Statistical inference on the difference in the means of two correlated functional processes: an application to sleep EEG power spectra

Ciprian M. Crainiceanu[1] Ana-Maria Staicu[2] Shubankar Ray[3] Naresh Punjabi[4]

Abstract

Nonparametric inference methods on the mean difference between two correlated functional processes are proposed. We compare methods that: 1) incorporate different levels of smoothing of the mean and covariance; 2) preserve the sampling design; and 3) use parametric and nonparametric estimation of the mean functions. We provide a comprehensive summary of our findings that could be used as a check list for other applications. We apply our method to estimating the mean difference between average normalized δ-power of sleep electroencephalograms for 51 subjects with severe sleep apnea and 51 matched controls in the first 4 hours after sleep onset. Data are obtained from the Sleep Heart Health Study (SHHS), the largest community cohort study of sleep.

Some key words: Penalized splines, sleep, measurement error.

1 Introduction

We propose nonparametric methods for estimating the mean difference and the associated variability between two correlated functional processes. There is a vast literature on estimating parametric fixed effects with correlated residuals, which led to two separate "schools" of thought in Statistics: estimating equations and covariance modeling. In short, estimating equations have focused on using a working covariance matrix to obtain unbiased estimators of the mean. The empirical covariance matrix is then used in a sandwich formula to obtain corrected covariance estimators; see [1, 2, 3] for more details. Covariance modeling can be done either explicitly using parametric, parametric mixtures or nonparametric methods or implicitly using random effects; see [4, 5, 6, 7, 8] for more details. Nonparametric smoothing with correlated residuals has a similar long history, with most papers being inspired and applied to smoothing of time series data [9, 10, 11, 12, 13].

The literature on functional data analysis also contains some papers on comparing the means of two functional processes. In particular, Benko et al. [14] provide theoretical arguments for using bootstrap tests for assessing the equality of means, eigenfunctions and

[1] Associate Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: ccrainic@jhsph.edu.

[2] Assistant Professor, Department of Statistics, North Carolina State University, 2311 Stinson Drive Campus, Raleigh, NC 27695. E-mail: ana-maria_staicu@ncsu.edu

[3] Research Associate, Biometrics Research, Merck & Company, RY 33-300, Rahway, NJ 07065 USA. E-mail: shubhankar_ray@merck.com

[4] Associate Professor, Department of Epidemiology, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD. E-mail: npunjabi@jhmi.edu

eigenvalues of the covariance functions for the two sample problem. Hall and Van Keilegom [15] use bootstrap for the hypothesis testing of the equality of distributions of two independent samples of curves. Zhang et al. [16] proposed L^2- and bootstrap-based statistics for testing the equality of two mean curves when curves are independent and observed without noise. The main novelty in our paper is that we propose methods for inference on the difference in the means of two functional processes that exhibit *complex correlation patterns*. In Section 1.1 we introduce our motivating example obtained from a matching case-control study, where subjects with severe sleep disrupted breathing were matched to controls. The correlation between cases and controls case is induced by matching. There are many other examples where correlated functional data appear naturally: 1) longitudinal observations of images or functions; 2) replication experiments; and 3) multilevel sampling experiments.

This is a new and important problem, as many new medical and public health studies contain non-independent samples of functions. Our primary aim is to estimate the difference in means between two correlated functional samples together with its associated variability. Our secondary aim is to test whether and where the difference in means is statistically different from zero. We provide three easy to use, fast, and statistically principled techniques that address our primary and secondary aims. These techniques contain variations, adaptations and refinements of ideas sprinkled throughout the literature and are easy to implement, computationally fast, scalable, and adaptable to increasingly complex designs.

Our methods and discussion will be general, but we consider a motivating example from the Sleep Heart Health Study [17], the largest community cohort study of sleep.

1.1 Short description of the data and problem

The SHHS collected in-home polysomnogram (PSG) data on thousands of subjects at multiple visits. Two-channel Electroencephalograph (EEG) data was collected as part of the PSG at a frequency of 125Hz, or 125 observations per second. Thus, for each subject, visit, and EEG channel a total of 3.6 millions observations were collected for a typical 8 hour sleep interval. Here we focus on modeling a particular characteristic of the spectrum of the Electroencephalograph (EEG) data, the proportion of δ-power. For more details on the definition and interpretation of δ-power see, for example, [18, 19, 20]. For our purpose, it is sufficient to know that percent δ-power is a summary measure of the spectral representation of the EEG signal; in this paper we use percent δ-power calculated in 30-second intervals. Figure 1 displays the sleep EEG proportion of δ-power in each of the 30-second intervals of the first 4 hours after sleep onset for 8 matched pairs of subjects. Each panel displays a matched pair with the red lines corresponding to subjects with sleep disrupted breathing (SDB) and the blue lines corresponding to their matched controls. The x axis represents time in hours since sleep onset and the y axis represents the estimated proportion of δ-power. Observations are shown in adjacent 30-second intervals with missing observations indicating wake periods.

A total of 51 matched pairs were obtained using propensity score matching [21]. Subjects with severe SDB were identified as those with a respiratory disturbance index (RDI) greater than 30 events/hour. Subjects without SDB were identified as those with an RDI smaller than 5 events/hour. Other exclusion criteria included prevalent cardiovascular dis-

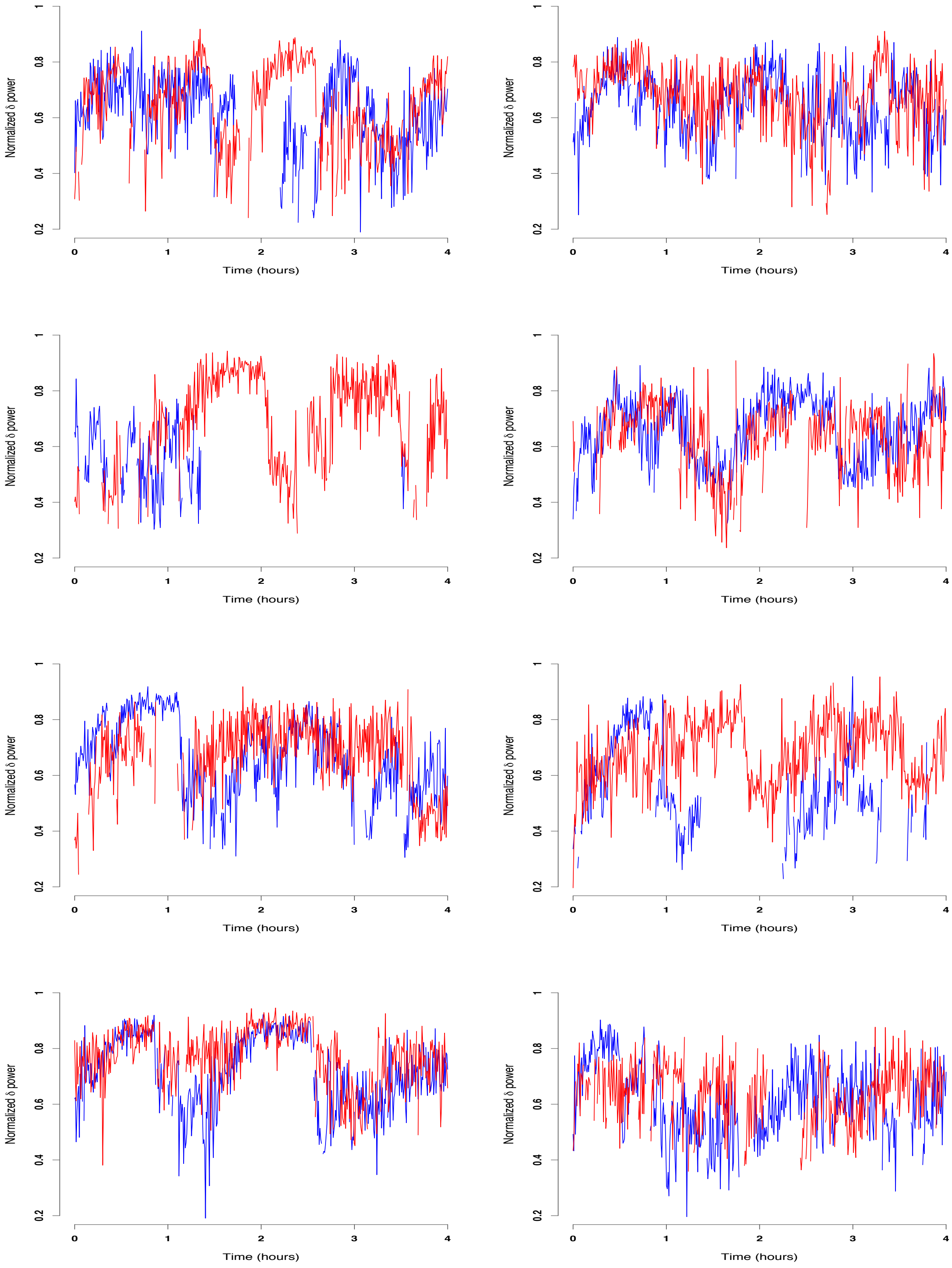

Figure 1: Normalized δ power for the first 4 hours after sleep onset for 8 matched pairs of controls (blue) and sleep apneics (red).

ease, hypertension, chronic obstructive pulmonary disease, asthma, coronary heart disease, history of stroke, and current smoking. Propensity score matching was utilized to balance the groups on demographic factors and to minimize confounding. SDB subjects were matched with no-SDB subjects on the factors of age, BMI, race, and sex. Race and sex were exactly matched, while age and BMI were matched using the nearest neighbor Mahalanobis technique with a caliper of 0.10. The resultant match was 51 pairs that met the strict inclusion criteria outlined above and exhibiting very low standardized biases, a vast improvement on the imbalance of BMI between diseased and non-diseased groups of past studies [22].

Inspection of the 8 pairs displayed in Figure 1 reveals several notable features of the data. First, there is large within/between-subject as well as within-group variability. Second, there are no readily recognizable patterns within groups. Third, and more conspicuous, functions are correlated within groups due to the matching process. Fourth, missing data patterns are subject-specific with the proportion of missing observations varying dramatically across subjects; note, for example, that more than 50% of data are missing for the healthy subject in the third plot. Thus, simply taking the within-group differences would be inefficient by throwing away data.

1.2 Short description of challenges

Data of the type shown in Figure 1 has many of the complexities encountered in similar applications: missing observations, correlated functions, complex dependence structures and noise. The problem we are interested in is estimating the difference in the mean functions corresponding to sleep apneics (red lines in Figure 1) and matched controls (blue lines in Figure 1).

To better understand the problem and the various assumptions it helps to provide a reasonable statistical framework. The data in our study are pairs of functions $\{Y_{iA}(t), Y_{iC}(t)\}$, where i denotes subject, $t = t_1, \ldots, t_T = 480$ denotes the time measured in 30 second intervals from sleep onset, A stands for apneic and C stands for control. For each subject some of the observations might be missing. We write both processes as

$$\left\{ \begin{array}{rcl} Y_{iA}(t) & = & \mu_A(t) + V_{iA}(t); \\ Y_{iC}(t) & = & \mu_C(t) + V_{iC}(t), \end{array} \right. \tag{1}$$

and we are interested in estimators of $d(t) = \mu_A(t) - \mu_C(t)$ and their associated variability. All functional data papers concerned with estimating $d(\cdot)$ [14, 15, 16] assume that $V_{iA}(\cdot)$ and $V_{iC}(\cdot)$ are independent, an unreasonable assumption in our and other contexts. Thus, the main challenge is to estimate the function $d(\cdot)$ when the residual processes $V_{iA}(\cdot)$ and $V_{iC}(\cdot)$ have complex covariance structures and are correlated. In most cases, assuming a parametric covariance function, such as working independence, autoregressive or exchangeable, would badly misfit the observed functional covariance. Taking mixtures of such families tends to fail equally badly due to the complex nature of functional data. A secondary challenge is that making a-priori parametric assumptions about either the mean functions or the difference between them would likely be misleading.

To address these issues we propose three strategies. The first strategy uses nonparametric estimators of the mean functions based on penalized splines [23, 24] under the independence assumption. The variability of the difference function estimate is then obtained via a nonparametric bootstrap of pairs. We call this the "Nonparametric estimation using nonparametric bootstrap" and we describe it in details in Section 2.1. The second strategy uses the same nonparametric estimators of the mean functions. The procedure then relies on modeling and smoothing the error processes, $V_{iA}(\cdot)$ and $V_{iC}(\cdot)$, using multilevel functional techniques [20]. Thus, instead of a nonparametric bootstrap of pairs we simulate data from the joint distribution of the error processes. We call this the "Nonparametric estimation using parametric bootstrap" because it uses parametric simulations from the functional distributions. The method is described in Section 2.2. The third strategy uses parametric estimation of the mean functions where the number of degrees of freedom is fixed a-priori. Nonparametric bootstrap of pairs is then used to estimate estimators variability. The method is described in Section 2.3.

2 Functional bootstrap

Because subjects are matched it is reasonable to assume that the processes $V_{iA}(t)$ and $V_{iC}(t)$ are correlated. In this section we propose two bootstrap methods that preserve the pair-specific correlation. The first approach employs a fully nonparametric bootstrap whereas the second combines elements of nonparametric modeling of covariance operators and parametric simulations from the induced mixed effects model.

Both methods use estimators of the mean function under the independence assumption. We start by describing two smooth estimators of $\mu_A(t)$; the estimator for $\mu_C(t)$ is obtained similarly. The first estimator, denoted $\widetilde{\mu}_A(t)$, is obtained by using penalized spline smoothing of all pairs $\{t, Y_{iA}(t)\}$ under the independence assumption, that is assuming that $V_{iA}(t)$ is a mean zero, uncorrelated, homoscedastic process. The second estimator, denoted by $\widehat{\mu}_A(t)$, is obtained by using penalized spline smoothing of $\{t, \overline{Y}_{\cdot A}(t)\}$, where $\overline{Y}_{\cdot A}(t) = \sum_{i=1}^{I} Y_{iA}(t)/I$ for all t. Penalized splines are one of the most successful and practical automatic smoothing techniques; we refer here to the excellent monographs [24, 25]. A penalized spline approach represents the mean function as $\mu_A(t) = \boldsymbol{B}_A(t)\boldsymbol{\beta}_A$, where $\boldsymbol{B}_A(t)$ is a low-rank spline basis obtained by fixing the number and location of knots and achieves smoothing by imposing a normal prior on the spline coefficients $\boldsymbol{\beta}_A \sim N(0, \boldsymbol{D}_A)$. The penalty matrix $\boldsymbol{D}_A$ is intrinsically related to the choice of spline basis, $\boldsymbol{B}_A(t)$, and typically depends on one smoothing parameter that is estimated from the data. In this paper we use thin-plate splines with 20 knots positioned at the empirical quantiles of the observed time points ([24], Chapter 13.4). We used the function `spm` implemented in the implemented in `R` [26] package `SemiPar` [27] .

There are some important points to make before we proceed. First, note that $\overline{Y}_{\cdot A}(t)$ is a consistent estimator of $\mu_A(t)$. Second, obtaining $\widetilde{\mu}_A(t)$ is more computationally expensive than obtaining $\widehat{\mu}_A(t)$ as it requires smoothing of IT pairs compared to only T pairs. This is especially important when the number of subjects, I, is large and one uses the nonparametric bootstrap to estimate the variability of mean estimators. However, both estimators can be used in most applications. We will show that they provide almost identical results in our application and in simulations. The intuition for this result is quite simple: the mean of local means is the local mean.

2.1 Nonparametric estimation using nonparametric bootstrap

Draw from the pool of people again? Get estimates for d() curve in each bootstrapped sample

We applied the bootstrap methods to the 51 matched pairs of controls and sleep apneics. The top-left panel in Figure 2 displays the average normalized δ-power for the 51 subjects with severe sleep apnea (red) and 51 matched controls (blue). Raw means are depicted as dots, while penalized spline smoothers of the raw means are depicted as lines. Similar curves could be shown using a penalized spline smoother of the entire data set, but the results are indistinguishable from the ones shown. We used $B = 1000$ nonparametric bootstrap samples of matched pairs and we repeated the penalized spline fitting of the raw means described above; the total computation time was 27 minutes (Dual Core Processor 3GHz, 32Gb RAM PC). This created B bootstrap estimators $\widehat{d}_b(t) = \widehat{\mu}_{A,b}(t) - \widehat{\mu}_{C,b}(t)$ of $d(t)$, $b = 1, \ldots, B$. The top-right panel in Figure 2 displays the bootstrap estimator of mean diffeernces $\widehat{d}_B(t) = \sum_{b=1}^{B} \widehat{d}_b(t)/B$ as a solid black line. The estimator $\widehat{d}_B(t)$ tends to be negative during most of the 4-hour interval, which suggests that subjects with severe sleep apnea tend to have lower normalized δ power. However, $\widehat{d}_B(t)$ is far from being a flat line indicating that the difference is more pronounced during certain intervals.

To better visualize where these differences are likely to occur we construct 95% pointwise confidence intervals based on the bootstrap samples. At time point t a 95% bootstrap confidence interval for $d_B(t)$ is $[\widehat{q}_{B,0.025}(t), \widehat{q}_{B,0.975}(t)]$, where $\widehat{q}_{B,p}(t)$ is the p-quantile of the bootstrap sample $\widehat{d}_b(t)$, $b = 1, \ldots, B$. Because the distribution is symmetric we chose instead to use $\widehat{d}_B(t) \pm 2\widehat{s}_B(t)$, where $\widehat{s}_B(t)$ is the estimated standard deviation of the bootstrap sample $\widehat{d}_b(t)$, $b = 1, \ldots, B$. Here we focus on pointwise confidence intervals, but joint confidence intervals will be discussed in Section 3. The top-right panel in Figure 2 also displays the 95% pointwise confidence intervals as a shaded gray area. Statistically significant differences between normalized δ power of sleep apneics and controls can be detected between minutes 4 and 23 with the largest difference around minute 18 and between minutes 113 and 129 with the largest difference around minute 122. These findings seem to agree with the observed variability in the top-left panel of Figure 2.

We have conducted the same analysis using $\widetilde{d}(t)$ instead of $\widehat{d}(t)$. Recall that $\widetilde{d}(t)$ is based on nonparametric smooth estimates of the entire data set, whereas $\widehat{d}(t)$ is based on nonparametric smooth estimates of the empirical means. The bottom-left panel in Figure 2 displays the same information for $\widetilde{d}(t)$ as the top-right panel in the same figure. This indicates that results are practically indistinguishable with the two methods. We prefer using $\widehat{d}(t)$ because it is much faster to calculate.

The raw difference between normalized δ power displayed in the top-right panel in Figure 2 provides valuable information, but does not directly quantify the relative size of observed differences. To provide that, the bottom-right panel in Figure 2 displays the difference between normalized δ power between the two groups as a percentage of the range of the estidelta power is the outcome of interest

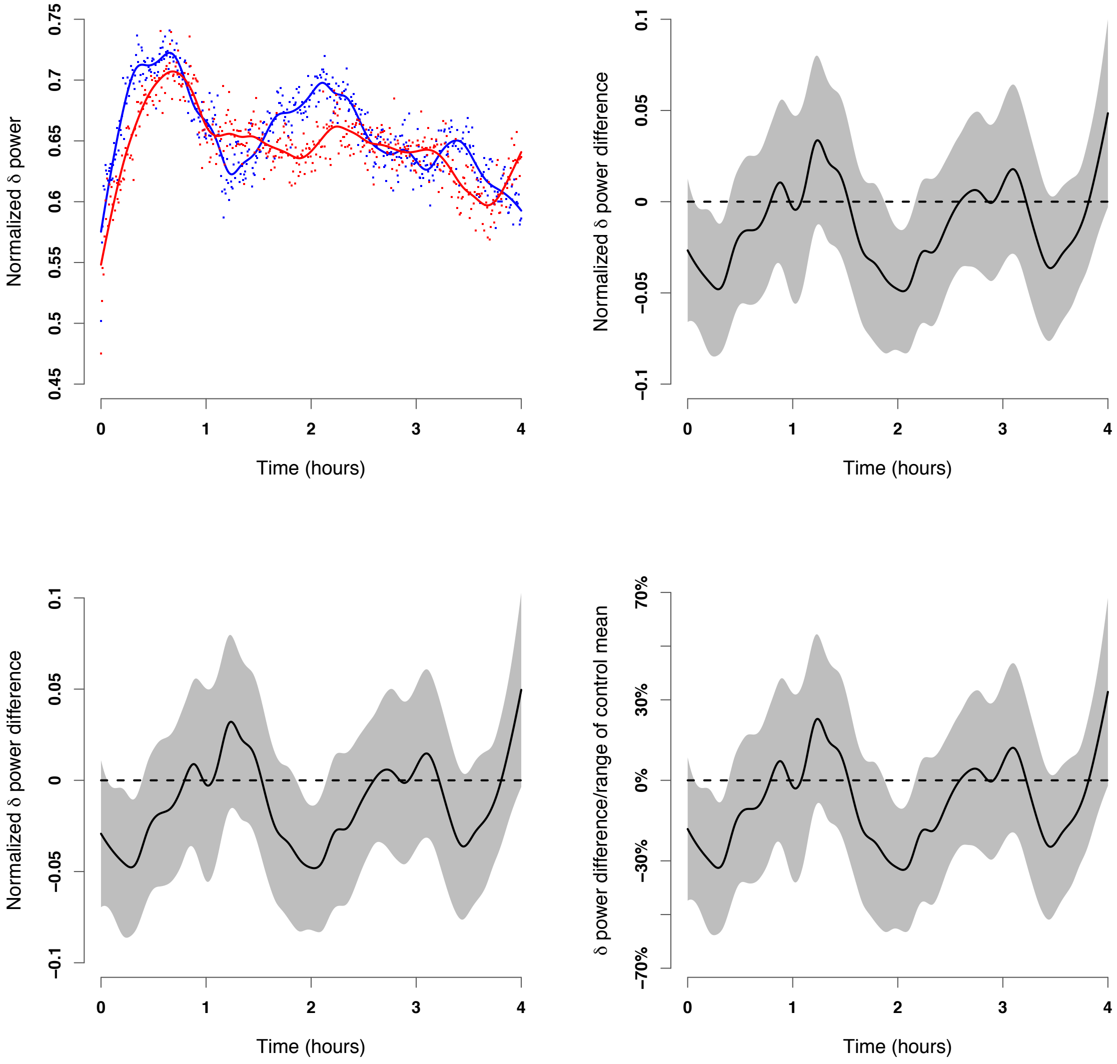

Figure 2: Top-left panel: average normalized δ-power for 51 subjects with severe sleep apnea (red) and 51 matched controls (blue). Raw means are depicted as dots, while penalized spline smoothers of the raw means are depicted as lines. Top-right panel: estimated mean difference between average normalized δ-power apneics and controls. The pointwise 95% confidence intervals (shaded gray area) are obtained by nonparametric bootstrapping of pairs of subjects, estimating the mean of the sleep-apneic and control groups using penalized splines, and taking the difference between the estimated means of the groups. Bottom-left panel: similar to the top-right panel, obtained using nonparametric bootstrap of pairs but with a nonparametric smooth estimates of the entire data set for each group. Bottom-right panel: shows the results in the top-right panel as a percent of the range of the mean normalized δ power of controls.

mated mean functions of controls. More precisely, we plot $100\widehat{d}(t)/\{\max_t \widehat{\mu}_C(t) - \min_t \widehat{\mu}_C(t)\}$ and its associated variability.

An alternative approach would be to bootstrap the pair differences $Y_{iA}(t) - Y_{iC}(t) = d(t) + V_{iA}(t) - V_{iC}(t)$. However, calculating the difference $Y_{iA}(t) - Y_{iC}(t)$ can only be done when both $Y_{iA}(t)$ and $Y_{iC}(t)$ are observed. Thus, missing data in either process would compound the problem and would lead to serious efficiency losses. When data are not missing, this approach provides similar results to the ones obtained with the method described above.

2.2 Nonparametric estimation using parametric bootstrap

In this section we will continue to use the smooth estimates of the mean functions under the independence assumption. The main difference is in how we estimate the variability of these estimators when the distribution of error processes $V_{iA}(t)$ and $V_{iC}(t)$ is unknown. We start by noting that the pairing of subjects induces within-pair correlation. We account for this by defining a multilevel functional model for both processes as described by [20]

$$\left\{ \begin{array}{lcl} V_{iA}(t) & = & X_i(t) + U_{iA}(t) + \epsilon_{iA}(t); \\ V_{iC}(t) & = & X_i(t) + U_{iC}(t) + \epsilon_{iC}(t), \end{array} \right. \tag{2}$$

where $X_i(t)$ is a functional process with smooth covariance operator $K^X(\cdot,\cdot)$, $U_{iA}(t)$ and $U_{iC}(t)$ are functional processes with the same smooth covariance operator $K^U(\cdot,\cdot)$, $\epsilon_{iA}(t)$ and $\epsilon_{iC}(t)$ are independent mean zero variance σ^2_ϵ random variables, and $X_i(t)$, $U_{iA}(t)$, $U_{iC}(t)$, $\epsilon_{iA}(t)$, and $\epsilon_{iC}(t)$ are assumed mutually independent within- and between-pairs. The role of the process $X_i(t)$ is to account for the within-pair correlation, as $\text{cov}\{Y_{iA}(t), Y_{iC}(s)\} = K^X(t,s)$. The processes $U_{iA}(t)$ and $U_{iA}(t)$ are assumed to share the same covariance operator, a reasonable assumption in this context. However, this assumption is not necessary and may be relaxed in other applications. Both $K^X(\cdot,\cdot)$ and $K^U(\cdot,\cdot)$ are left unspecified, are assumed to be smooth, and are estimated from the data.

Here we proceed in two stages. First, we obtain $W_{iA}(t) = Y_{iA}(t) - \widehat{\mu}_A(t) = V_{iA}(t) + \{\mu_A(t) - \widehat{\mu}_A(t)\}$ and $W_{iC}(t) = Y_{iC}(t) - \widehat{\mu}_C(t) = V_{iC}(t) + \{\mu_C(t) - \widehat{\mu}_C(t)\}$, where $\widehat{\mu}_A(t)$ and $\widehat{\mu}_C(t)$ are the nonparametric smooth estimates of $\mu_A(t)$ and $\mu_C(t)$ described in Section 2.1. Note that the covariance operators of the $W(\cdot)$ and $V(\cdot)$ are identical and $\sup_i |W_{iA}(t) - V_{iA}(t)| \leq |\mu_A(t) - \widehat{\mu}_A(t)|$. Thus, we can use the observed $W(\cdot)$ process to estimate the covariance operators of $V(\cdot)$.

Second, we use Multilevel Functional Principal Component Analysis (MFPCA) [20] to obtain the parsimonious bases that capture most of the functional variability of the space spanned by $X_i(t)$ and $U_{ij}(t)$ $j = A, C$, respectively. MFPCA is based on the spectral decomposition of the within- and between-visit functional variability covariance operators. We summarize here the main components of this methodology. Denote by $K^W_T(s,t) = \text{cov}\{\, W_{ij}(s), W_{ij}(t)\,\}$ and $K^W_B(s,t) = \text{cov}\{\, W_{ij}(s), W_{ik}(t)\,\}$ for $j \neq k$ the total and between covariance operator corresponding to the observed process, $W_{ij}(\cdot)$, respectively. Denote by $K^X(t,s) = \text{cov}\{X_i(t), X_i(s)\}$ the covariance operator of the $X_i(\cdot)$ process and by $K^U(t,s) = \text{cov}\{\, U_{ij}(s), U_{ij}(t)\,\}$ the covariance operator of the $U_{ij}(\cdot)$ process. By definition, $K^U_B(s,t) =$

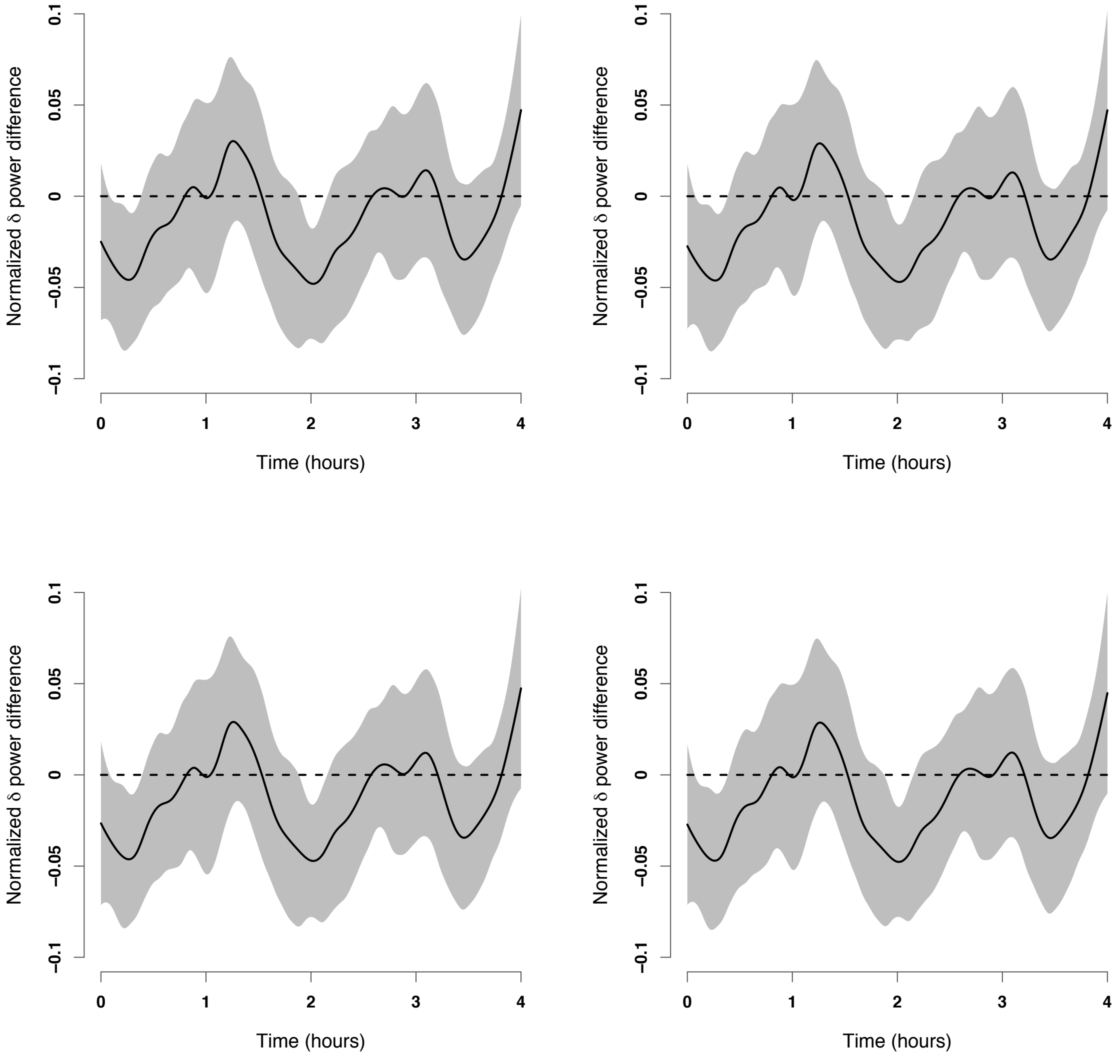

Figure 3: Mean difference and 95% pointwise confidence intervals via the quasi nonparametric functional bootstrap. Top-left panel: results for K and L chosen such that 90% of the variability described by $K^X(\cdot,\cdot)$ and $K^U(\cdot,\cdot)$ is explained. We also used the standard estimator of the variability, σ_ϵ. Top-right panel: explained variability is increased to 95% with the same estimator of σ_ϵ. Left-bottom panel: explained variability is 95% with an estimator that is roughly 20% larger than the standard estimator. Right-bottom panel: explained variability is 99% at both levels with the same conservative estimator of σ_ϵ.

$\text{cov}\{U_{ij}(s), U_{ik}(t)\} = 0$ for $j \neq k$. Moreover, $K_B^W(s,t) = K^X(s,t)$ and $K_T^W(s,t) = K^X(s,t) + K^U(s,t) + \sigma_\epsilon^2 \delta_{ts}$, where δ_{ts} is equal to 1 when $t = s$ and 0 otherwise. Thus, $K^X(s,t)$ can be estimated using a method of moments estimator of $K_B^W(s,t)$, say $\widehat{K}_B^W(s,t)$. For $t \neq s$ a method of moments estimator of $K_T^W(s,t) - K_B^W(s,t)$, say $\widehat{K}^U(s,t)$, can be used to estimate $K^U(s,t)$. To estimate $\widehat{K}^U(t,t)$ one predicts $K^U(t,t)$ using a bivariate thin-plate spline smoother of $\widehat{K}^U(s,t)$ for $s \neq t$. This method was proposed by Staniswalis and Lee [28] for nonparametric longitudinal data analysis and was shown to work well for MFPCA [18, 20, 29].

Once consistent estimators of $K^X(s,t)$ and $K^U(s,t)$ are available, the spectral decomposition and functional regression proceed as in the single-level case. More precisely, Mercer's theorem (see [30], Chapter 4) provides the following convenient spectral decompositions $K^X(t,s) = \sum_{k=1}^\infty \lambda_k^{(1)} \psi_k^{(1)}(t) \psi_k^{(1)}(s)$, where $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \ldots$ are the ordered eigenvalues and $\psi_k^{(1)}(\cdot)$ are the associated orthonormal eigenfunctions of $K^X(\cdot,\cdot)$ in the L^2 norm. Similarly, $K^U(t,s) = \sum_{l=1}^\infty \lambda_l^{(2)} \psi_l^{(2)}(t) \psi_l^{(2)}(s)$, where $\lambda_1^{(2)} \geq \lambda_2^{(2)} \geq \ldots$ are the ordered eigenvalues and $\psi_l^{(2)}(\cdot)$ are the associated orthonormal eigenfunctions of $K^U(\cdot,\cdot)$ in the L^2 norm. The Karhunen-Loève (KL) decomposition [31, 32] provides the following infinite decompositions $X_i(t) = \sum_{k=1}^\infty \xi_{ik} \psi_k^{(1)}(t)$ and $U_{ij}(t) = \sum_{l=1}^\infty \zeta_{ijl} \psi_l^{(2)}(t)$ where $\xi_{ik} = \int_0^1 X_i(t) \psi_k^{(1)}(t) dt$, $\zeta_{ijl} = \int_0^1 U_{ij}(t) \psi_l^{(2)}(t) dt$ are the principal component scores with $E(\xi_{ik}) = E(\zeta_{ijl}) = 0$, $\text{Var}(\xi_{ik}) = \lambda_k^{(1)}$, $\text{Var}(\zeta_{ijl}) = \lambda_l^{(2)}$. The zero-correlation assumption between the $X_i(\cdot)$ and $U_{ij}(\cdot)$ processes is ensured by the assumption that $\text{cov}(\xi_i, \zeta_{ijl}) = 0$. These properties hold for every i, j, k, and l. For simplicity we will refer to $\psi_k^{(1)}(\cdot)$, $\psi_l^{(2)}(\cdot)$ and $\lambda_k^{(1)}$, $\lambda_l^{(2)}$ as the level 1 and 2 eigenfunctions and eigenvalues, respectively.

Given these developments, we propose to parametrically simulate functional residuals from the model

$$\left\{ \begin{array}{lcl} V_{iA}(t) & = & \sum_{k=1}^K \xi_{ik} \psi_k^{(1)}(t) + \sum_{l=1}^L \zeta_{iAl} \psi_l^{(2)}(t) + \epsilon_{iA}(t); \\ V_{iC}(t) & = & \sum_{k=1}^K \xi_{ik} \psi_k^{(1)}(t) + \sum_{l=1}^L \zeta_{iCl} \psi_l^{(2)}(t) + \epsilon_{iC}(t), \end{array} \right. \tag{3}$$

where $\xi_{ik} \sim N\{0, \lambda_k^{(1)}\}$, $k = 1, \ldots, K$, $\zeta_{iAl}, \zeta_{iCl} \sim N\{0, \lambda_l^{(2)}\}$, $l = 1, \ldots, L$, and $\epsilon_{iA}(t), \epsilon_{iC}(t) \sim N(0, \sigma_\epsilon^2)$ are mutually independent. A quasi-nonparametric bootstrap is then obtained by calculating $Y_{iA}^{(b)}(t) = \widehat{\mu}_A(t) + V_{iA}^{(b)}(t)$ and $Y_{iC}^{(b)}(t) = \widehat{\mu}_C(t) + V_{iC}^{(b)}(t)$, where $V_{iA}^{(b)}(t)$ and $V_{iC}^{(b)}(t)$ are obtained by simulation from model (3) and $b = 1, \ldots, B$. This could replace the nonparametric bootstrap described in Section 2.1; the methods for obtaining the variability of estimators remain the same.

Simulating from model 3 is easy once MFPCA is used to estimate the eigenfunctions and eigenvalues. The only technical point is deciding what values of K and L to use in practice. In general the particular choice does not influence the confidence intervals provided that K and L are large enough. To show that this is, indeed, the case in our application we considered 4 different choices from reasonable to extreme. Figure 3 displays the 95% confidence intervals obtained using 4 different choices. The top-left panel displays results for K and L chosen such that 90% of the variability described by $K^X(\cdot,\cdot)$ and $K^U(\cdot,\cdot)$ is explained. We also used the standard estimator of the variability, σ_ϵ. The top-right panel displays results for the case when the explained variability is increased to 99% with the same estimator of σ_ϵ. The left-bottom panel displays results for the case when the explained variability is 99% with a conservative estimator of σ_ϵ, that is an estimator that is roughly 20% than the standard estimator. The right bottom panel shows results for the case when the explained variability is 99.95% at both levels with the same conservative estimator of σ_ϵ.

We conclude that the choices of K, L and estimator of σ_ϵ have a minimal impact on the inference about the mean function. Also, the confidence intervals tend to be shorter than in the case of the fully nonparametric bootstrap; compare Figures 2 and 3. The fact that the confidence intervals are shorter does not automatically make them better. It could be that the longer fully nonparametric bootstrap intervals are necessary to achieve the nominal coverage level. We investigate this further in Section 4.

2.3 Parametric estimation using nonparametric bootstrap

Both previous methods rely on nonparametric smoothing of the group-specific mean functions. A simple alternative is to parameterize the mean functions, estimate the means under the independence assumption, and use the nonparametric bootstrap of pairs described in Section 2.1 to estimate variability. Such a method is especially useful in the case when prior information about the shape of the mean functions exists before data analysis is conducted.

The top-left panel in Figure 4 displays the estimator of the mean difference based on the assumption that both means have 1 degree of freedom per hour (or 4 degrees of freedom for the 4 hour period). The other 3 panels display the estimators based on the assumptions that both means have 2, 3, and 4 degrees of freedom per hour, which correspond to 8, 12, and 16 total degrees of freedom, respectively. As the number of degrees of freedom increases, the function becomes wigglier while its variability increases. Indeed, the average pointwise standard deviation increases from 0.0124 for the 1 degree of freedom per hour fit to 0.0204, or 64.5%, for the 4 degrees of freedom per hour fit; for more details see Section 3.

The cases shown here are between two extremes. At one extreme is the model with 1/4 degrees of freedoms per hour, which would correspond to fitting a constant mean both to cases and controls. At the other extreme is the model with 120 degrees of freedoms per hour, which would fit a different mean to every time point. Both these cases are important in themselves and their results are provided in Section 3.

3 Pointwise and joint confidence intervals

We first clarify several facts about single and multiple testing. While these facts are well known in Statistics, there is still a level of confusion among Statisticians, which is highly amplified in the scientific literature. First, multiple testing is a more ambitious goal than single testing. The penalty for being ambitious is, typically, longer confidence intervals. The penalty increases with the number of tests one performs and decreases with the amount of correlation between tests. For example, performing one million tests with perfectly correlated

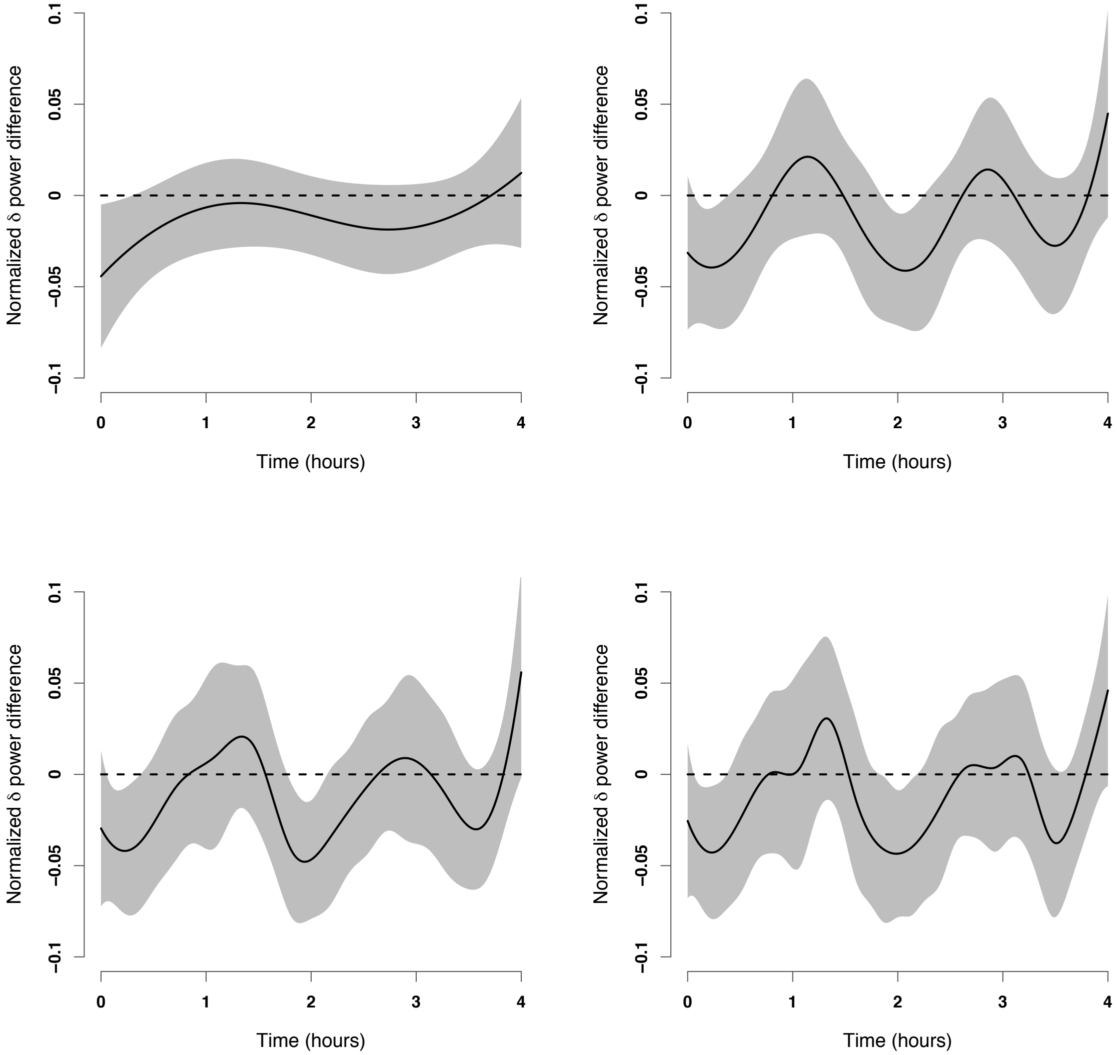

Figure 4: Mean difference estimators between apneics and controls using thin-plate regression splines with un-penalized coefficients. The number of degrees of freedom is fixed (not estimated) and is set equal to: 1 degree of freedom per hour (top-left panel); 2 degrees of freedom per hour (top-right panel); 3 degrees of freedom per hour (bottom-left panel); and 4 degrees of freedom per hour (bottom-right panel).

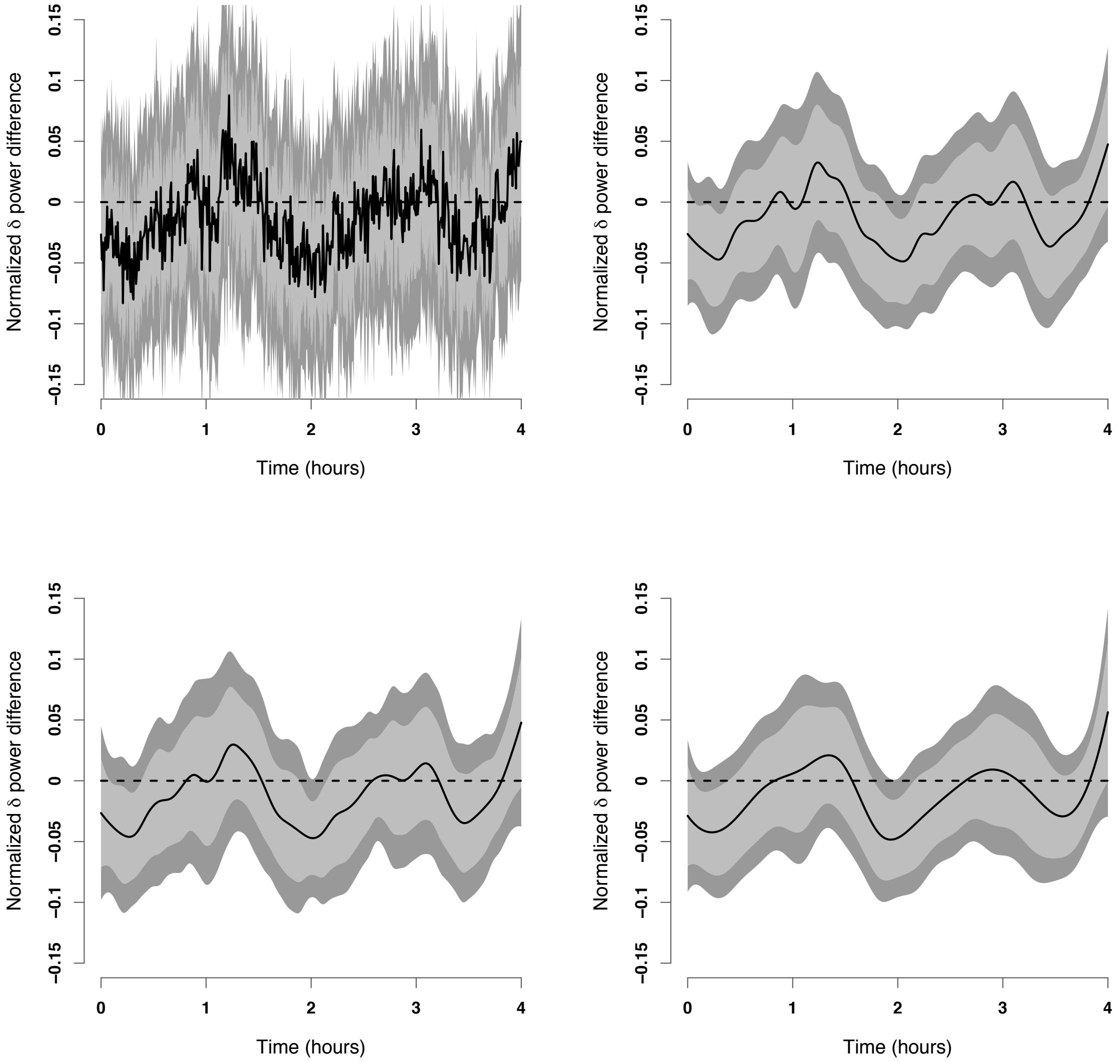

Figure 5: Pointwise (light gray) and joint (dark extensions) 95% confidence intervals for the difference in mean of normalized δ-power between sleep apneics and matched controls as a function of time from sleep onset. Top-left: t-testing without smoothing the mean functions. Top-right: nonparametric estimation using nonparametric bootstrap (NE/NB). Bottom-left: nonparametric estimation using parametric bootstrap (NE/PB). Bottom-right: parametric estimation with 3 degrees of freedom per hour using nonparametric bootstrap (PE/NB).

test statistics does not require any correction. Performing the same tests with independent test statistics will require a Bonferonni correction. Second, the Bonferonni correction for multiple testing ensures that the level of the multiple testing procedure does not exceed the specified α level irrespective of correlation and number of tests. In many cases the Bonferonni correction may be too aggressive. Depending on the problem and mood, the Bonferonni correction could be characterized either as "lazy" or "robust against false positives". Third, the false discovery rate (FDR) approach [33] is designed for *independent data* and, naturally, controls the FDR and not the α level of the test. A standard application of FDR to multiple testing can be viewed as "making α larger to allow for some positive findings" and is agnostic to correlation. Fourth, in many applications even when data are correlated the test statistics tend to have low or no correlation, which requires a Bonferonni correction.

With these facts in mind we now proceed with our application. Single testing for differences in means of two processes can be stated as

$$H_{0,t} : \mu_A(t) = \mu_C(t) \quad \text{versus} \quad H_{A,t} : \mu_A(t) \neq \mu_C(t) \text{ for a fixed } t.$$

When testing for one t then there are standard methods to preserve the α level of the test; for example, using normal or bootstrap approximations of the null distribution. These are called pointwise confidence intervals and are depicted as light gray bands in figures throughout this paper. Multiple testing for differences at all locations can be stated as

$$H_{0,M} : \mu_A(t) = \mu_C(t) \text{ for all } t \quad \text{versus} \quad H_{A,M} : \mu_A(t) \neq \mu_C(t) \text{ for at least one } t.$$

All three methods described in this paper produce samples from the joint distributions of an estimator of the difference function $d(t) = \mu_A(t) - \mu_C(t)$. These are either bootstrap or posterior samples from the distribution of $d(t)$ given the data.

Given these samples, there is a simple way to produce joint confidence intervals. Assume, that we have a $T \times B$ dimensional matrix S that stores the samples from the target distribution. Each row contains one sample of length T and corresponds to a particular estimator of $d(\cdot)$. The column mean, $\bar{d}(t)$, over all samples is an estimator of the mean function, whereas the covariance $\Sigma_S = \text{cov}(S)$ is a $T \times T$ dimensional matrix. With enough samples the sampling variability in $\bar{d}(t)$ and Σ_S can be ignored. To obtain the joint confidence intervals we use the following easy to implement algorithm:

T columns
B rows
(one row / target or bootstrapped sample)

1. Simulate $d_n(t)$ from a multivariate $N\{\bar{d}(t), \Sigma_S\}$

   Why not use the d_b's directly? Why re-generate them as normal?

   n = 1,...N
2. Calculate $x_n = \max_t\{|d_n(t) - \bar{d}(t)|/\sigma(t)\}$, where $\sigma^2(t)$ is the tth diagonal element of Σ_S

   What's the sup deviations the whole curve takes from the average, in terms of standard deviations??
3. Repeat for $n = 1, \ldots, N$ and obtain $q_{1-\alpha}$ the $1 - \alpha$ empirical quantile of the sample $\{x_n : n = 1, \ldots, N\}$

   N refers to # simulated
4. Obtain the joint confidence intervals $\bar{d}(t) \pm q_{1-\alpha}\sigma(t)$

   Multiply again by sd to put it in units of sd

   You find the quantile for how far betas shift from the mean at their max point

This was the algorithm used to obtain the joint confidence intervals in Figure 5; note the dark gray bands extending the light gray areas. The interpretation for the pointwise confidence

The idea is that if it was sig, you would expect it to differ at any SPECIFIC point to be more than the 95Q difference at that point

The pointwise CI is for

intervals shown in light gray is that, *at each location* in repeated samples, the true mean function will be covered by the shaded light gray interval $100(1-\alpha)\%$ of the time. The interpretation for the joint confidence intervals is that *at all locations* in repeated samples, the true mean function will be covered by the shown as dark or light gray areas $100(1-\alpha)\%$ of the time. One could think of the dark gray band extensions as the correction for multiple comparisons that takes into account the observed correlation between test statistics.

Figure 5 displays the estimated mean difference together with the 95% pointwise (light gray) and joint (dark gray extensions) under various estimation scenarios. The top-left panel corresponds to the case when no smoothing of the mean function is used. While the method is wasteful and results are extreme, this is the most popular approach and has been used extensively in genomics, where it is called "point-wise testing", and imaging, where it is termed "voxel-wise testing". The top-right panel displays the results for "Nonparametric estimation using nonparametric bootstrap" (NE/NB), the bottom-left panel displays the results for "Nonparametric estimation using parametric bootstrap" (NE/PB), and the bottom-right panel displays the results for "Parametric estimation using nonparametric bootstrap" (PE/NB) with 3 degrees of freedom per hour, or 12 degrees of freedom total.

The four plots display some obvious differences, but they convey the same general message: there is no statistically significant difference between the normalized δ-power in the first 4 hours after sleep onset between apneics and controls in this data set. Moreover, if a difference exists then it is more pronounced around minutes 20 and 120 after sleep onset. The distinction between pointwise and joint confidence intervals as well as between pointwise and joint tests for mean differences is not just academic. Indeed, ignoring this distinction would lead to fundamentally different conclusions from analyzing the same data set. A quick inspection of Figure 5 reveals that using pointwise confidence intervals would lead to the conclusion that there is a statistically significant difference between sleep apneics and controls.

METHOD:
Use bands to see if
the whole thing
is sig.

Use t-map for next
trial to see if it's sig
at those points

These findings should not be disappointing. Indeed, this study could be used to generate simple, plausible and easy to test hypotheses that could be analyzed in other studies. We hypothesize that there is a difference between average normalized δ power of sleep apneics and controls and this difference is localized between minutes 5 and 20 and between minutes 110 and 125. In fact, if this information were available a standard t-test for difference in the $[5, 20]$ minute period would have a p-value of 0.0190, whereas in the $[110, 125]$ minute would have a p-value of 0.0037 using a two-sided t-test. These tests are invalid after conducting 480 other tests to identify regions of large differences. However, they provide extremely interesting findings that could become more focused hypotheses for future studies. Note that a two sided t-test for the difference in the mean over all time points and subjects between apneics and controls has a p-value of 0.2, indicating that there is not enough statistical evidence to reject the null of no difference. We also intend to use the average δ-power between minutes $[5, 20]$ and $[110, 125]$ as potential health biomarkers in our future studies.

We now quantify more precisely the observed differences between the three procedures we applied to the SHHS data set. In particular, we report the average estimated standard deviation across all time periods, $\bar{\sigma} = \sum_{t=1}^{T} \hat{\sigma}_t / T$, where $\hat{\sigma}_t$ is an estimator of the variability of the estimator $\hat{d}(t)$ for a particular method. Visually, $\bar{\sigma}$ is a measure of width of the pointwise confidence intervals depicted as light gray areas. We also report the 0.95 quantile, $q_{0.95}$, of the distribution used for multiple test corrections. This quantile will depend on the method of estimation of the covariance of test statistics. Note that the average length of the joint confidence intervals shown as dark gray extensions is $2q_{0.95}\bar{\sigma}$.

Table 1 displays results for the three methods discussed in this paper and compares them with the results obtained using pointwise t-testing without smoothing the mean function. This is considered to be the reference procedure and is labeled "PE/NB (120 df/h)" because it is equivalent to using a parametric estimation with one degree of freedom for every time point. Rows two and three, labeled NE/NB, NE/PB, are the methods introduced in Sections 2.1 and 2.2, respectively. The last four rows correspond to the PE/NB method described in Section 2.3 using from 4 to 1 degrees of freedom per hour.

For sleep data it's pretty noisy

The results now quantify our findings. In particular, they indicate that not smoothing the mean is wasteful. Indeed, nonparametrically smoothing the mean across time reduces the average estimated standard deviation of the mean from 0.0286 to 0.0204, or roughly 30%. Moreover, the quantile used for multiple corrections, also decreases from 3.79 to 3.15, or roughly 17%. This decrease is likely due to the increased correlation after smoothing. For reference, the Bonferonni correction quantile for 480 two-sided tests with a family wide error rate (FWER) of 0.05 is 3.88. Thus, the average length of the joint 95% confidence intervals decreased from 0.108 to 0.064, which is a $100(0.108-0.064)/0.108 = 41\%$ reduction. Parametric smoothing further reduces the average length of the joint confidence intervals. Indeed, average length is increasing as a function of degrees of freedom from 0.032, for 4 degrees of freedom per hour, to 0.062, for 4 degrees of freedom per hour. However, one should not conclude that a smaller number of degrees of freedom is better, as the estimator of the mean function is shrunk towards zero; see Figure 3 for more details.

We conclude that the reduction in average length of confidence intervals can be quite dramatic using very simple smoothing methods. In practice, if little information is available before conducting the analysis it makes sense to use nonparametric smoothing of the mean. However, if some information is available then it may make sense to use that information to commit to a particular smoothing method. For example, in future studies of differences in normalized sleep δ-power one can start by assuming that the functions have 3 degrees of freedom per hour. If in doubt, it is probably better to allow for more rather than less degrees of freedom.

4 Simulations

Here we investigate the performance of the observed methods in a simulation study. For all settings our results are based on simulation of 200 data sets from model (1), where $\mu_d(t)$ is detailed below and $V_{id}(t) = X_i(t) + U_{id}(t) + \epsilon_{id}(t)$, for $i = 1, \ldots, I$ and $d = A, C$ indicating whether the ith curve is from the case (A) or control (C) group. We set $\sigma_\epsilon^2 = 0.10$ and consider curves sampled at $T = 100$ points. We consider many scenarios that combine various choices:

Method	$\bar{\sigma}$	$q_{0.95}$	$q_{0.95}\bar{\sigma}$	% reduction
PE/NB(120df/h)	0.0286	3.79	0.108	reference
NE/NB	0.0204	3.15	0.064	41%
NE/PB	0.0204	3.17	0.064	41%
PE/NB(4df/h)	0.0204	3.06	0.062	43%
PE/NB(3df/h)	0.0191	3.05	0.058	46%
PE/NB(2df/h)	0.0177	2.96	0.052	51%
PE/NB(1df/h)	0.0124	2.59	0.032	70%

Table 1: Average estimated standard deviation, $\bar{\sigma} = \sum_{t=1}^{T} \hat{\sigma}_t$, of the mean estimator, $\widehat{d}(t)$, 95% quantile used to correct for multiple testing, $q_{0.95}$, average length of joint confidence intervals, $2q_{0.95}\bar{\sigma}$, and percent reduction in average length of confidence intervals compared to t-testing without smoothing of the mean function, labeled PE/NB (120df/h). For example, the percent reduction in average length was calculated for method NE/NB as $100(0.108 - 0.064)/0.108 = 41\%$. The label NE/NB is short for nonparametric estimation using nonparametric bootstrap described in Section 2.1. The label NE/PB is short for nonparametric estimation using parametric bootstrap as described in Section 2.2. The label PE/NB is short for parametric estimation using nonparametric bootstrap as described in Section 2.3. After the label PE/NB the number of degrees of freedom per hour for parametric estimation is provided within brackets.

1. Number of subjects: (a) $I = 30$, (b) $I = 50$, (c) $I = 100$, (d) $I = 200$;

2. Sample design: (a) equally spaced time points in $[0, 1]$, (b) unequally spaced time points in $[0, 1]$ obtained by deleting at random observations that are equally spaced;

3. Group mean function: (M1) $\mu_A(t) = \mu_C(t) = \sin(t\pi)$, (M2) $\mu_A(t) = 0.5(1 - t)^2$, $\mu_C(t) = 0.1(t + 1)^2$, (M3) $\mu_A(t) = 3^{t^2}/2 + t^3 - 1.5t$; $\mu_C(t) = -5(t^2 - t)/3 + 0.2$.

4. Variance processes: (CV1) $X_i(t) = \sum_{k=1}^{K} \xi_{ik}\psi_k^{(1)}(t)$, $U_{id}(t) = \sum_{l=1}^{L} \zeta_{idl}\psi_l^{(2)}(t)$; (CV2) $X_i(t) = \sum_{k=1}^{K} \xi_{ik}\psi_k^{(1)}(t)$, $U_{iD} \sim GP\{0, \sigma_U^2 \rho_U(\cdot)\}$ where GP denotes a Gaussian process with mean 0, variance σ_U^2 and Matern auto-correlation function $\rho_U(t)$. The Matern auto-correlation function is defined as

$$\rho(\Delta; \phi, \kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{2\kappa^{1/2}\Delta}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{2\kappa^{1/2}\Delta}{\phi}\right) \quad (4)$$

where ϕ and κ are unknown parameters and K_κ is the modified Bessel function of order κ. We set $\sigma_U^2 = 1$ and the parameters of the Matern correlation function, $\rho_U(t)$, equal to $\kappa = 5$ and $\phi = 0.07$. Where appropriately, $\xi_{ik} \sim N(0, \lambda_k^{(1)})$, $\zeta_{idl} \sim N(0, \lambda_l^{(2)})$ for $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $\epsilon_{id}(t) \sim N(0, \sigma_\epsilon^2)$, for $d = A, C$. We set $K = 2$ and $L = 3$, $\lambda_k^{(1)} = 0.6 \times 2^{1-k}$; $\lambda_l^{(2)} = 2^{1-l}$ for $k = 1, 2$ and $l = 1, 2, 3$. Legendre polynomials were used for the process X, in particular $\psi_1^{(1)}(t) = \sqrt{3}(2t^2 - 1)$, $\psi_2^{(1)}(t) = \sqrt{5}(6t^2 - 6t + 1)$; and Fourier basis functions were used for the process U, in particular $\psi_1^{(2)}(t) = \sqrt{2}\sin(2\pi t)$, $\psi_2^{(2)}(t) = \sqrt{2}\cos(4\pi t)$ and $\psi_3^{(2)}(t) = \sqrt{2}\sin(4\pi t)$.

We used inferential methods described in Section 2: nonparametric estimation using nonparametric bootstrap (NE/NB), nonparametric estimation using parametric bootstrap (NE/PB), and parametric estimation using nonparametric bootstrap (PE/NB) using a different number of degrees of freedom. For PE/NB we used 4 (small), 7 (moderate) and 22 (large) number of degrees of freedom. Pointwise and joint confidence intervals were obtained and methods were compared in terms of actual coverage and length of the corresponding confidence intervals.

Integrated actual coverage for pointwise confidence intervals is calculated as $\text{IAC}_P = E[\int_0^1 1\{d(t) \in CI_P(t)\}\, dt]$, where $CI_P(t)$ is the pointwise confidence interval at time t, the expectation is taken with respect to the distribution of the confidence intervals and $1\{\cdot\}$ denotes the indicator function. In repeated samples IAC_P is estimated as $\widehat{\text{IAC}}_P = \sum_{b=1}^{B}\sum_{t=1}^{T} 1\{d(t) \in CI_P^{(b)}(t)\}/BT$, where B is the number of samples, T is the number of grid points and $CI_P^{(b)}(t)$ are the pointwise confidence intervals obtained in the bth simulation. Similarly, integrated actual coverage for joint confidence intervals is calculated as $\text{IAC}_J = E[1\{d(t) \in CI_J(t)\} : \text{for every } t \in [0,1]]$, where $CI_J(t)$ is the joint confidence interval at time t and the expectation is taken with respect to the distribution of the confidence intervals. In repeated samples IAC_J is estimated as $\widehat{\text{IAC}}_J = \sum_{b=1}^{B} 1\{d(t) \in CI_J^{(b)}(t) : \text{for every } t \in [0,1]\}/B$, where B is the number of samples and $CI_J^{(b)}(t)$ are the joint confidence intervals obtained in the bth simulation.

One can interpret IAC_P as the *average coverage probability* of pointwise confidence intervals (light gray areas throughout this paper) across grid points t. In contrast, IAC_J is the *probability that the entire function* is covered by the joint confidence intervals (dark gray extensions throughout this paper). The performance of both intervals is important, though only joint confidence intervals and IAC_J are directly related to answering the scientifically important questions: 1) "is there statistical evidence of difference between cases and controls?"; and 2) "if there is statistical evidence of difference then where is this evidence localized and how can it be quantified?"

Results are presented for sample sizes ranging from 30 to 200, with and without missing data. For the missing data scenarios 30% of the complete set of observations per subject were removed at random. We consider the case when the two group mean functions are equal and investigate the confidence intervals for different covariance structures. Tables 2 and 3 provide the integrated actual coverage and expected length of the various pointwise 90% and 95% confidence intervals. Results are compared to the t-test method based on empirical mean estimates that do not take into account smoothing. Because there are 100 observations per function this method is a particular case of parametric estimation with nonparametric bootstrap, where the mean function is estimated using 100 degrees of freedom. Thus, we denote this method PE/NB (100df).

Tables 3 and 5 indicate that confidence intervals that do not account for the smoothness of the mean function tend to be unnecessarily wide; compare columns labeled PE/NB (100df) with all other columns. The problem is even more serious when data are missing; compare results shown within brackets. The nonparametric estimation of the mean function with parametric or nonparametric bootstrap yield relatively similar confidence intervals, with respect to both coverage and length; compare results in columns labeled NE/NB and NE/PB in Tables 2-5. Parametric estimation using nonparametric bootstrap (labeled PE/NB) tend to have good coverage probability especially when the number of degrees of freedom of the fit is in a neighborhood of the true number of degrees of freedom. Our simulations seem to indicate that there is a wide range of number of degrees of freedom that provide reasonable results. This matches our experience that as long as the main features of the mean functions are captured, the coverage probabilities are remarkably robust to the choice of degrees of freedom. Thus, in general, it seems reasonable to choose a number of degrees of freedom that is likely to exceed the complexity of the functions. However, the length of PE/NB confidence intervals increases slowly with the number of degrees of freedom of the fit and becomes extreme when the maximum complexity of the model is reached. Thus, PE/NB could be recommended in situations where previous information about the expected complexity of the mean function exists or in cases where one expects to have very noisy empirical mean estimators. When the PE/NB strategy is employed the number of degrees of freedom has to be chosen a-priori; hunting for a number of degrees of freedom that provides some statistically significant results should be viewed as scientific cheating. The NE/NB and NE/PB methods provide a reasonable compromise for those situations when the scientist knows very little about the expected shape of the mean functions. They tend to trade some of the length of the confidence interval for the "piece of mind" provided by automatic smoothing.

Tables 4 and 5 provide the integrated actual coverage (IAC) and expected length of the joint 90% and 95% confidence intervals, respectively. While both NE/NB and NE/PB performed similarly, NE/PB required careful covariance modeling in the CV2 scenario. In this case the covariance decays slowly and a large number of eigenfunctions had to be retained to ensure at least 99% variance explained. While, in practice, more liberal thresholds, such as 90% or 95%, are use to model functional data, we argue that higher thresholds should be used for obtaining close to nominal coverage probabilities in most scenarios. As a final point, missing data is handled well by the three methods, at least when data are missing at random.

5 Discussion

The importance of multiple testing is established beyond a reasonable doubt, widely acknowledged, and almost universally ignored. For example, most, if not all, NIH grant applications contain phrases of the type "results will be corrected to account for multiple comparisons." At the other extreme most scientific papers report pointwise confidence intervals, or p-values that ignore the hundreds, thousands, or even millions of hypotheses that were tested before a "positive" result is found. This contributes to inflating the number of falsely positive "findings" and published research papers.

Table 2: Estimates of the integrated actual coverage (IAC) of the pointwise $(1-\alpha)100\%$ confidence intervals obtained with NB/NE, PB/NE and NB/PE with various degrees of freedom (df) for the fit. Results are presented for the two types of covariance structures for the case that data are observed completely (incompletely).

$1-\alpha$	I	CV	NB/PE (100)	NB/NE	PB/NE	NB/PE (4)	NB/PE (7)	NB/PE (22)
0.90	30	CV1	0.88 (0.89)	0.88 (0.89)	0.87 (0.90)	0.88 (0.89)	0.88 (0.89)	0.88 (0.88)
		CV2	0.88 (0.88)	0.87 (0.88)	0.89 (0.89)	0.87 (0.87)	0.87 (0.88)	0.88 (0.88)
	50	CV1	0.88 (0.89)	0.88 (0.89)	0.89 (0.90)	0.89 (0.89)	0.88 (0.88)	0.88 (0.88)
		CV2	0.88 (0.89)	0.88 (0.90)	0.90 (0.90)	0.86 (0.89)	0.87 (0.89)	0.88 (0.89)
	100	CV1	0.88 (0.90)	0.89 (0.92)	0.90 (0.89)	0.88 (0.92)	0.88 (0.91)	0.88 (0.91)
		CV2	0.89 (0.89)	0.89 (0.90)	0.89 (0.89)	0.89 (0.87)	0.88 (0.89)	0.89 (0.89)
	200	CV1	0.91 (0.90)	0.91 (0.89)	0.91 (0.89)	0.91 (0.89)	0.91 (0.89)	0.91 (0.89)
		CV2	0.89 (0.90)	0.89 (0.89)	0.91 (0.89)	0.91 (0.88)	0.90 (0.88)	0.89 (0.89)
0.95	30	CV1	0.93 (0.94)	0.93 (0.94)	0.93 (0.94)	0.93 (0.93)	0.93 (0.94)	0.93 (0.93)
		CV2	0.93 (0.94)	0.93 (0.94)	0.94 (0.94)	0.93 (0.94)	0.93 (0.93)	0.93 (0.93)
	50	CV1	0.94 (0.94)	0.94 (0.94)	0.94 (0.95)	0.95 (0.94)	0.94 (0.94)	0.94 (0.94)
		CV2	0.93 (0.94)	0.93 (0.95)	0.95 (0.95)	0.92 (0.95)	0.92 (0.94)	0.93 (0.94)
	100	CV1	0.94 (0.95)	0.95 (0.96)	0.95 (0.94)	0.94 (0.96)	0.95 (0.96)	0.94 (0.96)
		CV2	0.94 (0.95)	0.94 (0.95)	0.94 (0.94)	0.94 (0.93)	0.94 (0.94)	0.94 (0.94)
	200	CV1	0.95 (0.95)	0.95 (0.95)	0.95 (0.94)	0.96 (0.95)	0.96 (0.95)	0.95 (0.95)
		CV2	0.94 (0.95)	0.95 (0.94)	0.95 (0.94)	0.95 (0.93)	0.95 (0.94)	0.94 (0.94)

Table 3: Estimates of the integrated expected length (IEL) of the pointwise $(1-\alpha)100\%$ confidence intervals obtained with NB/NE, PB/NE and NB/PE with various degrees of freedom (provided between brackets) for the fit. Results are presented for the two types of covariance structures for the case that data are observed completely (incompletely).

$1-\alpha$	I	CV	NB/PE (100)	NB/NE	PB/NE	NB/PE (4)	NB/PE (7)	NB/PE (22)
0.90	30	CV1	1.11 (1.46)	1.08 (1.07)	1.11 (1.11)	1.00 (1.01)	1.08 (1.11)	1.09 (1.17)
		CV2	0.87 (1.18)	0.69 (0.69)	0.70 (0.69)	0.49 (0.51)	0.62 (0.65)	0.80 (0.88)
	50	CV1	0.87 (1.13)	0.85 (0.84)	0.85 (0.86)	0.78 (0.78)	0.85 (0.86)	0.85 (0.91)
		CV2	0.68 (0.91)	0.53 (0.54)	0.55 (0.54)	0.38 (0.40)	0.48 (0.51)	0.62 (0.69)
	100	CV1	0.62 (0.80)	0.60 (0.60)	0.61 (0.60)	0.56 (0.56)	0.61 (0.62)	0.61 (0.65)
		CV2	0.48 (0.65)	0.38 (0.39)	0.39 (0.39)	0.27 (0.28)	0.35 (0.36)	0.45 (0.49)
	200	CV1	0.44 (0.57)	0.43 (0.43)	0.43 (0.43)	0.40 (0.40)	0.43 (0.44)	0.43 (0.46)
		CV2	0.34 (0.46)	0.27 (0.28)	0.27 (0.27)	0.19 (0.20)	0.25 (0.26)	0.32 (0.35)
0.95	30	CV1	1.33(1.74)	1.28 (1.28)	1.32 (1.32)	1.19 (1.20)	1.29 (1.32)	1.30 (1.40)
		CV2	1.03 (1.41)	0.82 (0.82)	0.83 (0.82)	0.58 (0.61)	0.74 (0.78)	0.95 (1.05)
	50	CV1	1.04 (1.34)	1.01 (1.00)	1.02 (1.03)	0.93 (0.94)	1.01 (1.03)	1.02 (1.09)
		CV2	0.81 (1.09)	0.64 (0.65)	0.65 (0.64)	0.45 (0.47)	0.58 (0.61)	0.74 (0.82)
	100	CV1	0.74 (0.96)	0.72 (0.72)	0.72 (0.72)	0.67 (0.67)	0.72 (0.74)	0.73 (0.78)
		CV2	0.58 (0.77)	0.45 (0.46)	0.46 (0.46)	0.32 (0.34)	0.41 (0.43)	0.53 (0.58)
	200	CV1	0.53 (0.68)	0.51 (0.51)	0.51 (0.51)	0.47 (0.47)	0.51 (0.52)	0.51 (0.55)
		CV2	0.41 (0.55)	0.32 (0.33)	0.33 (0.33)	0.23 (0.24)	0.29 (0.31)	0.38 (0.41)

Table 4: Estimates of the integrated actual coverage (IAC) of the joint $(1-\alpha)100\%$ confidence intervals obtained with with NB/NE, PB/NE and NB/PE with various degrees of freedom (provided between brackets) for the fit. Results are presented for the two types of covariance structures for the case that data are observed completely (incompletely).

$1-\alpha$	I	CV	NB/PE (100)	NB/NE	PB/NE	NB/PE (4)	NB/PE (7)	NB/PE (22)
0.90	30	CV1	0.82 (0.76)	0.86 (0.84)	0.84 (0.83)	0.85 (0.85)	0.84 (0.82)	0.84 (0.82)
		CV2	0.68 (0.72)	0.77 (0.82)	0.82 (0.75)	0.84 (0.86)	0.78 (0.80)	0.72 (0.76)
	50	CV1	0.84 (0.82)	0.86 (0.88)	0.84 (0.85)	0.86 (0.88)	0.88 (0.88)	0.86 (0.82)
		CV2	0.80 (0.79)	0.85 (0.92)	0.86 (0.84)	0.86 (0.87)	0.84 (0.84)	0.85 (0.86)
	100	CV1	0.88 (0.88)	0.88 (0.92)	0.89 (0.82)	0.88 (0.92)	0.87 (0.92)	0.88 (0.90)
		CV2	0.84 (0.81)	0.86 (0.92)	0.90 (0.86)	0.86 (0.82)	0.88 (0.88)	0.86 (0.88)
	200	CV1	0.91 (0.88)	0.90 (0.91)	0.88 (0.84)	0.89 (0.87)	0.92 (0.88)	0.89 (0.88)
		CV2	0.88 (0.85)	0.90 (0.91)	0.89 (0.86)	0.90 (0.86)	0.90 (0.84)	0.88 (0.86)
0.95	30	CV1	0.89 (0.84)	0.90 (0.89)	0.86 (0.90)	0.91 (0.90)	0.90 (0.88)	0.90 (0.88)
		CV2	0.80 (0.85)	0.84 (0.90)	0.92 (0.88)	0.90 (0.92)	0.86 (0.86)	0.82 (0.88)
	50	CV1	0.90 (0.91)	0.92 (0.94)	0.91 (0.90)	0.92 (0.93)	0.92 (0.92)	0.92 (0.92)
		CV2	0.88 (0.90)	0.92 (0.95)	0.94 (0.91)	0.92 (0.94)	0.94 (0.91)	0.90 (0.93)
	100	CV1	0.92 (0.95)	0.94 (0.98)	0.96 (0.92)	0.95 (0.96)	0.93 (0.98)	0.94 (0.96)
		CV2	0.92 (0.88)	0.92 (0.94)	0.94 (0.92)	0.92 (0.92)	0.92 (0.94)	0.91 (0.94)
	200	CV1	0.94 (0.90)	0.94 (0.95)	0.94 (0.91)	0.93 (0.96)	0.94 (0.94)	0.94 (0.94)
		CV2	0.95 (0.90)	0.95 (0.95)	0.94 (0.92)	0.94 (0.92)	0.95 (0.91)	0.94 (0.92)

Table 5: Estimates of the integrated expected length (IEL) of the joint $(1-\alpha)100\%$ confidence intervals obtained with with NB/NE, PB/NE and NB/PE with various degrees of freedom (provided between brackets) for the fit. Results are presented for the two types of covariance structures for the case that data are observed completely (incompletely).

$1-\alpha$	I	CV	NB/PE (100)	NB/NE	PB/NE	NB/PE (4)	NB/PE (7)	NB/PE (22)
0.90	30	CV1	1.78 (2.74)	1.53 (1.57)	1.58 (1.59)	1.38 (1.41)	1.54 (1.63)	1.58 (1.90)
		CV2	1.63 (2.30)	1.14 (1.14)	1.16 (1.14)	0.72 (0.75)	1.00 (1.04)	1.41 (1.57)
	50	CV1	1.40 (2.12)	1.21 (1.23)	1.22 (1.23)	1.09 (1.10)	1.21 (1.27)	1.25 (1.47)
		CV2	1.28 (1.79)	0.89 (0.90)	0.91 (0.89)	0.56 (0.59)	0.78 (0.82)	1.11 (1.23)
	100	CV1	1.00 (1.52)	0.86 (0.88)	0.87 (0.86)	0.78 (0.79)	0.87 (0.91)	0.89 (1.06)
		CV2	0.92 (1.27)	0.63 (0.64)	0.64 (0.64)	0.40 (0.42)	0.56 (0.59)	0.79 (0.88)
	200	CV1	0.71 (1.07)	0.61 (0.62)	0.61 (0.61)	0.55 (0.56)	0.61 (0.65)	0.63 (0.75)
		CV2	0.65 (0.91)	0.45 (0.46)	0.46 (0.45)	0.28 (0.30)	0.40 (0.42)	0.56 (0.62)
0.95	30	CV1	1.96 (2.95)	1.72 (1.75)	1.77 (1.78)	1.56 (1.59)	1.73 (1.82)	1.77 (2.09)
		CV2	1.75 (2.45)	1.24 (1.25)	1.27 (1.24)	0.80 (0.83)	1.09 (1.14)	1.53 (1.70)
	50	CV1	1.54 (2.28)	1.35 (1.37)	1.37 (1.38)	1.22 (1.24)	1.36 (1.42)	1.39 (1.63)
		CV2	1.38 (1.91)	0.97 (0.98)	0.99 (0.98)	0.62 (0.65)	0.85 (0.90)	1.20 (1.33)
	100	CV1	1.10 (1.63)	0.97 (0.98)	0.97 (0.97)	0.87 (0.89)	0.97 (1.02)	0.99 (1.17)
		CV2	0.98 (1.36)	0.69 (0.70)	0.70 (0.70)	0.44 (0.47)	0.61 (0.64)	0.86 (0.95)
	200	CV1	0.78 (1.15)	0.68 (0.69)	0.69 (0.69)	0.62 (0.63)	0.69 (0.72)	0.70 (0.82)
		CV2	0.70 (0.96)	0.49 (0.50)	0.50 (0.50)	0.32 (0.33)	0.43 (0.46)	0.61 (0.68)

In this paper we provide simple and fast methods for testing if and where the means of two correlated functional processes are different. Joint confidence intervals that take into account the complexity of the data and sampling mechanisms are used as the basic inferential tools for making decisions. We contend that this and not the pointwise approach is the statistically principled approach to testing for a difference along the domain of the function. This approach is not conservative; instead, pointwise testing is too liberal.

We conclude that nonparametric estimation using nonparametric bootstrap that respects the data correlation structure is a powerful, simple, and practical method for making inference about the fixed effects of longitudinal functional models. In our case study a bootstrap of pairs is necessary to account for the sampling mechanism induced by matching. Nonparametric estimation using parametric bootstrap based on liberal choices of the number of eigenvectors is a viable alternative. This approach is slightly more computationally intensive but provides an excellent platform for generalization to more complex models. Parametric estimation using nonparametric bootstrap is a simple methodology that is especially appealing when some prior information about the shape of the mean functions is available.

References

[1] Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica* 1982; **50**:1029–1054.

[2] Hardin J, Hilbe J. *Generalized Estimating Equations.* Chapman & Hall/CRC, 2003.

[3] Liang KY, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**(1):13–22.

[4] Demidenko E. *Mixed Models: Theory and Applications.* John Wiley & Sons: Hoboken, New Jersey, 2004.

[5] McCulloch C, Shayle R Searle S, Neuhaus J. *Generalized, Linear, and Mixed Models.* John Wiley & Sons: Hoboken, New Jersey, 2008.

[6] Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**:823–836.

[7] Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data.* Springer Verlag: New York, 2000.

[8] Yin G. Bayesian Generalized Method of Moments. *Bayesian Analysis* 2009; **4(1)**:1–17.

[9] Beran J, Feng Y. Local Polynomial Estimation with FARIMA GARCH Error Process. *Bernoulli* 2001; **7**:733750.

[10] Currie I, Durban M. Flexible Smoothing with PSplines: A Unified Approach. *Statistical Modelling* 2002; **2**:333–349.

[11] Krivobokova T, Kauermann G. A note on penalized splines with correlated errors. *Journal of the American Statistical Association* 2007; **102**(480):1328–1337.

[12] Ray B, Tsay R. Bandwidth Selection for Kernel Regression with Long-Range Dependent Errors. *Biometrika* 1997; **84**:791–802.

[13] Wang Y. Smoothing Spline Models with Correlated Random Errors. *Journal of the American Statistical Association* 1998; **93**:341–348.

[14] Benko M, Härdle W, Kneip A. Common functional principal components. *Annals of Statistics* 2009; **37**:1–34.

[15] Hall P, Van Keilegom I. Two sample tests in functional data analysis, starting from discrete data. *Statistica Sinica* 2007; **17**:1511–1531.

[16] Zhang C, Peng H, Zhang JT. Two sample inference in functional linear models. *Communications in Statistics - Theory and Methods* 2010; **39**:559–578.

[17] Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, *et al.*. The Sleep Heart Health Study: design, rationale, and methods. *Sleep* 1997; **20**:1077–85.

[18] Crainiceanu CM, Staicu AM, Di C. Generalized multilevel functional regression. *Journal of the American Statistical Association* 2009; **104**:15501561.

[19] Crainiceanu C, Caffo B, Di CZ, Punjabi N. Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *Journal of the American Statistical Association* 2009; **104**(486):541–555.

[20] Di C, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. *Annals of Applied Statistics* 2009; **3(1)**:458–488. Online access 2008.

[21] Swihart B, Caffo B, Crainiceanu C, Punjabi N. Modeling multilevel sleep transitional data via poisson log-linear multilevel models. *Collection of Biostatistics Research Archive (COBRA)* November 2009; **Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 212.** URL `http://www.bepress.com/jhubiostat/paper212`.

[22] Swihart B, Caffo B, Bandeen-Roche K, Punjabi N. Characterizing sleep structure using the hypnogram. *Journal of Clinical Sleep Medicine* 2008; **4**(4):349–355.

[23] O'Sullivan F. A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science* 1986; **1**:505–527.

[24] Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press, 2003.

[25] Wood S. *Generalized Additive Models. An Introduction with R.* Chapman & Hall/CRC, 2006.

[26] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2009. URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

[27] Wand M, Coull B, French J, Ganguli B, Kammann E, Staudenmayer J, Zanobetti A. *SemiPar 1.0. R package.*

[28] Staniswalis J, Lee J. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 1998; **93**(444):14031418.

[29] Staicu AM, Crainiceanu CM, Carroll RJ. Fast methods for spatially correlated multilevel functional data. *Biostatistics* 2010; **11**.

[30] Indritz J. *Methods in Analysis.* Macmillan & Collier-Macmillan: New York, 1963.

[31] Karhunen K. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiæ Scientiarum Fennicæ, Series A1: Mathematica-Physica, Suomalainen Tiedeakatemia* 1947; **37**:3–79.

[32] Loève M. Functions Aleatoire de Second Ordre. *Comptes Rendus de l'Acadmie des Sciences* 1945; **220**.

[33] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 1995; **57**(1):289300.