,016
2	0.0018	0.0532	0.0514	0.0006	0	0	4,016
3	0.0024	0.0273	0.0193	0.0028	0	0	33
4	0.0002	0.0414	0.0468	0.0003	0	0	33
5	0.0076	0.0184	0.1014	0.0022	0	0.0061	86
6	0.1833	0.3142	0.3216	0.2104	0	0	87
7	0.0001	0.0091	0.0103	0	0.2683	0.1361	51
8	0.0881	0.4537	0.0617	0.0054	0	0	49

obtain p-values for all 6 alternatives. Then we generate the random Gaussian processes again and repeat the process above 100 times, so that we obtain 100 p-values for each alternative. Finally, we plot the empirical c.d.f.'s of the 100 p-values, to give us some idea of the power at these alternatives.

The results are presented in Figure 2. The way to interpret the picture is that, if a curve dominates (lies above) other curves in a plot, then the test that corresponds to the dominating curve is more powerful than other tests. For the constant and linear alternatives, the comparative performance among the tests are consistent with those of sine and beta alternatives, so we reach the same conclusion for these alternatives.

From looking at Figure 2, the general conclusion is that test 7 is the best and test 6 performs the worst in the alternatives A to D, while the situation is reversed for alternatives E and F. Our tests, 3 and 4, do well in all situations. The max-T test (test 5) performs reasonably well (except for Alternative F), but our tests are usually more powerful. This confirms our findings from previous simulation results that tests 3 and 4 perform the best overall. Since the test 4 is slightly easier to compute and often outperforms the test 3, we recommend the use of test 4.

We have also run the simulation at different magnitudes of the alternatives and produced graphs similar to those in Figure 2. The results were consistent with the results shown in Figure 2, in that the dominance (and hence the power) remained in the same order at different magnitudes for every alternative. The results of the extended simulation results will be available from the authors upon request.

5. Applications

The application of interest concerns a medical device that uses fluorescence spectroscopy for cervical cancer screening. This technology involves measuring fluorescence spectra at a few sites for each patient, and we obtain 16 smooth curves for each observation. Figure 3 shows an example of a fluorescence spectrum for the application. The domain for these functional data is the 2-dimensional region in the right panel of Figure 3. See Cantor *et al.* (2011) for more details on this technology.

For the clinical trial, researchers have gathered information on the spectroscopic measurements at several different sites of cervix, as well other information (such as the device and probe used for the measurements and biographical information on the subject). They also obtained pathologist's diagnosis of disease state at each site for the purpose of classification of the disease states. The pathologists also determined the tissue type (squamous, columnar, or mixed). Unfortunately, some preliminary studies have shown that there are many sources of variability in the measurements. In a previous work (Pikkula *et al.*, 2007), an experiment was conducted to assess the variability due to the fluorescence

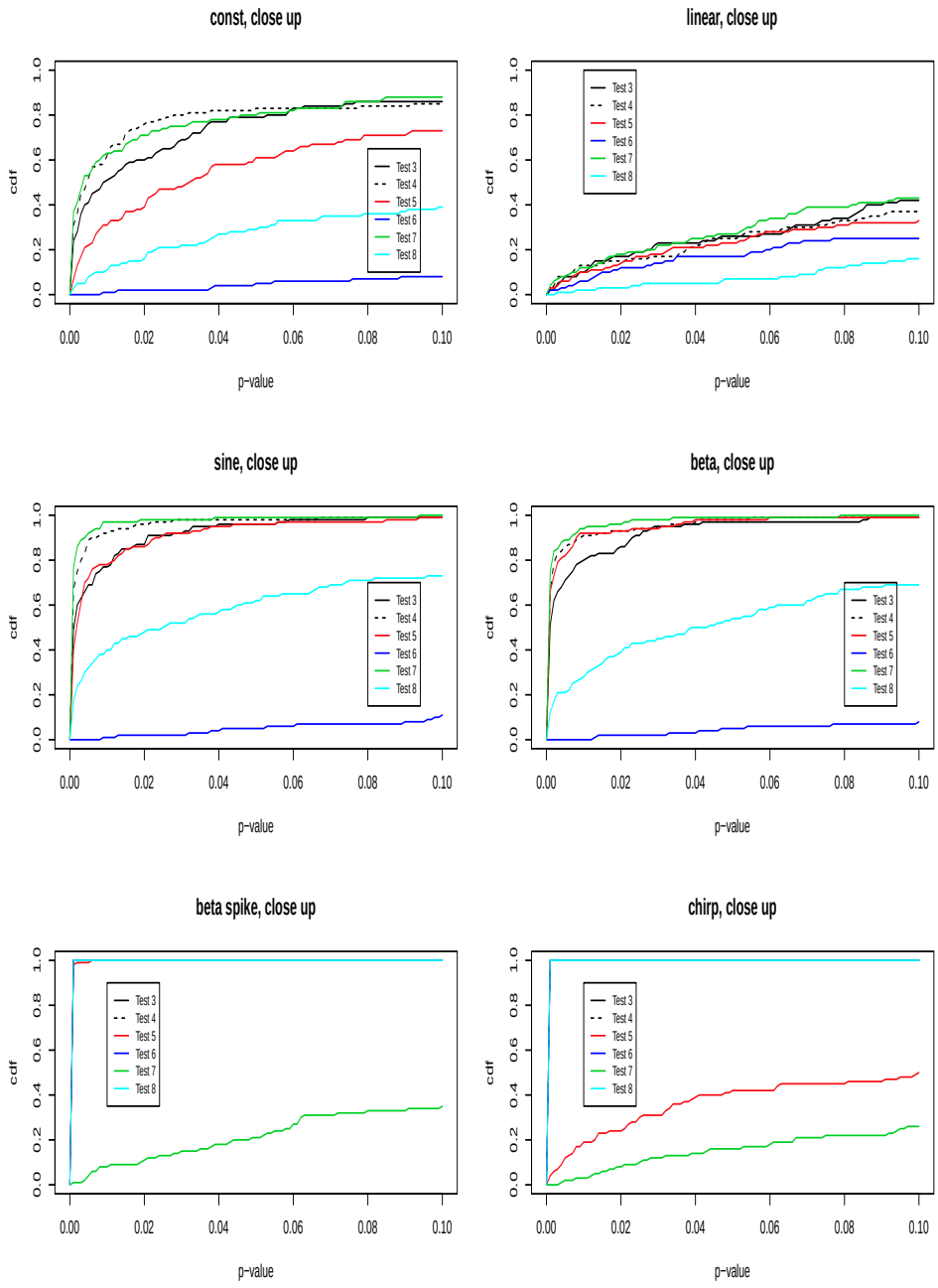

Figure 2: *Plots of the empirical c.d.f.'s of the p-values for all 6 alternatives (comparing Tests 3 to 8).*

spectroscopy devices, and it was reported that the device differ significantly in their output, even when they measured the same thing. We would like to perform an another study, controlling for different factors that may have affected the analysis previously and possibly gave misleading results. We subset

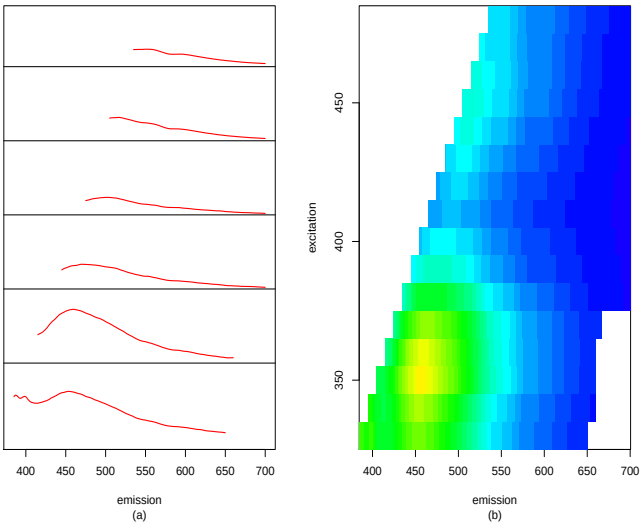

Figure 3: *An example of fluorescence spectra. The left panel shows six sample curves from one observation, where each curve corresponds to intensity of an excitation wavelength (values on the vertical axis of right panel) as a function of emission wavelengths (horizontal axis of right panel). The right panel is a representation of spectra by a heat map, where intensity is represented as a function of emission and excitation wavelengths.*

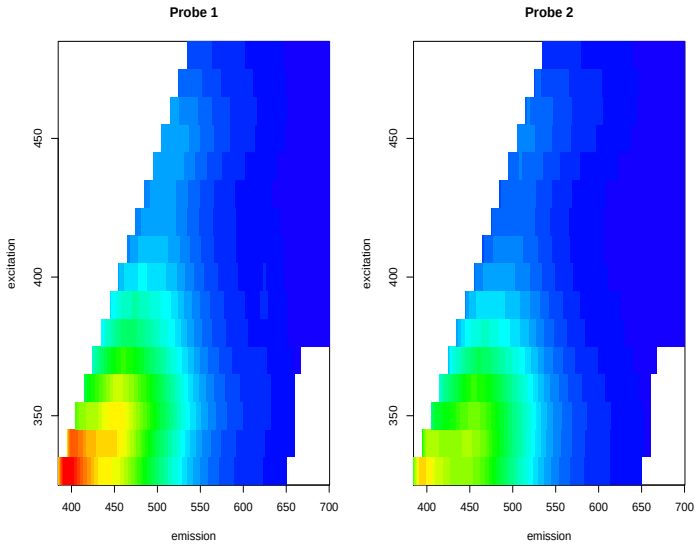

Figure 4: *Comparing pointwise means of the two probes.*

the data to make sure that all known possible sources of variability are controlled. Specifically, we consider only one device and one tissue type (squamous), and select one normal site from the multiple measurements of a patient.

With everything else controlled as described above, we first look at the 2 probes used with that device, where some patients were measured using one probe (probe 1) while others were measured using the other probe (probe 2). In this way we obtain 74 observations from probe 1 and 84 observations

from probe 2. See Figure 4 for the mean spectra for each probe.

Based on the figures, there seem to be some difference of means in the probes. For an empirical study, this descriptive conclusion may by enough, but we would like to analytically determine the significance. Thus, we perform the adaptive Hotelling's T^2 test with the χ^2-transform version. The result is significant (p-value = 0.0002), which provides a strong evidence that the probes are different. Hence, for the future investigation, we need to somehow either control for this effect or remove it (*e.g.*, by better probe design).

6. Conclusion

The overall result of the adaptively truncated Hotelling's T^2 method is satisfactory in the simulation study and applications.

There are unresolved issues, some of which we now list as potential future research questions. First, there may be a concern of bias of the sample eigenvalues. In general, the leading sample covariance eigenvalues tend to overestimate the true eigenvalues, and the trailing sample eigenvalues tend to underestimate their population counterparts, but eigenvalue bias correction is a difficult task (see, *e.g.* Johnstone, 2001). Also, establishing the consistency for the adaptively truncated Hotelling's T^2 test is challenging. Since we are dealing with functional data, we want to investigate when $p \to \infty$, i.e., the case where our sampled vector converges to a sample function as we make the grid finer. Hence, it will be desirable to investigate the behavior of both the sample sizes n_1 and n_2, as well as dimension (number of grid points) p, of our test statistics to go to infinity. Additionally, it would be desirable to relax the assumption of equal covariances. One could use the same statistic, although there may be better test statistics, but certainly not use permutations to get a null distribution. A simulation based approach could be used based on an assumption of Gaussian processes.

Despite these unresolved issues, we conclude from our study that the method gives good results in a variety of situations and is a useful functional data analytic tool.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Third Edition, Wiley, New York.

Cantor, S. B., Yamal, J. M., Guillaud, M., Cox, D. D., Atkinson, E. N., Benedet, J. L., Miller, D., Ehlen, T., Matisic, J., van Niekerk, D., Bertrand, M., Milbourne, A., Rhodes, H., Malpica, A., Staerkel, G., Nader-Eftekhari, S., Adler-Storthz, K., Scheurer, M. E., Basen-Engquist, K., Shinn, E., West, L. A., Vlastos, A. T., Tao, X., Beck, J. R., MacAulay, C. and Follen, M. (2011). Accuracy of optical spectroscopy for the detection of cervical intraepithelial neoplasia: Testing a device as an adjunct to colposcopy, *International Journal of Cancer*, **128**, 1151–1168.

Cox, D. D. and Lee, J. S. (2008). Pointwise testing with functional data using the Westfall-Young randomization method, *Biometrika*, **95**, 621–634.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation, *Journal of American Statistical Association*, **91**, 674–688.

Fan, J. and Lin, S. (1998). Test of significance when data are curves, *Journal of the American Statistical Association*, **93**, 1007–1021.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer, New York.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis, *Journal of the Royal Statistical Society, Series B*, **68**, 109–126.

Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1994). Power approximations to and power com-

parison of smooth goodness-of-fit tests, *Scandinavian Journal of Statistics*, **21**, 131–145.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalues in principal components analysis, *Annals of Statistics*, **29**, 295–327.

Lopes, M. E., Jacob, L. and Wainwright, M. J. (2012). A more powerful two-sample test in high dimensions using random projection (arXiv:1108.2401v2)

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Neyman, J. (1937). Smooth test for goodness of fit, *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

Pikkula, B. M., Shuhatovich, O., Price, R. L., Serachitopol, D. M., Follen, M., McKinnon, N., MacAulay, C., Richards-Kortum, R., Lee, J. S., Atkinson, E. N. and Cox, D. D. (2007). Instrumentation as a source of variability in the application of fluorescence spectroscopy devices for detecting cervical neoplasia, *Journal of Biomedical Optics*, **12**, 034014.

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

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, Second edition, Wiley, New York.

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

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

Wald, A. and Wolfowitz, J. (1944). Statistical tests based on permutations of the observations, *Annals of Mathematical Statistics*, **15**, 358–372.

Zhang, J. (2011). Statistical inferences for linear models with functional responses, *Statistica Sinica*, **21**, 1431–1451

Received November 24, 2014; Revised January 30, 2015; Accepted January 31, 2015