A Two Sample Distribution-Free Test for Functional Data with Application to a Diffusion Tensor Imaging Study of Multiple Sclerosis

Gina-Maria Pomann

Duke University, Durham, USA.

Ana-Maria Staicu*

North Carolina State University, Raleigh, USA.

Sujit Ghosh

North Carolina State University, Raleigh and Statistical and Applied Mathematical Sciences Institute, RTP, NC. USA.

Summary. Motivated by an imaging study, this paper develops a nonparametric testing procedure for testing the null hypothesis that two samples of curves observed at discrete grids and with noise have the same underlying distribution. The objective is to formally compare white matter tract profiles between healthy individuals and multiple sclerosis patients, as assessed by conventional diffusion tensor imaging measures. We propose to decompose the curves using functional principal component analysis of a mixture process, which we refer to as *marginal functional principal component analysis*. This approach reduces the dimension of the testing problem in a way that enables the use of traditional nonparametric univariate testing procedures. The procedure is computationally efficient and accommodates different sampling designs. Numerical studies are presented to validate the size and power properties of the test in many realistic scenarios. In these cases, the proposed test has been found to be more powerful than its primary competitor. Application to the diffusion tensor imaging data reveals that all the tracts studied are associated with multiple sclerosis and the choice of the diffusion tensor image measurement is important when assessing axonal disruption.

1. Introduction

In this paper, we consider testing the hypothesis that two or multiple samples of curves have the same distribution, when the observed data are realizations of the curves at finite grids and possibly corrupted by noise. The motivating application is a diffusion tensor imaging study, where the objective is to formally assess whether several imaging modalities vary differently between patients with multiple sclerosis and healthy controls. We address this problem by introducing a novel framework based on functional principal component analysis of an appropriate mixture process, which we refer to as marginal functional principal component analysis (marginal FPCA). We develop a computationally stable and theoretically valid procedure. Our approach is applicable to a variety of realistic scenarios, including 1) curves observed at dense or sparse grids of points, with or without measurement error; 2) different sampling designs for different samples; and 3) different sample sizes. The methodology scales well with the total sample size and it can be extended to test for the equality of distributions of more than two samples of curves.

1.1. Diffusion Tensor Imaging Data

Multiple sclerosis (MS) is a disease that affects the central nervous system and in particular damages white matter tracts in the brain through lesions, myelin loss, and axonal damage. One of the recent approaches to visualize the white matter tracts is the use of diffusion tensor imaging (DTI), a magnetic resonance imaging technique that measures water diffusivity in the brain. In the brain, water diffuses isotropically in both gray matter and cerebrospinal fluid and anisotropically in the white matter regions, which makes DTI an ideal approach to study the white matter tracts (Goldsmith et al. (2012)). There are several well identified white matter tracts in the brain: right/left corticospinal tract (rCST, lCST), corpus callosum (CCA), and right/left optic radiations tract (rOPR, lOPR). Along each tract, there are several measures that DTI provides. Two of these measures (referred to as modalities) that are most commonly used are fractional anisotropy (FA) and parallel diffusivity

(L0), which characterize the tissue microstructure. Briefly, FA describes the degree of anisotropy of the water diffusion process while LO measures the diffusivity along the principal axis. In this paper, we focus on continuous summaries of these measures of water diffusivity, as parameterized by distance along the tract and refer to them as tract-specific modality profiles.

The DTI study comprises 162 subjects with MS and 42 healthy controls observed at multiple hospital visits. Details of this study have been described in Staicu et al. (2012); Gertheiss et al. (2013a), Gertheiss et al. (2013b); they addressed some of the scientific questions by focusing on specific aspects of the data. Our objective is to study whether the water diffusivity, as measured by the conventional FA or LO, along each of several well identified white matter tracts varies between MS patients and controls. In this work, we consider the data collected at the patients' baseline visit: Figure 1 displays the two DTI measures, FA and L0, along the CCA, rCST, lCST, and lOPR tracts for all the subjects in the study. For example the top left plot shows the FA profile along the CCA tract for the MS patients. Each line corresponds to an MS patient and is obtained by connecting the patient's FA at the 93 locations along the CCA. A subject's data can be viewed as arising from a subject-specific latent smooth function that is evaluated at a grid of locations and the evaluations are corrupted by noise. It is this latent function that characterizes the FA of the water diffusivity along the CCA for each subject. The interest is to formally assess whether the profiles of the true FA or L0 of water diffusivity along each of the five tracts vary differently between the MS patients and healthy controls. The result of this investigation has the potential to shed new light onto our understanding of the neurodegeneration associated with this disease. In particular, it facilitates the identification of the white matter tracts that are likely targeted by MS, as well as the modalities that capture such neurodegeneration.

In studies involving multiple groups, formally assessing whether the distribution of the 'characteristics' of interest is the same across the groups is the first step of the analysis. In a recent study of the neuronal electrophysiological properties of

rodents of different sexes, Dorris et al. (2014) compare the distribution of each of several characteristics of interest across the two gender groups. For scalar or vector 'characteristics', there is extensive literature on this topic. However, this problem is beginning to attract increasing interest when the 'characteristics' are curves that are observed with error, at finite grids of points. Hall and Van Keilegom (2007) and Corain et al. (2014) have considered this problem when the sampling design is dense, e.g. the curves are observed at very fine grids of points. To the best of the authors' knowledge no methodology exists outside of this case. In the DTI study, although both FA and L0 are sampled on a regular grid, some subjects have missing data, with the percentage of missing values ranging up to 22%. This sampling design does not fall under the situations studied by the existing methods. In this paper, we propose testing methodology that is applicable to noisy curves observed under both dense and sparse sampling designs.

While our research is motivated by the DTI study, it is increasingly common to observe curves that are separated into groups and to test whether the distribution of the curves is the same across the groups. One example is studied by Annette Moller et al. (2015), where the interest is to test whether the Raman spectra of boar-tainted or non boar-tainted mean samples have the same distribution. Another example is a recent study of activity in cats suffering from degenerative joint disease where the interest is to test whether the minute-by-minute activity profiles of the cats who receive treatment vary differently compared to those who receive placebo.

1.2. Statistical framework

In this section, we formally describe the statistical framework for this problem. Suppose we observe data arising from two groups, $\{(t_{1ij}, Y_{1ij}) : j = 1, \ldots, m_{1i}\}_{i=1}^{n_1}$ and $\{(t_{2ij}, Y_{2ij}) : j = 1, \ldots, m_{2i}\}_{i=1}^{n_2}$, where $t_{1ij}, t_{2ij} \in \mathcal{T}$, a compact interval; for simplicity we take $\mathcal{T} = [0, 1]$. For example Y_{1ij} could be the FA at location t_{1ij} along the CCA for the ith MS patient, while Y_{2ij} could be the FA at location t_{2ij} along the CCA for the ith healthy control. The notation of the time-points, t_{1ij} and t_{2ij}, allows for different observation points in the two groups. It is assumed that

the Y_{1ij}'s and the Y_{2ij}'s are independent realizations of two underlying (stochastic) processes observed with noise on a finite grid of points. Specifically, assume

$$Y_{1ij} = X_{1i}(t_{1ij}) + \epsilon_{1ij}, \text{ and } Y_{2ij} = X_{2i}(t_{2ij}) + \epsilon_{2ij}, \tag{1}$$

where $X_{1i}(\cdot) \overset{iid}{\sim} X_1(\cdot)$ and $X_{2i}(\cdot) \overset{iid}{\sim} X_2(\cdot)$ are independent and square-integrable random functions over $\mathcal{T}$, for some underlying (latent) random processes $X_1(\cdot)$ and $X_2(\cdot)$. In the DTI example, $X_{1i}(\cdot)$ and $X_{2i}(\cdot)$ would be the true latent smooth FA along the CCA for the ith MS patient and ith control, respectively. It is assumed that $X_1(\cdot)$ and $X_2(\cdot)$ are second order stochastic processes with mean functions assumed to be continuous and covariance functions assumed to be continuous and positive semi-definite, both being unknown. The measurement errors, $\{\epsilon_{1ij}\}$ and $\{\epsilon_{2ij}\}$, are independent and identically distributed with mean zero, and with variances σ_1^2 and σ_2^2 respectively, and are assumed independent of $X_{1i}(\cdot)$ and $X_{2i}(\cdot)$. Our objective is to test the null hypothesis,

$$H_0 : X_1(\cdot) \overset{d}{=} X_2(\cdot) \tag{2}$$

versus the alternative $H_A : X_1(\cdot) \overset{d}{\neq} X_2(\cdot)$, where "$\overset{d}{=}$" denotes that the processes on either side have the same distribution. In this paper we develop a non-parametric and computationally stable method to test the null hypothesis in (2).

Since $X_1(\cdot)$ and $X_2(\cdot)$ are processes defined over a continuum, hypothesis (2) implies that the two infinite dimensional objects have the same generating distribution. This is different from the two sample testing in a multivariate framework, where the dimension of the random objects of interest is finite and the same. In the case where the sampling design is common to all the subjects (i.e. $t_{1ij} = t_{2ij} = t_j$ and $m_{1i} = m_{2i} = m$), the dimension of the testing problem could potentially be reduced by testing an approximate null hypothesis - that the multivariate distribution of the processes evaluated at the observed grid points are equal. Some of the popular multivariate testing procedures (eg. Friedman and Rafsky (1979); Read and Cressie (1988); Aslan and Zech (2005)) could be employed in this situation. However, these procedures have only been illustrated for cases when $m = 4$ or 5

in our notation. In dense functional data, the number of unique time-points, m, is orders of magnitude larger, often even larger than the sample size, precluding the straightforward use of these finite-dimensional multivariate methods.

1.3. Existing approaches for two sample testing

Two sample hypothesis testing for functional data has been considered in many contexts; ranging from testing for specific types of differences, such as differences in the mean or covariance functions, to testing for overall differences in the cumulative density functions. To detect differences in the mean functions of two independent samples of curves Ramsay and Silverman (2005) introduced a pointwise t-test, Zhang et al. (2010) presented an L^2-norm based test, Horváth et al. (2013) proposed a test based on the sample means of the curves, and Staicu et al. (2014) developed a pseudo likelihood ratio test. Extension to k independent samples of curves was discussed in Cuevas et al. (2004), Estévez-Pérez and Vilar (2008), and Laukaitis and Račkauskas (2005), who proposed ANOVA-like testing procedures for testing the equality of mean functions. For detecting differences in the covariance functions of independent samples of curves Ferraty et al. (2007) proposed a factor-based test, Kraus and Panaretos (2012) introduced a regularized M-test, and Fremdt et al. (2012) proposed a chi-squared test.

Literature on testing the equality of the distributions of two sets of curves observed at a discrete grid of points and possibly with error is rather scarce. Corain et al. (2014) proposed a permutation testing framework for comparing two sets of functions. However, their method requires the functions to be observed on a regularly spaced grid of points. Thus, pre-smoothing would be required to apply this test to the DTI data; such extension has not been studied. Hall and Van Keilegom (2007) investigated the effect of pre-smoothing on testing procedures. They proposed an extension of the multivariate Cramer-von Mises (CVM) test and use of bootstrapping to approximate the null distribution of the test. These methods are again developed for noisy functional data observed on dense designs.

Benko et al. (2009) attempted to address this problem by considering a com-

mon functional principal components model for the two samples and testing for the equality of the corresponding model components using bootstrapping. In the proposed form, this test does not account for multiple comparisons and is not directly applicable to our general testing problem. We use their proposed bootstrap test in the DTI application for testing the equality of the mean profiles of the MS patients and healthy controls. The testing procedure is compared with our proposed test using simulated data and the results are presented in the appendix. The use of bootstrap techniques, while advantageous because they do not rely on specific distributional assumptions, comes with the price of a large computational burden. Therefore, it is often numerically unfeasible to perform extensive empirical power analysis when the sample size is moderately large.

In this paper, we propose an approach based on the so-called marginal FPCA, which facilitates representation of the curves using the marginal eigenbasis. This reduces the original infinite dimensional two-sample functional testing problem to an approximate simpler finite dimensional testing problem. The methodology is applied using the two-sample Anderson-Darling statistic (Pettitt, 1976); however, any other two-sample distribution-free tests can also be used. Our simulation results show that the method performs well, and in cases where the approach of Hall and Van Keilegom (2007) applies, our proposed test is considerably more powerful.

2. Two Sample Testing for Functional Data

2.1. Preliminary

To begin with, we describe how to test hypothesis (2) under the assumption that the curves are observed entirely and without noise (Hall et al., 2006). Extension to practical settings is discussed in Section 3. Consider two sets of independent curves $\{X_{11}(\cdot), \ldots, X_{1n_1}(\cdot)\}$ and $\{X_{21}(\cdot), \ldots, X_{2n_2}(\cdot)\}$, defined on $[0, 1]$, where $X_{1i}(\cdot) \sim X_1(\cdot)$ and $X_{2i}(\cdot) \sim X_2(\cdot)$ are square integrable and have continuous mean and covariance functions respectively.

The large sample validity of our methodology is developed under the assumption that both $n_1, n_2 \to \infty$ such that $\lim_{n_1,n_2\to\infty} n_1/(n_1 + n_2) = p \in (0, 1)$. Let

$X(\cdot)$ be the mixture process of $X_1(\cdot)$ and $X_2(\cdot)$ with mixture probabilities p and $1-p$ respectively. Furthermore, let Z be a binary random variable taking values in $\{1,2\}$ such that $P(Z=1)=p$. Then $X_1(\cdot)$ is the conditional process $X(\cdot)$ given $Z=1$, and $X_2(\cdot)$ is the conditional process $X(\cdot)$ given $Z=2$. It follows that $X(\cdot)$ is square integrable and its marginal distribution has continuous mean and positive semi-definite covariance functions. Let $\mu(t)=E[X(t)]$ be the (marginal) mean function and let $\Sigma(t,s)=\text{cov}\{X(t),X(s)\}$ be the (marginal) covariance function of $X(\cdot)$. Mercer's theorem yields the spectral decomposition of the marginal covariance function, $\Sigma(t,s)=\sum_{k\geq 1}\lambda_k\phi_k(t)\phi_k(s)$ in terms of non-negative eigenvalues $\lambda_1\geq\lambda_2\geq\ldots\geq 0$ and orthogonal eigenfunctions $\phi_k(\cdot)$, with $\int_0^1\phi_k(t)\phi_{k'}(t)dt=1(k=k')$, where $1(k=k')$ is the indicator function which is 1 when $k=k'$ and 0 otherwise (Bosq, 2000). We refer to $\{\phi_k(\cdot)\}_k$ as the *marginal eigenbasis*, and to λ_k's as the corresponding *marginal eigenvalues*. The decomposition implies that $X(\cdot)$ can be represented via Karhunen-Loève (KL) expansion as $X(t)=\mu(t)+\sum_{k=1}^{\infty}\xi_k\phi_k(t)$ where $\xi_k=\int_0^1\{X(t)-\mu(t)\}\phi_k(t)dt$ are commonly called FPC scores and are uncorrelated random variables with zero mean and variance equal to λ_k. For practical as well as theoretical reasons (see for example Yao et al. (2005), Hall et al. (2006), or Di et al. (2009)), the infinite expansion of $X(\cdot)$ is often truncated. Let $X^K(t)=\mu(t)+\sum_{k=1}^{K}\xi_k\phi_k(t)$ be the truncated KL expansion of $X(\cdot)$. It follows that $X^K(t)$ converges to $X(t)$ uniformly in quadratic mean as $n\to\infty$. Define $X_z^K(t)=\mu(t)+\sum_{k=1}^{K}\xi_{zk}\phi_k(t)$ to be the finite-dimensional approximation of $X_z(t)$ such that $\xi_{zk}=\int_0^1\{X_z(t)-\mu(t)\}\phi_k(t)dt$ for $k\geq 1$ and $z=1,2$. Let $K=K_n$ be a suitably large integer such that $X^K(\cdot)$ approximates $X(\cdot)$ accurately using L^2 norm. It follows that $X_z^K(\cdot)$ approximates $X_z(\cdot)$ well for $z=1,2$. Since $\{\phi_k(\cdot)\}_k$ are the eigenbasis of the covariance of the marginal distribution of $X(\cdot)$ we refer to the analysis based on this basis as '*marginal FPCA*'. Then the null hypothesis in (2) reduces to

$$H_0^K:\{\xi_{1k}\}_{k=1}^K\stackrel{d}{=}\{\xi_{2k}\}_{k=1}^K; \tag{3}$$

where the superscript K in H_0^K emphasizes the dependence of the reduced null hypothesis on the finite truncation K. See the Supplementary Material for further justification.

One possible approach to test hypothesis (3) is to consider two-sample multivariate procedures; see for example Wei and Lachin (1984), Schilling (1986) or Bohm and Zech (2010), Ch.10. For simplicity, we consider multiple two-sample univariate tests combined with a multiple comparison adjustment (e.g. a Bonferroni correction). In particular, testing the null hypothesis (3) can be conducted by testing of the null hypotheses H_{k0}^K, for $k = 1, \ldots, K$, where

$$H_{k0}^K : \xi_{1k} \stackrel{d}{=} \xi_{2k}. \tag{4}$$

There are several common univariate two sample tests: for example the Kolmogorov-Smirnov test (KS, Massey Jr (1951)) or the Anderson-Darling test (AD, Pettitt (1976)). The KS and AD tests are both capable of detecting higher order moment shifts between two univariate distributions, by using differences in the empirical cumulative distributions. Empirical studies have shown that the AD test tends to have higher power than the KS test (Stephens, 1974; Bohm and Zech, 2010). Thus, we present the proposed testing procedure using the AD test.

Recall that the $X_{1i}(\cdot)$'s and the $X_{2i}(\cdot)$'s are two samples of independent and identically distributed curves observed from $X_1(\cdot)$ and $X_2(\cdot)$ respectively, and assume that both the mean function, $\mu(t)$, and the eigenbasis $\{\phi_k(\cdot)\}_{k\geq 1}$ of the mixture process $X(\cdot)$ are known. Then, the corresponding basis coefficients, ξ_{zik}, can be determined as $\xi_{1ik} = \int\{X_{1i}(t) - \mu(t)\}\phi_k(t)dt$ and $\xi_{2ik} = \int\{X_{2i}(t) - \mu(t)\}\phi_k(t)dt$. Let $\widetilde{F}_{1k}(\cdot)$ and $\widetilde{F}_{2k}(\cdot)$ be the corresponding empirical conditional distribution functions of $\{\xi_{1ik}\}_i$ and $\{\xi_{2ik}\}_i$ respectively. The AD test statistic is defined as,
$AD_k^2 = (n_1 n_2/n)\int_{-\infty}^{\infty}\{\widetilde{F}_{1k}(x) - \widetilde{F}_{2k}(x)\}^2/[\widetilde{F}_k(x)\{1 - \widetilde{F}_k(x)\}]d\widetilde{F}_k(x)$, where $n = n_1 + n_2$ and $\widetilde{F}_k(x) = \{n_1\widetilde{F}_{1k}(x) + n_2\widetilde{F}_{2k}(x)\}/n$ (Pettitt, 1976; Scholz and Stephens, 1987). Under the null hypothesis H_{k0}^K of (4), AD_k^2, converges to the same limiting distribution as the AD test statistic for one sample (Pettitt, 1976). Given a univariate two-sample test, we define an α-level testing procedure to test hypothesis (2) as

follows: hypothesis (2) is rejected if $\min_{1\leq k\leq K} p_k \leq (\alpha/K)$, where p_k is the p-value obtained using the chosen univariate two sample test for H_{k0}, for $k = 1, \ldots, K$. The use of the Bonferroni correction ensures that the testing procedure maintains its nominal size, conditional on the truncation level K. Because we apply it to functional data we call this test the *Functional Anderson-Darling (FAD)*. The proposed testing methodology allows us to extend any univariate testing to the case of functional data. Of course, any advantages or drawbacks of the univariate tests, such as the ability to detect higher order moment shifts or weak power in small sample sizes, will carry over to the functional extension.

3. Extension to Practical Situations

Extending the testing procedure described in Section 2.1 to practical applications is not straightforward, as the true smooth trajectories $X_i(\cdot)$, and the true scores ξ_{ik}, are not directly observable. In the DTI study, a subject's observed data, say FA for the 93 locations along the CCA, are noisy measurements of the smooth subject-specific FA profile along CCA evaluated at the 93 locations. In this case, we propose to replace ξ_{zik} from the previous section with appropriate estimators, $\widehat{\xi}_{zik}$. Thus we test the hypotheses H_{0k} in (4) using the $\widehat{\xi}_{zik}$'s instead of the ξ_{zik}'s for $z = 1, 2$. Define $\widehat{AD}_k^2$ to be the estimate for AD_k^2, obtained by replacing ξ_{1ik} and ξ_{2ik} with the basis coefficients $\widehat{\xi}_{1ik}$ and $\widehat{\xi}_{2ik}$, respectively. If the null hypothesis H_{k0} (4) is true, then it is expected that the asymptotic distribution of $\widehat{AD}_k^2$ is the same as the asymptotic null distribution of AD_k^2.

Our logic is based on the result that under null hypothesis (2) $\widehat{\xi}_{1ik} - \widehat{\xi}_{2ik} \xrightarrow{p} 0$ as $n \to \infty$ where "$\xrightarrow{p}$" denotes convergence in probability, for $k = 1, \ldots, K$. Thus, to test (4), one can use the testing procedure described in Section 2, but with the estimated basis coefficients $\widehat{\xi}_{1ik}$ and $\widehat{\xi}_{2ik}$ instead of the true ones.

3.1. Dense design

First, consider the situation when the grid of points for each subject is dense in $[0, 1]$, that is m_{1i} and m_{2i} are very large. Zhang and Chen (2007) proved that one

can reconstruct the curves $X_i(t)$ with negligible error by smoothing the observed functional observations $\{Y_{i1}, \ldots, Y_{im_i}\}$ using local polynomial kernel smoothing. Let $\widehat{X}_{1i}(\cdot)$ and $\widehat{X}_{2i}(\cdot)$ be the reconstructed trajectories in group one and two respectively.

Consider the pooled sample $\{\widehat{X}_{1i}(\cdot) : i = 1, \ldots, n_1\} \cup \{\widehat{X}_{2i}(\cdot) : i = 1, \ldots, n_2\}$, and let $\widehat{X}_i(t)$ be a generic curve in this sample. Let $\widehat{\mu}(t)$ be the (marginal) sample average and let $\widehat{\Sigma}(t, s)$ be the (marginal) sample covariance functions of the reconstructed trajectories $\widehat{X}_i(t)$. Under regularity assumptions, these sample estimators are asymptotically identical to the ideal estimators based on the true trajectories (Zhang and Chen, 2007). The spectral decomposition of the estimated marginal covariance yields the pairs of estimated marginal eigenfunctions and eigenvalues $\{\widehat{\phi}_k(\cdot), \widehat{\lambda}_k\}_k$, with $\lambda_1 > \lambda_2 > \ldots \geq 0$. It follows that $\widehat{\xi}_{ik} = \int \{\widehat{X}_i(t) - \widehat{\mu}(t)\}\widehat{\phi}_k(t)dt$ are consistent estimators of the FPC scores ξ_{ik} (Hall et al., 2006; Zhang and Chen, 2007); $\widehat{\xi}_{1ik} = \widehat{\xi}_{ik}$ if $Z_i = 1$ and $\widehat{\xi}_{2ik} = \widehat{\xi}_{ik}$ if $Z_i = 2$, where Z_i denotes the group membership of the ith curve. Thus, for large sample sizes n_1 and n_2, the distribution of $\widehat{\xi}_{zik}$ approximates that of ξ_{zik}; $\widehat{\xi}_{zik}$ is used to test the hypothesis (3). In applications, $\widehat{\xi}_{ik}$ can be calculated via numerical integration. The finite truncation K, of the estimated eigenfunctions $\{\widehat{\phi}_k(\cdot)\}_k$, can be chosen using model selection based-criteria. We found that the commonly used cumulative explained variance criterion (Di et al., 2009; Staicu et al., 2010) works very well, in practice.

3.2. Sparse design

Next, consider the situation when the grid of points for each subject is sparse (m_{1i}, m_{2i} are as small as few observations) but the overall sets of observed points $\{t_{1ij} : j = 1, \ldots, m_{1i}, i = 1, \ldots, n_1\}$ and $\{t_{2ij} : j = 1, \ldots, m_{2i}, i = 1, \ldots, n_2\}$ are dense sets in $[0, 1]$. The sparse setting requires different methodology for several reasons. First, the bounding constraint on the number of repeated observations m_{1i} and m_{2i} implies a sparse setting at the curve level and does not provide accurate estimators by smoothing each curve separately. Secondly, estimation of the basis coefficients ξ_{ik} via numerical integration is no longer reliable. Instead, we consider the pooled sample $\{(t_{1ij}, Y_{1ij}) : j\}_{i=1}^{n_1} \cup \{(t_{2ij}, Y_{2ij}) : j\}_{i=1}^{n_2}$, where

$\{(t_{ij}, Y_{ij}) : j = 1, \ldots, m_i\}$ is a generic functional observation in the pooled set. The observed measurements $\{Y_{i1}, \ldots, Y_{im_i}\}$ are viewed as the evaluations of a latent process $X_i(\cdot)$ at time points $\{t_{i1}, \ldots, t_{im_i}\}$ and are contaminated with error. Specifically, it is implied that $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$, for $X_i(\cdot)$, the mixture process as described earlier with $E[X(t)] = \mu(t)$ and $\text{cov}\{X(t), X(s)\} = \Sigma(t, s)$. Here the ϵ_{ij}'s are independent and identically distributed (iid) measurement errors with zero mean and variance σ^2.

Common FPCA-techniques can be applied to reconstruct the underlying subject-trajectories, $\widehat{X}_i(\cdot)$ from the observed $\{(t_{ij}, Y_{ij}) : j = 1, \ldots, m_i\}$ (Yao et al., 2005; Di et al., 2009). The key idea is to first obtain estimates of the (marginal) smooth mean and covariance functions, $\widehat{\mu}(t)$ and $\widehat{\Sigma}(t, s)$ respectively. The spectral decomposition of the estimated marginal covariance yields the marginal eigenfunction/eigenvalue pairs, $\{\widehat{\phi}_k(\cdot), \widehat{\lambda}_k\}_{k \geq 1}$, where $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \ldots \geq 0$. Next, the variance of the noise is estimated based on the difference between the pointwise variance of the observed Y_{ij}'s and the estimated pointwise variance $\widehat{\Sigma}(t, t)$ (Staniswalis and Lee, 1998; Yao et al., 2005). There are several methods in the literature to select (or estimate) the finite truncation K, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). From our empirical experience the simple criterion based on percentage of explained variance (such as 90% or 95%) gives satisfactory results in terms of recovering the smooth latent process. In our simulation experiments we use the cumulative explained variance to select K. Sensitivity with regards to this parameter is studied in Section 4.

Once the marginal mean function, marginal eigenfunctions, eigenvalues, and noise variance are estimated, the model for the observed data $\{Y_{ij} : 1 \leq j \leq m_i\}$ becomes a linear mixed effects model $Y_{ij} = \widehat{\mu}(t_{ij}) + \sum_k \xi_{ik} \widehat{\phi}_k(t_{ij}) + \epsilon_{ij}$, where $var(\xi_{ik}) = \widehat{\lambda}_k$ and $var(\epsilon_{ij}) = \widehat{\sigma}^2$. The coefficients ξ_{ik} can be predicted using the conditional expectation formula $\widehat{\xi}_{ik} = \widehat{E}[\xi_{ik} | (Y_{i1}, \ldots, Y_{im_i})]$. Under the assumption that the responses and errors are jointly Gaussian, the predicted coefficients are in fact the empirical best linear unbiased predictors: $\widehat{\xi}_{ik} = \widehat{\lambda}_k \widehat{\Phi}_i^T (\widehat{\Sigma}_i + \widehat{\sigma}^2 I_{m_i \times m_i})^{-1} (Y_i - \widehat{\mu}_i)$.

Here Y_i is the m_i-dimensional vector of Y_{ij}, $\widehat{\mu}_i$ and $\widehat{\Phi}_i$ are m_i-dimensional vectors with the jth entries $\widehat{\mu}(t_{ij})$ and $\widehat{\phi}(t_{ij})$ respectively, $\widehat{\Sigma}_i$ is a $m_i \times m_i$-dimensional matrix with the (j, j')th entry equal to $\widehat{\Sigma}(t_{ij}, t_{ij'})$, and $I_{m_i \times m_i}$ is the $m_i \times m_i$ identity matrix. Yao et al. (2005) proved that $\widehat{\xi}_{ik}$ converges in probability to $\widetilde{\xi}_{ik} = E[\xi_{ik}|(Y_{i1}, \ldots, Y_{im_i})]$ as n grows to infinity. Define $\widetilde{\xi}_{1ik} = \widetilde{\xi}_{ik}$ and $\widehat{\xi}_{1ik} = \widehat{\xi}_{ik}$ if $Z_i = 1$; similarly define $\widetilde{\xi}_{2ik}$ and $\widehat{\xi}_{2ik}$. If $\xi_{1i} \stackrel{d}{=} \xi_{2i}$ then $\widetilde{\xi}_{1i} \stackrel{d}{=} \widetilde{\xi}_{2i}$; the reverse implication is true under a joint Gaussian assumption. It follows that for large sample sizes n_1 and n_2 the sampling distributions of $\widehat{\xi}_{1ik}$ and $\widehat{\xi}_{2ik}$ are asymptotically close to those of ξ_{1ik} and ξ_{2ik}, respectively. Therefore, we can use $\widehat{\xi}_{zik}$ to test hypothesis (4). This approach is employed in the analysis of our motivating data. In general, we expect the sparsity of the sampling design to affect the rejection probability. The power to correctly reject the null hypothesis, when the curves are observed on sparse sampling design would typically be smaller than when the curves are observed on dense design. Of course the magnitude of this effect also depends on the smoothness of the true process.

4. Simulation Studies

The performance of the proposed testing procedure is studied under a variety of settings and for varying sample sizes. We focus on a small number of FPCs first and later study more complex data settings. Section 4.1 studies the Type I error rate of the FAD test, and the sensitivity to sample size and the percentage of explained variance, τ, used to estimate the parameter, K. Section 4.2 provides a numerical comparison of the proposed approach with the closest available competitor - the Cramér-von Mises (CVM) -type test introduced by Hall and Van Keilegom (2007).

4.1. Type One Error and Power Performance

We construct datasets $\{(t_{1ij}, Y_{1ij}) : j\}_{i=1}^{n_1}$ and $\{(t_{2ij}, Y_{2ij}) : j\}_{i=1}^{n_2}$ using model (1) for $t_{1ij} = t_{2ij} = t_j$ observed on an equally spaced grid of $m = 100$ points in $[0, 1]$. Here $X_{1i}(t) = \mu_1(t) + \sum_k \phi_{1k}(t)\xi_{1ik}$ and $X_{2i}(t) = \mu_2(t) + \sum_k \phi_{2k}(t)\xi_{2ik}$, where $\phi_{11}(t) = \phi_{21}(t) = \sqrt{2}\sin(2\pi t)$, $\phi_{12}(t) = \phi_{22}(t) = \sqrt{2}\cos(2\pi t)$ and so on, are the

Fourier basis functions, ξ_{1ik} and ξ_{2ik} are uncorrelated respectively with $var(\xi_{1ik}) = \lambda_{1k}$, $var(\xi_{2ik}) = \lambda_{2k}$ and $\lambda_{1k} = \lambda_{2k} = 0$ for $k \geq 4$. We set $\epsilon_{1ij} \sim N(0, 0.25)$ and $\epsilon_{2ij} \sim N(0, 0.25)$. Setting $\phi_{zk}(t) = \phi_k(t)$ allows us to study different types of departures from the null hypothesis, that the underlying processes of the two datasets are the same. In Section 4.2 we consider generating processes $X_{1i}(\cdot)$ and $X_{2i}(\cdot)$ that have different basis functions representations.

The FAD test is employed to test the null hypothesis in (2); the marginal mean functions, the marginal eigenbasis functions, and the corresponding basis coefficients are estimated using the methods described in Section 3. The number of marginal eigenbasis functions is estimated using the percentage of explained variance, τ, for the pooled data. The estimates for all the model components, including the basis coefficients, are obtained using the `R` (Team, 2015) package `refund` (Crainiceanu et al., 2012). The `R` package `AD` (Scholz, 2011) is used to test the equality of the corresponding univariate distributions for each pair of basis coefficients. The Bonferroni multiple testing correction is used to maintain the desired level of the FAD test. The null hypothesis is rejected/not rejected according to the approach described in Section 2.1. All the results in this section are based on $\alpha = 0.05$ significance level.

First, we assess the Type I error rate for various threshold parameter values, τ. For simplicity, we set $\mu_1(t) = \mu_2(t) = 0$ for all t and consider $\xi_{1ik}, \xi_{2ik} \sim N(0, \lambda_k)$, for $\lambda_1 = 10$, $\lambda_2 = 5$, and $\lambda_3 = 2$ and $\lambda_k = 0$ for $k > 3$. The Type I error rate is studied for varying values of τ from 80% to 99% and for increasing equal/unequal sample sizes. For the simulated data, the proposed criterion to select the number of marginal eigenfunctions K yields $K = 2$ if $\tau = 80\%$ and $K = 3$ when $\tau = 99\%$. Different values for τ may result in different choices for K, which in turn may lead to different Type I error estimates. Table 1 displays the empirical size of the FAD test using varying thresholds τ when the total sample size ranges from 200 to 2000. The results are based on 5000 MC replications. They show that the size of the test is close to or slightly above the nominal level and is not sensitive to τ.

Next, we set $\tau = 0.95$ and study the power performance of the FAD test. Setting

the percentage of explained variance to be too high, or implicitly selecting an unnecessarily large number of components may result in a loss of power of the testing procedure. In particular, if we set the threshold value at a level that results in K marginal eigenfunctions, and the difference between the sets of curves is captured by the distribution of the coefficients of the $(K+1)$th eigenfunction, then the proposed testing procedure does not have any power.

The distribution of the true processes under the alternative is described by the mean functions, as well as by the distributions of the basis coefficients. The following scenarios allow us to study the FAD test for specific types of changes in the two data sets. Settings A, B and C correspond to deviations in the first, second and third moments, respectively, of the corresponding distribution of the first set of basis coefficients. Throughout this section it is assumed that $\lambda_{1k} = \lambda_{2k} = 0$ for all $k \geq 3$. **[A] Mean Shift:** Set the mean functions as $\mu_1(t) = t$ and $\mu_2(t) = t + \delta t^3$. Generate the coefficients as $\xi_{1i1}, \xi_{2i1} \sim N(0, 10)$, $\xi_{1i2}, \xi_{2i2} \sim N(0, 5)$. The index δ controls the departure in the mean behavior of the two distributions. **[B] Variance Shift:** Set $\mu_1(t) = \mu_2(t) = 0$. Generate the coefficients $\xi_{1i1} \sim N(0, 10)$, $\xi_{2i1} \sim N(0, 10 + \delta)$, and $\xi_{1i2}, \xi_{2i2} \sim N(0, 5)$. Here δ controls the difference in the variance of the first basis coefficient between the two data sets. **[C] Skewness Shift:** $\xi_{1i1} \sim T_4(0, 10)$ and $\xi_{2i1} \sim ST_4(0, 10, 1 + \delta)$, and $\xi_{1i2}, \xi_{2i2} \sim T_4(0, 5)$. Here, $T_4(\mu, \sigma)$ denotes the common student T distribution with 4 degrees of freedom, that is standardized to have mean μ and standard deviation σ and $ST_4(\mu, \sigma, \gamma)$ is the standardized skewed T distribution (Wurtz et al., 2006) with 4 degrees of freedom, mean μ, standard deviation σ, and shape parameter $0 < \gamma < \infty$, which controls skewness. The `rstd` and `rsstd` functions in the `R` package `fgarch` (Wurtz et al., 2006) are used to generate random data from a standardized T and standardized skewed T distribution respectively. The shape parameter γ is directly related to the skewness of this distribution and the choice $\gamma = 1$ corresponds to the symmetric T distribution. In this case, the index δ controls the difference in the skewness of distribution of the first basis coefficient.

For all the settings, $\delta = 0$ corresponds to the null hypothesis that the two samples of curves have the same generating distribution, whereas $\delta > 0$ corresponds to the alternative hypothesis that the two sets of curves have different distributions. Thus, δ represents the departure from the null hypothesis and it will be used to characterize empirical power curves. The estimated power is based on 1000 MC replications. Results are presented in Figure 2 for the case of equal/unequal sample sizes in the two groups, and for various total sample sizes.

Column A of Figure 2 displays the empirical power curves of the FAD test when the mean discrepancy index δ ranges from 0 to 8. It appears that the performance of the power is affected more by the combined sample size, $n = n_1 + n_2$, than the magnitude of each sample size n_1 or n_2. Column B shows the empirical power, when the variance discrepancy index δ ranges from 0 to 70. The empirical power increases at a faster rate for equal sample sizes than unequal sample sizes, when the total sample size is the same. However, the differences become less pronounced as the total sample size increases. Finally, column C displays the power behavior for observed data generated under setting C for δ between 0 and 6. For moderate sample sizes, irrespective of their equality, the probability of rejection does not converge to 1 no matter how large δ is; see the results corresponding to a total sample size equal to $n = 200$ or 400. This is in agreement with our intuition that detecting differences in higher order moments of the distribution becomes more difficult and requires increased sample sizes. In contrast, for larger total sample sizes, the empirical power curve has a fast rate of increase.

Following a suggestion by an anonymous reviewer, we numerically compared our testing procedure with a simpler test for the mean; we used the L^2-based mean test proposed in Benko et al. (2009). The two testing procedures are developed for different null hypotheses; to compare them for testing the null hypothesis that the distribution of the two curves is the same across groups, the mean detection test requires an additional working assumption that the distribution of the curves in each group differs by at most a mean shift. In practice one typically does not

know if the way two sets of curves vary is captured solely by the mean function. Thus, we generate two sets of curves as described in settings A (mean shift) and B (variance shift) above. In both cases we record the rejection probabilities. Note that the necessary assumption for the mean test is true for setting A but not true for B. One expects a higher power for the mean test in setting A, as this test uses the additional information that the distribution of the sets of curves differ solely in the mean function, while our proposed test does not use such information. On the contrary, when the distribution of the curves does not differ just by the mean function, the loss of power may be substantial. In setting B, the mean of the two distributions is the same across groups. Therefore, the test for the mean does not have power to detect the alternative hypothesis. These additional results are included in the Supplementary Material, due to space limitation. In the DTI study, we use the FAD test for initial screening and consider the "targeted" test for the mean as a second stage analysis.

4.2. Comparison with available approaches

To our best of our knowledge, the work of Hall and Van Keilegom (2007) is the only available alternative that considers testing that the distributions of two samples of curves are the same, when the observed data are noisy, discrete and irregular realizations of the latent curves. In this section, we compare the performance of our proposed FAD test compared to their Cramer-von Mises (CVM) - type test, based on the empirical distribution functions after local-polynomial smoothing of the two samples of curves.

We generate data $\{(t_{1ij}, Y_{1ij}) : j\}_{i=1}^{n_1}$ and $\{(t_{2ij}, Y_{2ij}) : j\}_{i=1}^{n_2}$ exactly as in Hall and Van Keilegom (2007), and for completeness we describe it next: $Y_{1ij} = X_{1i}(t_{1ij}) + \epsilon_{1ij}$ and $Y_{2ij} = X_{2i}(t_{2ij}) + \epsilon_{2ij}$, where , $\epsilon_{1ij} \sim N(0, 0.01)$, $\epsilon_{2ij} \sim N(0, 0.09)$, $X_{1i}(t) = \sum_{k=1}^{15} e^{-k/2} N_{k1i}\psi_k(t)$ and $X_{2i}(t) = \sum_{k=1}^{15} e^{-k/2} N_{k21i}\psi_k(t) + \delta \sum_{k=1}^{15} k^{-2} N_{k22i}\psi_k^*(t)$, such that $N_{k1i}, N_{k21i}, N_{k22i} \overset{iid}{\sim} N(0,1)$. Here $\psi_1(t) \equiv 1$ and $\psi_k(t) = \sqrt{2}sin\{(k-1)\pi t\}$ are orthonormal basis functions. Also $\psi_1^*(t) \equiv 1$, $\psi_k^*(t) = \sqrt{2}sin\{(k-1)\pi(2t-1)\}$ if $k > 1$ is odd and $\psi_k^*(t) = \sqrt{2}cos\{(k-1)\pi(2t-1)\}$

if $k > 1$ is even. As before, the index δ controls the deviation from the null hypothesis; $\delta = 0$ corresponds to the null hypothesis, that the two samples have identical distribution. Finally, the sampling design for the curves is assumed balanced ($m_1 = m_2 = m$), but irregular, and furthermore different across the two samples. Specifically, it is assumed that $\{t_{1ij} : 1 \leq i \leq n_1, 1 \leq j \leq m_1\}$ are iid realizations from $Uniform(0, 1)$, and $\{t_{2ij} : 1 \leq i \leq n_2, 1 \leq j \leq m_2\}$ are iid realizations from the distribution with density $0.8 + 0.4t$ for $0 \leq t \leq 1$. Two scenarios are considered: i) $m = 20$ points per curve, and ii) $m = 100$ points per curve. Notice this is an example where different basis function expansions are used for $X_{1i}(\cdot)$ and $X_{2i}(\cdot)$ for $\delta > 0$. Figure S2 displays an example of data set $X_{1i}(\cdot)$ for $\delta = 0$ and $X_{2i}(\cdot)$ for $\delta = 1$ (top panels), and the noisy evaluations at irregular grids (bottom panels).

The null hypothesis that the underlying distribution is identical in the two samples is tested using CVM (Hall and Van Keilegom, 2007) and FAD testing procedures for various values of δ. Figure 3 illustrates the comparison between the approaches for significance level $\alpha = 0.05$; the results are based on 500 Monte Carlo replications. In practice, we observe that as the number of observations on each curve decreases, the empirical size of the test seems to increase slightly. The CVM test is conducted using the procedure described in Hall and Van Keilegom (2007), and the p-value is determined based on 250 bootstrap replicates; the results are obtained using the `R` code provided by the authors. To apply the marginal FPCA and determine the marginal eigenbasis, we use the `refund` package (Crainiceanu et al., 2012) in `R`, which requires that the data are formatted corresponding to a common grid of points, with possible missingness. Thus, a pre-processing step is necessary. For each scenario, we consider a common grid of $m = 100$ equally spaced points in $[0, 1]$ and bin the data of each curve according to this grid. This procedure introduces missingness for the points where data are not observed. This pre-processing step is not necessary if one uses `PACE` package (Yao et al., 2005) in `Matlab`. However, our preference for using open-source software motivates the use of `refund`. Comparison of `refund` and `PACE` revealed that the two methods lead to similar results when

smoothing trajectories from noisy and sparsely observed 'functional' data.

As Figure 3 illustrates, both procedures maintain the desired level of significance. However, the empirical power of the FAD test increases at a faster rate than the CVM test (Hall and Van Keilegom, 2007) under all the settings considered. This should not seem surprising, since representing the data using the marginal eigenbasis expansion, as detailed in Section 3, removes extraneous components. In contrast, the CVM test attempts to estimate all basis functions by smoothing the data. This could lead to cumulative estimation errors that can ultimately lower the power of the test. Additionally, due to the usage of bootstrapping to approximate the null distribution of the test, the CVM test has a much higher computational burden than the FAD test. Thus we restrict our simulation studies to 500 MC replications.

5. Diffusion Tensor Image Data Analysis

We now return to our motivating brain tractography application. Recall that for each patient we measure the water diffusivity along five well identified tracts - right/left corticospinal tract (rCST, lCST), corpus callosum (CCA), and right/left optic radiation tract (rOPR, lOPR) - by two common DTI modalities fractional anisotropy (FA) and parallel diffusivity (L0). Three of these 10 tract-modality profiles are available in the `R`-package refund (Crainiceanu et al., 2012). Our interest is to study if FA or L0 profiles along the five tracts have different distributions for subjects affected by MS when compared to controls. Such an assessment would provide researchers with valuable information about the neurodegeneration of MS caused by the axonal disruption along the white mater tracts and the usefulness of the commonly acquired modalities to capture this process. In the following, we focus first on the CCA, as measured by FA and L0, since there is scientific evidence that MS, in advanced stages, is associated with significant neuronal loss in the corpus callosum (Ozturk et al., 2010).

5.1. Corpus Callosum

The top four panels in Figure 1 display the FA profiles (left most two panels) and L0 profiles (right most two panels) sampled at 93 locations along CCA for the 162 MS patients and 42 healthy controls. Though the modalities are observed on a regular sampling design, there are missing observations and the data are likely corrupted with measurement error. It seems reasonable to suspect that the distribution of the modalities has a relatively similar mean profile for MS cases compared to controls. Furthermore, it seems the pointwise variance in the two groups may be different though this observation may be an artifact of the larger group size of the MS patients compared to the healthy controls. More importantly, L0 seems to exhibit group-specific distributional characteristics that vary across the locations along the tract (such as skewness). Staicu et al. (2012) investigated this point in detail and proposed different semi-parametric Gaussian copula-based models to describe the distribution of the L0 profiles along CCA (CCA-L0) in MS patients and healthy controls. Based on bootstrapping the subjects, they inferred that the marginal distribution parameters - pointwise mean, variance and skewness functions - are significantly different in MS cases and controls. In our analysis we go one step further and formally assess whether the distribution of the CCA-L0 profiles in the MS population is different than the distribution of the CCA-L0 profiles in the healthy controls, without making any parametric assumptions.

We begin with the study of the CCA-FA profiles and test the null hypothesis that the CCA-FA profiles have the same distribution for the MS patients and controls. We also investigate if there are differences in the mean of CCA-FA between cases and controls using the L^2-based mean test presented in (Benko et al., 2009); this can be considered as a more 'targeted' testing procedure. The null distribution of the test is approximated by bootstrapping the curves. The method does not directly account for missingness. Therefore, to employ it for our application we need some pre-processing. Specifically, we reconstruct the latent smooth profiles within each group over equally spaced grids. To do this, we use the *fpca.sc* function

in the `R` package `refund` with the percentage of explained variance parameter set to 99% (Crainiceanu et al. (2012)). The p-value with this approach is based on 5000 bootstrap replications. The p-value is 0.00, indicating that the means profiles are statistically different in the two groups.

We turn next to the CCA-L0 profiles and study whether they vary according to the same distribution in the MS patients compared to controls; we use FAD as described above. The analysis follows roughly the same procedure as above; to avoid repetition we focus only on the part that is new. Five eigenfunctions are selected to estimate 95% of the total variability of the combined dataset. Figure S3 displays the leading three eigenfunctions, along with the boxplots of the corresponding coefficients presented separately for the MS and control groups. The leading eigenfunction accounts for 76% of the total variability; it is negative and has a concave shape with a dip around location 60 of the CCA. This component gives the direction along which the two curves differ the most. Near location 60, the distribution of the CCA-L0 is highly skewed for the MS group, but not as skewed as in the control group. Examination of the boxplot of the coefficients corresponding to the first eigenfunction (left, bottom panel of Figure S3) shows that most healthy individuals (controls) are positive, yielding CCA-L0 profiles that are lower than the population average profile. In contrast, over half of the MS cases have a negative coefficient for this function, leading to CCA-L0 profiles that are larger than the population average. Our findings are consistent with Staicu et al. (2012), who discuss possible reasons for which L0 tends to be higher in the MS cases than controls. We apply the AD testing procedure to assess whether the distribution of the coefficients is different for the two groups. The p-values of the univariate tests are $p_1 = 0.00001$, $p_2 = 0.07165$, $p_3 = 0.02479$, $p_4 = 0.19887$, and $p_5 = 0.29161$, leading to the p-value $p = 5 \times \min_{1 \leq k \leq 5} p_k = 0.00005$ of the FAD test. This shows significant evidence that the parallel diffusivity has a different distribution in MS subjects compared to controls. The L^2-based mean test (Benko et al., 2009) is used to assess if the mean of the CCA-L0 profiles is the same in the two groups; the p-value is equal to 0.00,

suggesting significant difference in the mean profiles.

Overall the results provide strong support that along the CCA, the fractional anisotropy and parallel diffusivity properties of the water diffusion vary differently in the MS patients relative to controls. This is in agreement with the scientists' expectations that MS is generally associated with lesions and axonal demyelination on the CCA (Evangelou et al. (2000); Ozturk et al. (2010)). Thus, the water diffusivity is affected along this tract for diseased subjects. The "targeted" mean test (Benko et al. (2009)) yields that both FA and L0 have a different mean along the CCA for MS cases and controls.

5.2. Additional White Mater Tracts

In the attempt to gain insight into the neurodegeneration of the white mater tracts associated with MS, we study whether the measured properties of water diffusivity, FA and L0, vary differently in MS cases and controls over the remaining four tracts. For each tract modality data set, the `fpca.sc` function in the R package `refund` is employed for the marginal FPCA and the percentage of explained variance is fixed at $\tau = 95\%$. Table 2 shows the p-values corresponding to testing the equality of the distributions of the FA and L0 profiles along each of the five tracts. When interpreting the results, Bonferroni correction is used to ensure a family error rate equal to the desired nominal level; the analysis uses 0.05 level of significance.

The results show that the water diffusivity varies differently along the lCST tract in MS patients compared to healthy individuals, as assessed through FA and L0. The p-values for these cases are very small and less than 0.005 - which is the threshold level using the Bonferroni adjustment for 10 comparisons. The situation is not as clear for rCST, lOPR, and rOPR tracts, where only one of the DTI modalities seems to capture the difference in the distribution of the water diffusivity properties. Take for example the rCST tract. When assessing the null hypothesis that the FA profiles vary in the same way for MS patients and controls using the FAD test, we obtain a p-value equal to 0.0103, which does not provide support against this null hypothesis. On the other hand, testing the null hypothesis that the L0 profiles vary

in the same way for MS patients and controls using the FAD test yields a p-value equal to 0.00002, which indicates that if the null hypothesis were true, obtaining such a value of the test would be very unlikely. These two results show evidence that the rCST tract is also affected by MS and that the parallel diffusivity property of the water diffusivity appropriately captures this degeneration. Furthermore, our findings suggest that both the lOPR and the rOPR tracts are associated with MS, and that FA is the more appropriate modality to capture this effect.

As before, we consider a second stage analysis to investigate whether the difference in the distribution of the profiles along various tracts is in the mean, or whether it is in the higher order moments. The second stage analysis is only conducted on the tract-modality profiles for which the FAD test determines significant difference in the distributions. We test the null hypothesis that the means of various modalities along corresponding tracts is the same for MS patients and controls, using the test proposed by Benko et al. (2009) with 5000 bootstrap samples. As discussed above, this procedure requires some pre-processing of the data, to reconstruct the true latent signals. The `fpca.sc` function was used for this purpose with the percentage of explained variance set at 99%. The results show that CCA-FA, lOPR-FA, rOPR-FA, CCA-L0, lCST-L0, and lCST-L0 have mean profiles that are significantly different in MS patients and controls (p-values of 0.00). In contrast the difference between the distribution of lCST-FA in MS subjects and controls seems to be in the higher order moments (p-value of 0.0672).

Earlier versions of this study consisting of data from fewer and possibly different patients were considered in Greven et al. (2011) and Goldsmith et al. (2011). Greven et al. (2011) proposed a longitudinal functional principal component framework analysis for the CCA-FA profiles observed over time. Goldsmith et al. (2011) studied the association between CCA-L0 profiles of MS observed at the baseline and the patients' cognitive scores. In the earlier version of the study, a different registration technique was used for the various tracts: the tracts were sampled at 120 equidistant locations - based on a processing algorithm that used 20 equally spaced points

between known landmarks. In more recent versions of this study (Crainiceanu et al. (2012), Staicu et al. (2012); Gertheiss et al. (2013a); McLean et al. (2014), Ivanescu et al. (2014); Li et al. (2014)) including the one considered in this paper, 93 locations are used for the CCA tract - based on slice numbers in the atlas space after registering images across subjects (Reich et al. (2010)). This process aligns the functions across patients using anatomical characteristics of the brain. Nevertheless, this work is the first one that formally assesses whether the way the several modalities along the main tracts vary is different in MS patients relative to healthy controls.

Perhaps the closest work to ours is that of Gertheiss et al. (2013a), who considered a simultaneous study of the magnetic resonance imaging indices along the various tracts in an attempt to find which ones are important in predicting disease status. Their study included six magnetic resonance imaging indices, including the two DTI modalities, FA and L0. Their results indicated that CCA-FA and CCA-L0 are not deemed important in predicting the disease status. Our presented analysis does find that these tract-modality profiles differ between patients with MS and healthy controls. Gertheiss et al. (2013a) did identify a different modality - perpendicular diffusivity - to be more relevant. Some of our results do seem to agree with theirs, since lCST-FA, rCST-FA, lOPR-FA, rOPR-FA, and rOPR-L0 are found to be associated with MS. However their functional variable selection approach centered and scaled the functional covariates (modalities along tracts), to have pointwise zero mean and unit variance in the overall sample; no such transformation was used by our approach. While their transformation is not required, the comparison of our results with their findings requires further exploration.

6. Discussion

We develop methodology to compare the white matter tract neurodegeneration in MS patients and healthy individuals using state-of-the-art DTI. We formally assess whether two conventional DTI modalities along each of several well identified white matter tracts have different distributions in MS patients compared to controls. Our

findings confirm that the CCA and lCST tracts are associated with the disease. Our results suggest evidence of axonal disruption along the other three tracts in the study - rCST, lOPR and rOPR - though only one of the two DTI measures actually captures it. Furthermore, while the mean of most DTI profiles is one of the reasons (or the reason) that the modalities vary significantly differently in MS patients than controls, the way that FA along the lCST tract varies differently in the two groups is captured by higher order moments.

When dealing with functional data, which theoretically are infinite dimensional objects, it is important to use data reduction techniques. We propose marginal FPCA to represent the curves using the eigenbasis of the marginal covariance of an appropriate mixture process. This reduces the dimension of the testing problem and allows the application of well known lower-dimensional testing procedures. The proposed testing approach combines marginal FPCA and classical univariate procedures, scales well to large samples sizes, and can be easily extended to test the null hypothesis that multiple (more than two) groups of curves have identical distribution. We used Anderson-Darling test and showed through simulation studies that the proposed FAD test is very competitive when compared with alternatives.

Acknowledgments

We thank Daniel Reich and Peter Calabresi for the DTI data and Ingrid Van Keilegom for sharing the R code used in Hall and Van Keilegom (2007). Pomann was funded by the AT&T Graduate Research Fellowship and NSF grant DGE-0946818. Staicu was funded by the NSF grants DMS 1007466 and DMS 0454942 and NIH grants R01 NS085211 and R01 MH086633. Ghosh was funded by the NSF grant DMS-1358556 (IPA assignment) and DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute.

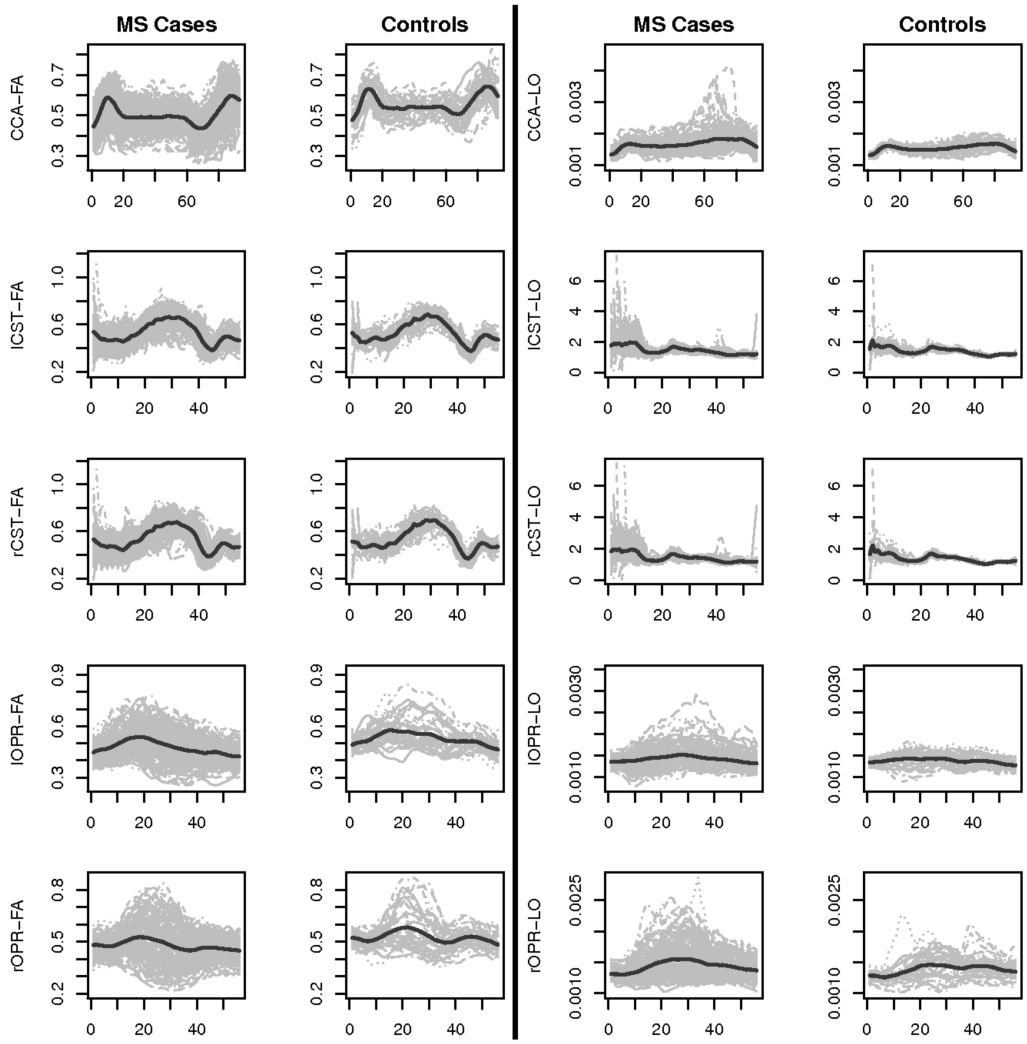

Fig. 1. Top: Fractional anisotropy (left) and parallel diffusivity (right) for MS cases and healthy controls. The pointwise means are displayed by the solid black lines.

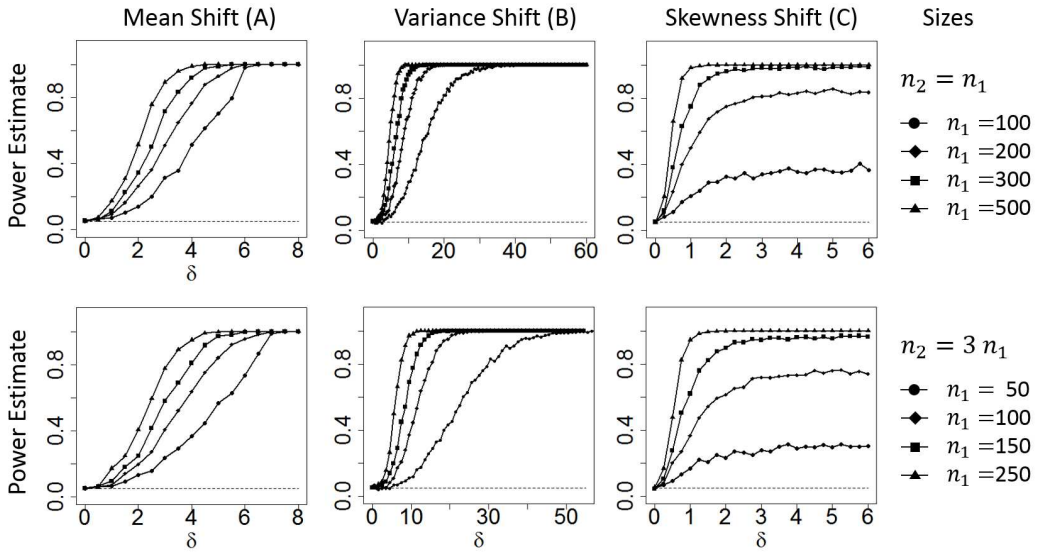

Fig. 2. Empirical power curves under simulation setting A (leftmost panels), B (middle panels) and C (rightmost panels). The top panels consist of the case in which both sets of curves have equal sample sizes. The bottom panels correspond to unequal sample sizes as displayed in the legends. The overall sample size $n = n_1 + n_2$ varies from $n = 200$ to $n = 1000$. The maximum standard error is 0.007.

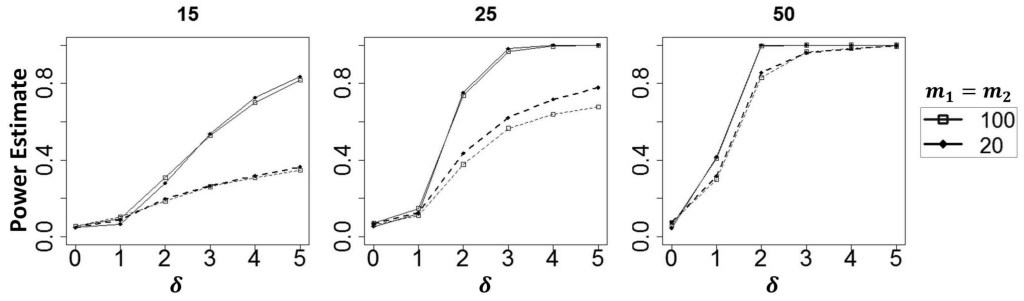

Fig. 3. Estimated power curves for the FAD (solid line) and CVM (dashed) for equal sample sizes, varying from $n_1 = n_2 = 15, 25$ to 50 and for $m_1 = m_2 = 20$ (dot) and $m_1 = m_2 = 100$ (square). Results use 5% significance level and 500 MC replications.

Table 1. Estimated Type I error rate of FAD, based on 5000 replications for several values τ, leading to different estimates of the truncation parameter K.

(n_1, n_2) \ τ	0.80	0.85	0.90	0.95	0.99
(100,100)	0.056	0.056	0.059	0.060	0.060
(200,200)	0.050	0.050	0.052	0.053	0.053
(300,300)	0.053	0.053	0.049	0.049	0.049
(500,500)	0.049	0.049	0.050	0.050	0.050
(1000,1000)	0.055	0.055	0.058	0.058	0.058
(50,150)	0.055	0.055	0.058	0.058	0.057
(100,300)	0.053	0.053	0.049	0.049	0.049
(150,450)	0.048	0.048	0.055	0.054	0.054
(250,750)	0.051	0.051	0.054	0.054	0.054
(500,1500)	0.055	0.055	0.053	0.053	0.053

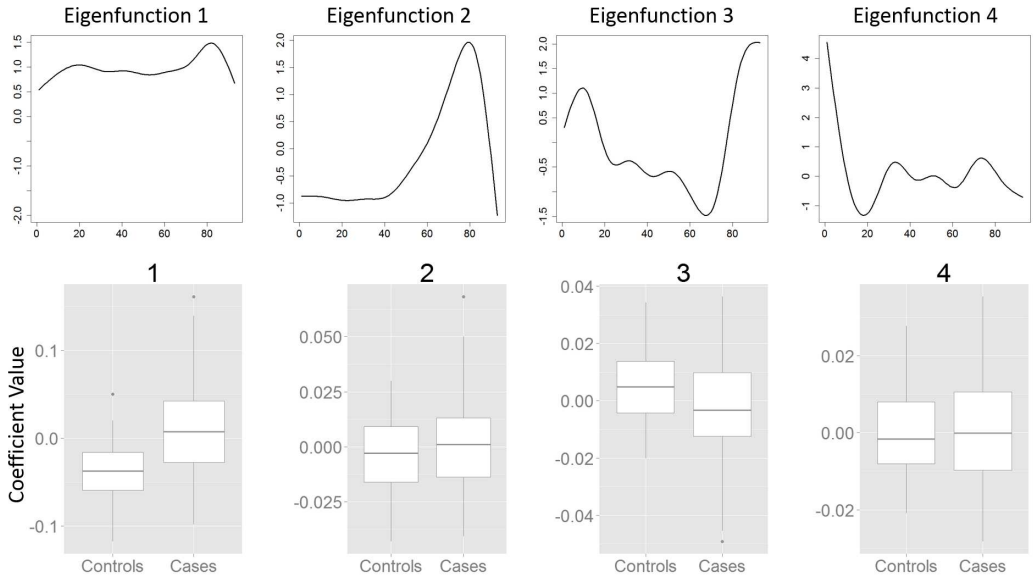

Fig. 4. Fractional Anisotropy (FA). Top: Four leading estimated eigenfunctions that explain over 90% of the total variation in the combined FA data set. Bottom: Boxplots of the first four estimated basis coefficients.

Table 2. FAD-based p-values for testing the same distribution in MS patients and controls.

Tract-modality	CCA-FA	lCST-FA	rCST-FA	lOPR-FA	rOPR-FA
FAD p-value	0.00	0.52×10^{-3}	10.3×10^{-3}	0.01×10^{-3}	0.01×10^{-3}
Tract-modality	CCA-L0	lCST-L0	rCST-L0	lOPR-L0	rOPR-L0
FAD p-value	0.01×10^{-3}	1.95×10^{-3}	0.02×10^{-3}	6.51×10^{-3}	0.01

References

Annette Moller, A., G. Tutz, and J. Gertheiss (2015). Random forests for functional covariates. *in preparation*.

Aslan, B. and G. Zech (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation 75*(2), 109–119.

Benko, M., W. Hardle, and A. Kneip (2009). Common functional principal components. *The Annals of Statistics 37*, 1–34.

Bohm, G. and G. Zech (2010). *Introduction to statistics and data analysis for physicists*. DESY.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, Volume 149. Springer.

Corain, L., V. B. Melas, A. Pepelyshev, and L. Salmaso (2014). New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification 8*(3), 339–356.

Crainiceanu, C., P. Reiss, J. Goldsmith, L. Huang, L. Huo, F. Scheipl, S. Greven, J. Harezlak, M. G. Kundu, and Y. Zhao (2012). refund : Regression with functional data. *R Package 0.1-6*.

Cuevas, A., M. Febrero, and R. Fraiman (2004). An anova test for functional data. *Computational statistics & data analysis 47*(1), 111–122.

Di, C., C. M. Crainiceanu, B. S. Caffo, and N. M. Naresh M. Punjabi (2009). Multilevel functional principal component analysis. *The annals of Applied Statistics 3*(1), 458–488.

Dorris, D. M., J. Cao, J. A. Willett, C. A. Hauser, and J. Meitzen (2014). Intrinsic excitability varies by sex in pre-pubertal striatal medium spiny neurons. *Journal of neurophysiology*, jn–00687.

Estévez-Pérez, G. and J. A. Vilar (2008). Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics 20*, 495–515.

Evangelou, N., D. Konz, M. Esiri, S. Smith, J. Palace, and P. Matthews (2000). Regional axonal loss in the corpus callosum correlates with cerebral white matter lesion volume and distribution in multiple sclerosis. *Brain 123*(9), 1845–1849.

Ferraty, F., P. Vieu, and S. Viguier-Pla (2007). Factor-based comparison of groups of curves. *Computational Statistics & Data Analysis 51*(10), 4903–4910.

Fremdt, S., L. Horvath, P. Kokoszka, and J. Steinebach (2012). Testing the equality of covariance operators in functional samples. *Scand. J. Statist. 40*, 138–152.

Friedman, J. and L. Rafsky (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics 7*, 697–717.

Gertheiss, J., A. Maity, and A.-M. Staicu (2013a). Variable selection in generalized functional linear models. *Stat 2*(1), 86–101.

Gertheiss, J., J. Goldsmith, C. Crainiceanu, and S. Greven (2013b). Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics 14*, 447–461.

Goldsmith, J., J. Bobb, C. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics 20*, 850–851.

Goldsmith, J., C. M. Crainiceanu, B. Caffo, and D. Reich (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 453–469.

Greven, S., C. Crainiceanu, B. Caffo, and D. Reich (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics*, pp. 149–154. Springer.

Hall, P., H.-G. Muller, and J.-L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics 34*, 1493–1517.

Hall, P. and I. Van Keilegom (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica 17*(4), 1511–1531.

Horváth, L., P. Kokoszka, and R. Reeder (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 103–122.

Ivanescu, A. E., A.-M. Staicu, F. Scheipl, and S. Greven (2014). Penalized function-on-function regression. *Computational Statistics*, 1–30.

Kraus, D. and V. Panaretos (2012). Dispersion operators and resistant second-order analysis of functional data. *Biometrika 99*, 813–832.

Laukaitis, A. and A. Račkauskas (2005). Functional data analysis for clients segmentation tasks. *European journal of operational research 163*(1), 210–216.

Li, M., A.-M. Staicu, and H. D. Bondell (2014). Incorporating covariates in skewed functional data models. *Biostatistics*, kxu055.

Massey Jr, F. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association 46*(253), 68–78.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics 23*(1), 249–269.

Ozturk, A., S. Smith, E. Gordon-Lipkin, D. Harrison, N. Shiee, D. Pham, B. Caffo, P. Calabresi, and D. Reich (2010). Mri of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis 16*(2), 166–177.

Pettitt, A. N. (1976). A two-sample anderson-darling rank statistic. *Biometrika 63*(1), 161–168.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. Springer.

Read, T. and N. Cressie (1988). *Goodness-of-fit statistics for discrete multivariate data*, Volume 7. Springer-Verlag New York.

Reich, D. S., A. Ozturk, P. A. Calabresi, and S. Mori (2010). Automated vs. conventional tractography in multiple sclerosis: variability and correlation with disability. *NeuroImage 49*(4), 3047–3056.

Schilling, M. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association 81*(395), 799–806.

Scholz, F. (2011). Anderson darling k sample test. R Package adk.

Scholz, F. and M. Stephens (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association 82*(399), 918–924.

Staicu, A.-M., C. M. Crainiceanu, and R. J. Carroll (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics 11*(2), 177–194.

Staicu, A.-M., C. M. Crainiceanu, D. S. Reich, and D. Ruppert (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics 68*(2), 331–343.

Staicu, A.-M., Y. Li, C. M. Crainiceanu, and D. Ruppert (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics to appear*.

Staniswalis, J. G. and J. J. Lee (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association 93*(444), 1403–1418.

Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association 69*(347), pp. 730–737.

Team, R. C. (2015). R: A language and environment for statistical computing. vienna, austria. *URL http://www. R-project. org*.

Wei, L. and J. Lachin (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association 79*(387), 653–661.

Wurtz, D., Y. Chalabi, and L. Luksan (2006). Parameter estimation of arma models with garch/aparch errors an r and splus software implementation. University of Pennsylvania Online Article.

Yao, F., H. Muller, and J. Wang (2005). Functional data analysis for sparse longitudinal data. *JASA 100*, 577–591.

Zhang, J.-T. and J. Chen (2007). Statistical inferences for functional data. *The Annals of Statistics 35*(3), 1052–1079.

Zhang, J.-T., X. Liang, and S. Xiao (2010). On the two-sample behrens-fisher problem for functional data. *Journal of Statistical Theory and Practice 4*(4), 571–587.