Likelihood Ratio Tests for Dependent Data with Applications to Longitudinal and Functional Data Analysis

Ana-Maria Staicu* Yingxing Li† Ciprian M. Crainiceanu‡
David Ruppert§

October 13, 2013

Abstract

The paper introduces a general framework for testing hypotheses about the structure of the mean function of complex functional processes. Important particular cases of the proposed framework are: 1) testing the null hypotheses that the mean of a functional process is parametric against a nonparametric alternative; and 2) testing the null hypothesis that the means of two possibly correlated functional processes are equal or differ by only a simple parametric function. A global pseudo likelihood ratio test is proposed and its asymptotic distribution is derived. The size and power properties of the test are confirmed in realistic simulation scenarios. Finite sample power results indicate that the proposed test is much more powerful than competing alternatives. Methods are applied to testing the equality between the means of normalized

*Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203 USA. (email: staicu@stat.ncsu.edu). Research supported by NSF grant DMS1007466.

†The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China. (email: yxli@xmu.edu.cn). Research supported by NIH grant R01NS060910.

‡Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, E3636 Baltimore, MD, 21205, USA. (email: ccrainic@jhsph.edu). Research supported by NIH grant R01NS060910.

§Department of Statistical Science and School of Operations Research and Information Engineering, Cornell University, 1170 Comstock Hall, Ithaca, NY 14853, USA. (email: dr24@cornell.edu). Research supported by NIH grant R01NS060910.

δ-power of sleep electroencephalograms of subjects with sleep-disordered breathing and matched controls.

Some Key Words: functional data, longitudinal data, pseudo likelihood, Sleep Health Heart Study, two sample problem.

1 Introduction

We introduce pseudo likelihood ratio testing (pseudo LRT) for hypotheses about the structure of the mean of complex functional or longitudinal data. The main theoretical results are: 1) the asymptotic distribution of the pseudo LRT under general assumptions; and 2) simple sufficient conditions for these general assumptions to hold in the cases of longitudinal and functional data. The methods are applied to testing whether there is a difference between the average normalized δ-power of 51 subjects with sleep-disordered breathing (SDB) and 51 matched controls.

Tests of a parametric null hypothesis against a nonparametric alternative when the errors are independent and identically distributed has been under intense methodological development. For example, Fan, Zhang, and Zhang (2001) introduced a generalized LRT, while Crainiceanu and Ruppert (2004) and Crainiceanu *et al.* (2005) introduced a LRT. In contrast, development for non-independent errors has received less attention, although there are some results. For example, Guo (2002) and Antoniadis and Sapatinas (2007) considered functional mixed effects models using preset smoothing splines and wavelets bases respectively and discussed testing of fixed effects via LRTs; both approaches assume that the fixed and random functions are in the same functional space. Zhang and Chen (2007) proposed hypothesis testing about the mean of functional data based on discrepancy measures between the estimated means under the null and alternative models; the approach requires a dense sampling design. We propose a pseudo LRT for testing polynomial regression versus a general alternative modeled by penalized splines, when errors are correlated. The pseudo LRT does not assume the same smoothness property for the mean function and the random functional deviations, and it can be applied to dense or sparse functional data, with or without missing observations. Our simulation results show that in cases where the approach of Zhang and Chen applies, the pseudo LRT is considerably more powerful.

We consider a wider spectrum of null hypotheses, which includes the hypothesis that the

means of two functional processes are the same. Several recent methodological developments address this problem: Fan and Lin (1998) developed an adjusted Neyman testing procedure for independent stationary linear Gaussian processes; Cuevas, Febrero, and Fraiman (2004) proposed an F-test for independent processes; Staicu, Lahiri, and Carroll (2012) considered an L^2-norm-based global testing procedure for dependent processes; Crainiceanu *et al.* (2012) introduced bootstrap-based procedures using joint confidence intervals. Our pseudo LRT procedure has the advantage that it is applicable to independent or dependent samples of curves with both dense and sparse sampling design.

Our approach is based on modeling the mean function as a penalized spline with a mixed effect representation (Ruppert *et al.* 2003). Various hypotheses of interest can then be formulated as a combination of assumptions that variance components and fixed effects parameters are zero. When errors are independent and identically distributed (i.i.d.), testing for a zero variance component in this context is non-standard, as the parameter is on the boundary of the parameter space (Self and Liang, 1987) and the vector of observations cannot be partitioned into independent subvectors. In this case, Crainiceanu and Ruppert (2004) derived the finite sample and asymptotic null distributions of the LRT for the hypothesis of interest. However, in many practical situations the i.i.d. assumption is not fulfilled; for example when for each subject the outcome consists of repeated measures, the observations on each subject are likely to be correlated within the subject. We consider the latter case, that the errors have a general covariance structure, and propose a pseudo LRT obtained from the LRT by replacing the error covariance by a consistent estimator. Pseudo LRTs with parameters of interest or nuisance parameters on the boundary are discussed by Liang and Self (1996) and Chen and Liang (2010), respectively. Their derivations of the asymptotic null distributions require that the estimated nuisance parameters are $\sqrt{n}$-consistent—this assumption does not usually hold when the nuisance parameters have infinite dimension, e.g., for functional data.

We demonstrate that, if an appropriate consistent estimator of the error covariance is used, then the asymptotic null distribution of the pseudo LRT statistic is the same as the distribution of the LRT using the true covariance. For longitudinal data, we discuss some commonly used models and show that under standard assumptions in longitudinal data analysis (LDA) literature one obtains such a suitable consistent estimator of the covariance. For both densely and sparsely sampled functional data, we use smoothness assumptions standard in functional data analysis (FDA) literature to derive appropriate consistent esti-

mators of the covariance function. The methodology is extended to testing for differences between group means of two dependent or independent samples of curves, irrespective of their sampling design. The main innovations of this paper are the development of a rigorous asymptotic theory for testing null hypotheses about the structure of the population mean for clustered data using likelihood ratio-based tests, and the demonstration of its applicability to settings where the errors are longitudinal with parametric covariance structure, as well as when errors are contaminations of functional processes with smooth covariance structure.

The remainder of the paper is organized as follows. Section 2 presents the general methodology and the null asymptotic distribution of the pLRT for dependent data. Section 3 discusses applications of the pLRT to LDA and FDA. The finite sample properties of the pLRT are evaluated by a simulation study in Section 4. Testing equality of two mean curves is presented in Section 5 and illustrated using the Sleep Heart Health Study data in Section 6. A brief discussion is in Section 7 and details on available extra material are in Section 8.

2 Pseudo LRT for dependent data

In this section we describe the models and hypotheses considered, introduce the pseudo LRT for dependent data, and derive its asymptotic null distribution. Although our developments focus on the penalized spline class of functions, the results are general and can be used for other types of bases (B-splines, Fourier basis etc.) and other types of quadratic penalties.

Let Y_{ij} be the jth measurement of the response on the ith subject at time point t_{ij}, $1 \leq j \leq m_i$ and $1 \leq i \leq n$, and consider the model $Y_{ij} = \mu(t_{ij}) + e_{ij}$, where $\mu(\cdot)$ is the population mean curve and e_{ij} is the random deviation from the population mean curve. We are interested in testing the null hypothesis that

$$H_0 : \mu(t) = \beta_0 + \beta_1 t + \ldots + \beta_{p-q} t^{p-q}, \quad 0 \leq q \leq p \tag{1}$$

versus the alternative $H_A : \mu(t) = \beta_0 + \beta_1 t + \ldots + \beta_p t^p + \sum_{k=1}^{K} b_k (t - \kappa_k)_+^p$ when the errors e_{ij} are correlated over j, and their correlation is complex. Here $x_+^p = \max(0, x)^p$, $\kappa_1, \ldots, \kappa_K$ are knots placed at equally spaced quantiles and K is assumed to be large enough to ensure the desired flexibility (see Ruppert, 2002; Ruppert *et al.*, 2003). Also $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ is the vector of polynomial parameters and $\mathbf{b} = (b_1, \ldots, b_K)$ is the vector of spline coefficients. To avoid overfitting, we consider the approach proposed in Crainiceanu and Ruppert (2004), Crainiceanu et al. (2005) and represent the mean function via an equivalent mixed effects

representation, $\mu(t_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}+\mathbf{Z}_{ij}\mathbf{b}$, where $\mathbf{X}_{ij} = (1, t_{ij}, \ldots, t_{ij}^p)$, $\mathbf{Z}_{ij} = \{(t_{ij}-\kappa_1)_+^p, \ldots, (t_{ij}-\kappa_K)_+^p\}$, and $\mathbf{b}$ is assumed $N(0, \sigma_b^2 \boldsymbol{I}_K)$.

Let $\boldsymbol{X}_i$ the $m_i \times (p+1)$ dimensional matrix with the jth row equal to $\mathbf{X}_{ij}$, by $\boldsymbol{Z}_i$ the $m_i \times K$ dimensional matrix with jth row equal to $\mathbf{Z}_{ij}$. If $\mathbf{Y}_i$ is the $m_i \times 1$ dimensional vector of Y_{ij}, and $\mathbf{e}_i$ is the $m_i \times 1$ dimensional vector of e_{ij} then our model framework is

$$\begin{cases} \mathbf{Y}_i & = & \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathbf{b} + \mathbf{e}_i, \text{ for } i = 1, \ldots, n, \\ \mathbf{b} & \sim & N(\mathbf{0}, \sigma_b^2 \boldsymbol{I}_K) \\ \mathbf{e}_i & \sim & N(\mathbf{0}, \boldsymbol{\Sigma}_i) \\ \mathbf{b}, \mathbf{e}_i & & \text{independent}, \end{cases} \tag{2}$$

where the error covariance, $\boldsymbol{\Sigma}_i$, is assumed unknown and captures the within-cluster variability. Note that (2) is not the Laird and Ware model for longitudinal data (Laird and Ware, 1982), which requires $\mathbf{b}$ to depend on the cluster i and for $\mathbf{b}_1, \ldots, \mathbf{b}_n$ to be mutually independent. Thus, unlike standard LMMs the data in model (2) cannot be partitioned into independent subvectors. Therefore, standard asymptotic theory of mixed effects models does not directly apply to model (2), and different asymptotic distributions are obtained than in the Laird and Ware model. Additionally, (2) does not fall in the framework analyzed by Crainiceanu and Ruppert (2004) because of the assumed unknown non-trivial covariance structures $\boldsymbol{\Sigma}_i$. Many hypotheses of interest about the structure of the mean function $\mu(\cdot)$ are equivalent to hypotheses about the fixed effects $\beta_0, \ldots, \beta_p$ and the variance component σ_b^2. In particular, if $Q = \{0, 1, \ldots, p-q\}$, the null hypothesis (1) can be formulated as

$$H_0 : \beta_\ell = 0 \text{ for } \ell \in Q \text{ and } \sigma_b^2 = 0 \quad \text{versus} \quad H_A : \exists\, q_0 \in Q \text{ such that } \beta_{q_0} \neq 0 \text{ or } \sigma_b^2 > 0 \tag{3}$$

When $\boldsymbol{\Sigma}_i = \sigma_e^2 \mathbf{I}_{m_i}$ such hypotheses have been tested by Crainiceanu and Ruppert (2004) and Crainiceanu *et al.* (2005) using LRTs. Here we extend these results to the case when $\boldsymbol{\Sigma}_i$ is not necessary diagonal to capture the complex correlation structures of longitudinal and functional data; see Sections 3 and 5 for examples of commonly used $\boldsymbol{\Sigma}_i$. In Section 5 we also extend testing to include null hypotheses of no difference between the means of two groups. For now we focus on the simpler case, which comes with its own set of subtleties.

Our theoretical developments are based on the assumption that the distribution of $\mathbf{e}_i$'s is multivariate Normal, but the simulation results in Section 4 and the Web Supplement indicate that the null distribution of the pseudo LRT is robust to this assumption. Let $\mathbf{e}$ be the stacked vector of $\mathbf{e}_i$'s, $\mathbf{Y}$ the stacked vector of $\mathbf{Y}_i$'s, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ be the stacked matrices of $\boldsymbol{X}_i$'s and $\boldsymbol{Z}_i$'s, respectively. Also let $N = \sum_{i=1}^n m_i$ be the total number of

observations and $\boldsymbol{\Sigma}$ be an $N \times N$ block diagonal matrix, where the ith block is equal to $\boldsymbol{\Sigma}_i$, for $i = 1, \ldots, n$. When $\boldsymbol{\Sigma}$ is known, twice the log-likelihood of $\mathbf{Y}$ is, up to an additive constant, $2 \log L_{\mathbf{Y}}(\boldsymbol{\beta}, \sigma_b^2) = -\log(|\boldsymbol{\Sigma} + \sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}^T|) - (\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{\Sigma} + \sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}^T)^{-1}(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and the LRT statistic is $LRT_N = \sup_{H_0 \cup H_A} 2 \log L_{\mathbf{Y}}(\boldsymbol{\beta}, \sigma_b^2) - \sup_{H_0} 2 \log L_{\mathbf{Y}}(\boldsymbol{\beta}, \sigma_b^2)$. Here $|\cdot|$ is the determinant of a square matrix.

In practice, $\boldsymbol{\Sigma}$ is typically unknown; we propose testing the hypothesis (3) using the pseudo LRT, obtained by replacing $\boldsymbol{\Sigma}$ in the LRT by an estimate $\widehat{\boldsymbol{\Sigma}}$. Denote by $\boldsymbol{A}^{-1/2}$ a matrix square root of $\boldsymbol{A}^{-1}$, where $\boldsymbol{A}$ is a positive definite matrix, and let $\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Y}$, $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{X}$, $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{Z}$. Thus, twice the pseudo log likelihood is, up to a constant, $2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(\boldsymbol{\beta}, \sigma_b^2) = -\log |\widehat{\mathbf{H}}_{\sigma_b^2}| - (\widehat{\mathbf{Y}} - \widehat{\mathbf{X}}\boldsymbol{\beta})^T \widehat{H}_{\sigma_b^2}^{-1}(\widehat{\mathbf{Y}} - \widehat{\mathbf{X}}\boldsymbol{\beta})$, where $\widehat{\boldsymbol{H}}_{\sigma_b^2} = \boldsymbol{I}_N + \sigma_b^2 \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{Z}}^T$; the pseudo LRT statistic for testing (3) is defined as

$$pLRT_N = \sup_{H_0 \cup H_A} 2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(\boldsymbol{\beta}, \sigma_b^2) - \sup_{H_0} 2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(\boldsymbol{\beta}, \sigma_b^2). \tag{4}$$

The asymptotic null distribution of the pseudo LRT is discussed next.

PROPOSITION 2.1. *Suppose that* $\mathbf{Y}$ *is obtained from model (2), and assume a Gaussian joint distribution for* $\mathbf{b}$ *and* $\mathbf{e}$*, where* $\mathbf{e} = (\mathbf{e}_1^T, \ldots, \mathbf{e}_n^T)^T$*. In addition, assume the following:*

(C1) The null hypothesis H_0 *defined in (3) holds.*

(C2) The minimum eigenvalue of $\boldsymbol{\Sigma}$ *is bounded away from* 0 *as* $n \to \infty$*. Let* $\widehat{\boldsymbol{\Sigma}}$ *be an estimator of* $\boldsymbol{\Sigma}$ *satisfying* $\mathbf{a}^T\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{a} - \mathbf{a}^T\boldsymbol{\Sigma}^{-1}\mathbf{a} = o_p(1)$, $\mathbf{a}^T\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{e} - \mathbf{a}^T\boldsymbol{\Sigma}^{-1}\mathbf{e} = o_p(1)$*, where* $\mathbf{a}$ *is any* $N \times 1$ *non random normalized vector.*

(C3) There exists positive constants ϱ *and* ϱ' *such that* $N^{-\varrho}\mathbf{Z}^T\mathbf{Z}$ *and* $N^{-\varrho'}\mathbf{X}^T\mathbf{X}$ *converge to nonzero matrices. For every eigenvalue* $\tilde{\xi}_{k,N}$ *and* $\tilde{\zeta}_{k,N}$ *of the matrices* $N^{-\varrho}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}$ *and* $N^{-\varrho}\{\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\}$ *respectively, we have* $\tilde{\xi}_{k,N} \xrightarrow{P} \xi_k$ *and* $\tilde{\zeta}_{k,N} \xrightarrow{P} \zeta_k$ *for some* $\xi_1, \ldots, \xi_K$*,* $\zeta_1, \ldots, \zeta_K$ *that are not all* 0*.*

Let $LRT_\infty(\lambda) = \sum_{k=1}^K \frac{\lambda}{1+\lambda\zeta_k} w_k^2 - \sum_{k=1}^K \log(1 + \lambda\xi_k)$*,* $w_k \sim N(0, \zeta_k)$ *for* $k = 1, \ldots, K$*,* $\nu_j \sim N(0, 1)$ *for* $j = 1, \ldots, p - q + 1$*, and the* w_k*'s and* ν_j*'s are mutually independent. Then:*

$$pLRT_N \xrightarrow{D} \sup_{\lambda \geq 0} LRT_\infty(\lambda) + \sum_{j=1}^{p-q+1} \nu_j^2, \tag{5}$$

where the right hand side is the null distribution of the corresponding LRT based on the true model covariance Σ *(Crainiceanu and Ruppert, 2004).*

Here we used $\xrightarrow{P}$ to denote convergence in probability and $\xrightarrow{D}$ to denote convergence in distribution. The proof of Proposition 2.1, like all proofs, is given in the Web Supplement. The approach is reminiscent of Crainiceanu and Ruppert (2004). The main idea is to show that the components of the spectral decomposition of the pLRT converge in distribution to the counterparts corresponding to the true LRT based on the true covariance Σ. Accounting for correlated errors with unknown covariance requires tedious matrix algebra, inequalities with matrix norms, as well as application of the continuity theorem. Assumption (C2) provides a necessary condition for how close the estimated $\widehat{\mathbf{\Sigma}}^{-1}$ and the true $\mathbf{\Sigma}^{-1}$ precision matrices have to be. This condition (see also Cai, Liu and Luo, 2011) is related to the rate of convergence between the inverse covariance estimator and the true precision matrix in the spectral norm. For example, if $\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = o_p(1)$ then the first part of (C2) holds, where $\|\boldsymbol{A}\|_2$ denotes the spectral norm of a matrix $\boldsymbol{A}$ defined by $\|\boldsymbol{A}\|_2 = \sup_{|\mathbf{x}|_2 \leq 1} |\boldsymbol{A}\mathbf{x}|_2$ and $|\mathbf{a}|_2 = \sqrt{\sum_{i=1}^{r} a_i^2}$ for $\mathbf{a} \in R^r$. Such an assumption may seem difficult to verify, but in Sections 3 and 5 we show that it is satisfied by many estimators of covariance structures commonly employed in LDA and FDA. Assumption (C3) is standard in LRT; for example, when $\mathbf{Z}$ is the design matrix for truncated power polynomials with equally spaced knots (see Section 3.2), taking $\varrho = 1$ is a suitable choice (Crainiceanu, 2003).

Consider the particular case when there are m observations per subject and identical design points across subjects, i.e., $t_{ij} = t_j$, so that $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ do not depend on i and $\mathbf{\Sigma} = \boldsymbol{I}_n \otimes \mathbf{\Sigma}_0$ where $\otimes$ is the Kronecker product. Then (C2) is equivalent to:

(C2′) *The minimum eigenvalue of* $\mathbf{\Sigma}_0$ *is bounded away from* 0. *Let* $\widehat{\mathbf{\Sigma}}_0$ *be its consistent estimator satisfying* $\mathbf{a}^T\widehat{\mathbf{\Sigma}}_0^{-1}\mathbf{a} - \mathbf{a}^T\mathbf{\Sigma}_0^{-1}\mathbf{a} = o_p(1)$, *and* $\mathbf{a}^T\widehat{\mathbf{\Sigma}}_0^{-1}\mathbf{e}_0 - \mathbf{a}^T\mathbf{\Sigma}_0^{-1}\mathbf{e}_0 = o_p(1)$, *where* $\mathbf{a}$ *is any* $m \times 1$ *non random normalized vector and* $\mathbf{e}_0 = n^{-1/2}\sum_{i=1}^{n}\mathbf{e}_i$.

The asymptotic null distribution of $pLRT_N$ is not standard. However, as Crainiceanu and Ruppert (2004) point out, the null distribution can easily be simulated, once the eigenvalues ξ_k's and ζ_k's are determined. For completeness, we review their proposed algorithm.

Step 1 For a sufficiently large L, define a grid $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_L$ of possible values for λ.

Step 2 Simulate independent $N(0, \zeta_k)$ random variables w_k, $k = 1, \ldots, K$.

Step 3 Compute $LRT_\infty(\lambda)$ in (5) and determine its maximizer $\lambda_{\max}$ on the grid.

Step 4 Compute $pLRT = LRT_\infty(\lambda_{\max}) + \sum_{j=1}^{p-q+1}\nu_j^2$, where the ν_j's are i.i.d. $N(0,1)$.

Step 5 Repeat Steps 2–4.

The R package `RLRsim` (Scheipl, Greven, and Küchenhoff, 2008) or a MATLAB function `http://www.biostat.jhsph.edu/~ccrainic/software.html` can be used for implementation of Algorithm 1. It takes roughly 1.8 seconds to simulate 100,000 simulations from the null distribution using `RLRsim` on a standard computer (64-bit Windows with 2.8 GHz Processors and 24 GB random access memory).

3 Applications to longitudinal and functional data

We now turn our attention to global tests of parametric assumptions about the mean function in LDA and FDA and describe simple sufficient conditions under which assumption (C2) or (C2′) holds. This will indicate when results in Section 2 can be applied for testing.

3.1 Longitudinal data

Statistical inference for the mean function has been one of the main foci of LDA research (Diggle *et al.* 2002). Longitudinal data are characterized by repeated measurements over time on a set of individuals. Observations on the same subject are likely to remain correlated even after covariates are included to explain observed variability. Accounting for this correlation in LDA is typically done using several families of covariances. Here we focus on the case of commonly used parametric covariance structures. Consider the general model

$$Y_{ij} = \mu(t_{ij}) + e_i(t_{ij}), \qquad \text{cov}\{e_i(t_{ij}), e_i(t_{ij'})\} = \sigma_e^2 \varphi(t_{ij}, t_{ij'}; \boldsymbol{\theta}), \tag{6}$$

where t_{ij} is the time point at which Y_{ij} is observed and $\mu(t)$ is a smooth mean function. The random errors $e_{ij} = e_i(t_{ij})$ are assumed to have a covariance structure that depends on the variance parameter, σ_e^2, and the function $\varphi(\cdot, \cdot; \boldsymbol{\theta})$, which is assumed to be a positive definite function known up to the parameter $\boldsymbol{\theta} \in \Theta \subset R^d$.

Using the penalized spline representation of the mean function, $\mu(t_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\mathbf{b}$, the model considered here can be written in a LMM framework (2), where the covariance matrix $\boldsymbol{\Sigma}_i = \sigma_e^2 \boldsymbol{C}_i(\boldsymbol{\theta})$, and $\boldsymbol{C}_i(\boldsymbol{\theta})$ is an $m_i \times m_i$ dimensional matrix with the (j, j')th entry equal to $\varphi(t_{ij}, t_{ij'}; \boldsymbol{\theta})$. Hypothesis testing can then be carried out as in Section 2. Proposition 3.1 below provides simpler sufficient conditions for the assumption (C2) to hold.

PROPOSITION 3.1. *Suppose that for model (6) the number of observation per subject m_i is bounded, for all $i = 1, \ldots, n$, the regularity conditions (A1)-(A3) in the Supplementary Material hold, $\sigma_\epsilon^2 > 0$, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p(1)$, and $\widehat{\sigma}_e^2 - \sigma_e^2 = o_p(1)$. Then condition (C2) holds for $\widehat{\Sigma} = \widehat{\sigma}_e^2 \text{diag}\{C_1(\widehat{\boldsymbol{\theta}}), \ldots, C_n(\widehat{\boldsymbol{\theta}})\}$.*

One approach that satisfies these assumption is quasi-maximum likelihood estimation, as considered in Fan and Wu (2008). The authors proved that, under regularity assumptions that include (A1)-(A3), the quasi-maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$, and the nonparametric estimator $\widehat{\sigma}^2$ are asymptotically normal, with $\widehat{\boldsymbol{\theta}}$ having $\sqrt{n}$ convergence rate.

3.2 Functional data

In contrast to longitudinal data, where the number of time points is small, and simple correlation structures are warranted, functional data require flexible correlations structures; see Rice (2004) for a thorough discussion of longitudinal and functional data and analytic methods. It is theoretically and practically useful to think of functional data as realizations of an underlying stochastic process contaminated with noise.

Let Y_{ij} be the response for subject i at time t_{ij} as before, and assume that $Y_{ij} = \mu(t_{ij}) + V_i(t_{ij}) + \epsilon_{ij}$, where $\mu(\cdot) + V_i(\cdot)$ is the underlying process written in a form that emphasizes the mean function $\mu(\cdot)$ and the zero-mean stochastic deviation $V_i(\cdot)$, which is assumed to be squared integrable on a bounded and closed time interval $\mathcal{T}$, and ϵ_{ij} is the contaminating measurement error. It is assumed that $V_i(\cdot)$ are i.i.d. with covariance function $\text{cov}\{V_i(t), V_i(t')\} = \Gamma(t, t')$ that is continuous over $[0, 1]$. Mercer's lemma (see for example Section 1.2 of Bosq, 2000) implies a spectral decomposition of the function $\Gamma(\cdot, \cdot)$, in terms of eigenfunctions, also known as functional principal components, $\theta_k(\cdot)$, and decreasing sequence of non-negative eigenvalues σ_k^2 , $\Gamma(t, t') = \sum_k \sigma_k^2 \theta_k(t)\theta_k(t')$, where $\sum_k \sigma_k^2 < \infty$. Following the usual convention, we assume that $\sigma_1^2 > \sigma_2^2 > \ldots \geq 0$. The eigenfunctions form an orthonormal basis in the space of squared integrable functions and we may represent each curve using the Karhunen-Loève (KL) expansion (Karhunen, 1947; Loève, 1945) as $V_i(t) = \sum_{k \geq 1} \xi_{ik}\theta_k(t), t \in [0, 1]$, where ξ_{ik} are uncorrelated random variables with mean zero and variance $E[\xi_{ik}^2] = \sigma_k^2$. Thus out model can be represented as

$$Y_{ij} = \mu(t_{ij}) + \sum_{k \geq 1} \xi_{ik}\theta_k(t_{ij}) + \epsilon_{ij}, \tag{7}$$

where ϵ_{ij} are assumed i.i.d. with zero-mean and finite variance $E[\epsilon_{ij}^2] = \sigma_\epsilon^2$. Our objective is to test polynomial hypotheses about $\mu(\cdot)$ using the proposed pseudo LRT. As argued in Proposition 2.1 this testing procedure relies on an accurate estimator of the model covariance, and, thus, of the covariance function $\Gamma(\cdot, \cdot)$ and the noise variance σ_ϵ^2.

The FDA literature contains several methods for obtaining consistent estimators of both the eigenfunctions/eigenvalues and the error variance; see for example Ramsay and Silverman (2005), Yao, Müller and Wang (2005). Furthermore, properties of the functional principal component estimators, including their convergence rates, have been investigated by a number of researchers (Hall and Hosseini-Nasab, 2006; Hall, Müller and Wang, 2006; Li and Hsing, 2010, etc.) for a variety of sampling design scenarios. In particular for a dense sampling design, where $m_i = m$, Hall *et al.* (2006) argue that one can first construct de-noised trajectories $\widehat{Y}_i(t)$ by running a local linear smoother over $\{t_{ij}, Y_i(t_{ij})\}_j$, and then estimate all eigenvalues and eigenfunctions by conventional PCA as if $\widehat{Y}_i(t)$ were generated from the true model and without any error. They point out that when $m = n^{1/4+\nu}$ for $\nu > 0$ and the smoothing parameter is appropriately chosen, one can obtain estimators of eigenfunctions/eigenvalues with $\sqrt{n}$ consistency. Of course, for a sparse sampling design, the estimators enjoy different convergence rates.

For our theoretical developments we assume that in (7), ξ_{ik} and ϵ_{ij} are jointly Gaussian. This assumption has been commonly employed in functional data analysis; see for example Yao, et al. (2005). Simulation results, reported in Section 4.1, indicate that the proposed method is robust to violations of the Gaussian assumption. Moreover, we assume that the covariance function Γ has M non-zero eigenvalues, where $1 \leq M < \infty$. The number of eigenvalues M is considered unknown and it can be estimated using the percentage of variance explained, AIC, BIC or testing for zero variance components, as discussed Staicu, Crainiceanu and Carroll (2010). We use the percentage variance explained in the simulation experiment and the data analysis. Next, we discuss the pseudo LRT procedure separately for the dense sampling design and for the sparse sampling design. More specifically we discuss conditions such that the requirement (C2) of Proposition 2.1 holds.

Dense sampling design. This design refers to the situation where the times, at which the trajectories are observed, are regularly spaced in $[0, 1]$ and increase to ∞ with n. We assume that each curve i is observed at common time points, i.e., $t_{ij} = t_j$ for all $j = 1, \ldots, m$. Thus $\mathbf{\Sigma}_i$ is the same for all subjects, say $\mathbf{\Sigma}_i = \mathbf{\Sigma}_0$ for all i.

PROPOSITION 3.2. *Consider that the above assumptions for model (7) hold. Assume the*

following conditions hold:

(F1) If $\widehat{\theta}_k(t)$, $\widehat{\sigma}_k^2$, *and* $\widehat{\sigma}_\epsilon^2$ *denote the estimators of the eigenfunctions, eigenvalues, and noise variance correspondingly, then*
$\|\|\widehat{\theta}_k - \theta_k\| = O_p(n^{-\alpha})$, $\widehat{\sigma}_k^2 - \sigma_k^2 = O_p(n^{-\alpha})$, *and* $\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = O_p(n^{-\alpha})$.

(F2) We have $m \sim n^\delta$ *where* $0 < \delta < 2\alpha$.

Then (C2) of Proposition 2.1 holds for the estimator $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}_0$ *of* $\boldsymbol{\Sigma}$, *where* $\widehat{\boldsymbol{\Sigma}}_0$ *is*

$$[\widehat{\boldsymbol{\Sigma}}_0]_{jj'} = \sum_{k=1}^{M} \widehat{\sigma}_k^2 \widehat{\theta}_k(t_j)\widehat{\theta}_k(t_{j'}) + \widehat{\sigma}_\epsilon^2 1(t_j = t_{j'}), \qquad 1 \le j, j' \le m. \tag{8}$$

The proposition is shown by using the Woodbury formula (Woodbury, 1950) to simplify the expression of $\widehat{\boldsymbol{\Sigma}}_0^{-1}$; the result follows then from employing triangle inequality for matrix norms, as well as central limit theorem and Chebyshev's inequality. Assumption (F1) concerns the L^2 convergence rate of the estimators; for local linear smoothing, Hall, et al. (2006) showed that the optimal L^2 rate is $n^{-\alpha}$ where $\alpha = 1/2$. Condition (F2) imposes an upper bound on the number of repeated measurements per curve: this requirement is needed in the derivation of the asymptotic null distribution of the pseudo LRT. In particular, when linear smoothing is used and $\alpha = 1/2$ (see Hall et al., 2006), condition (F2) reduce to $m = n^\delta$, for $1/4 < \delta < 1$. Nevertheless, empirical results showed that the pseudo LRT performs well, even when applied to settings where the number of repeated measurements is much larger than the number of curves. In particular, Section 4.1 reports reliable results for the pseudo LRT applied to data settings where m is up to eight times larger than n.

Remark 1. An alternative approach for situations where m is much larger than n is to use the following two-step procedure. First estimate the eigenfunctions / eigenvalues and the noise variance using the whole data, and then apply the pseudo LRT procedure only to a subset of the data that corresponds to suitably chosen subset of time points $\{\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}\}$ where $\widetilde{m}$ is such that it satisfies assumption (F2). Our empirical investigation of this approach shows that the power does not change with m and that there is some loss of power for smaller sample sizes n. However, the power loss decreases as n increases. The alternative approach is designed for use with large m and can be used for, say, $m > 1000$ with only a negligible loss of power. Even for smaller value of m, we find that our test is more powerful than its competitor, the test due to Zhang and Chen (2007).

Remark 2. The result in Proposition 3.2 can accommodate situations when data are missing at random. More precisely, let $t_1, \ldots, t_m$ be the grid of points in the entire data and denote by n_j the number of observed responses Y_{ij} corresponding to time t_j. Under the assumption that $n_j/n \to 1$ for all j, the conclusion of Proposition 2.1 still holds.

Sparse sampling design. Sparse sampling refers to the case when observation times vary between subjects and the number of observations per subject, m_i, is bounded and small. Examples of sparse sampling are auction bid prices (Jank and Shmueli, 2006), growth data (James, Hastie and Sugar, 2001), and many observational studies. The following proposition presents simplified conditions under which the requirement (C2) of Proposition 2.1 is satisfied. The main idea is to view the sparsely observed functional data as incomplete observations from dense functional data.

PROPOSITION 3.3. *Consider that the above assumptions about the model (7) are met. In addition assume the following conditions:*

(F1') *The number of measurements per subject is finite, i.e.,* $\sup_i m_i < \infty$. *Furthermore it is assumed that, for each subject* i, *the corresponding design points* $\{t_{ij} : j = 1, \ldots, m_i\}$ *are generated uniformly and without replacement from a set* $\{t_1, \ldots, t_m\}$, *where* $t_k = (k - 1/2)/m$, *for* $k = 1, \ldots, m$ *and* m *diverges with* n.

(F2') $\sup_{t \in \mathcal{T}} |\widehat{\theta}_k(t) - \theta_k(t)| = O_p(n^{-\alpha})$, $\widehat{\sigma}_k^2 - \sigma_k^2 = O_p(n^{-\alpha})$, *and* $\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = O_p(n^{-\alpha})$.

(F3') *We have* $m \sim n^\delta$ *where* $0 < \delta < 2\alpha$.

Then condition (C2) holds for the estimator $\widehat{\mathbf{\Sigma}} = \text{diag}\{\widehat{\mathbf{\Sigma}}_1, \ldots, \widehat{\mathbf{\Sigma}}_n\}$ *of* $\mathbf{\Sigma}$, *where the* $m_i \times m_i$ *matrix* $\widehat{\mathbf{\Sigma}}_i$ *is defined similarly to (8) with* $(t_j, t_{j'})$ *replaced by* $(t_{ij}, t_{ij'})$ *and* m *replaced by* m_i.

Proposition 3.3 is proved following roughly similar logic as the proof of Proposition 3.2. However the sparseness assumption makes the justification more challenging, especially when proving the second part of condition (C2). The key idea relies in the application of assumption (F1'); the result follows from using continuity theorem, as well as Bonferroni and Chebyshev's inequalities. Condition (F1') can be weaken for design points that are generated from a uniform distribution. In such cases, the design points are rounded to the nearest $t_k = (k - 1/2)/m$, and can be viewed as being sampled uniformly without replacement from $\{t_1, \ldots, t_m\}$ for some $m \to \infty$. Because of the smoothness intrinsic to functional data (observed without noise), the effect of this rounding is asymptotically negligible when $m \to \infty$

at a rate faster than $n^{-\alpha}$. Thus condition (F1') can be relaxed to assuming that t_{ij}'s are uniformly distributed between 0 and 1. Because $0 < \delta < 2\alpha$, if m to grows at rate n or faster then the alternative approach is needed. Condition (F2') regards uniform convergence rates of the covariance estimator; see also Li and Hsing (2010). For local linear estimators Yao *et al.* (2005) showed that, under various regularity conditions, the uniform convergence rate is of order $n^{-1/2}h_\Gamma^{-2}$, where h_Γ is the bandwidth for the two-dimensional smoother and is selected such that $nh_\Gamma^{2\ell+4} < \infty$, $\ell > 0$. When the smoothing parameter is chosen appropriately, and $\ell = 4$, the convergence rate is of order $O_p(n^{-1/3})$; thus conditions (F1') and (F3') reduce to $m = n^\delta$, for $\delta < 2/3$.

In summary, tests of the mean function in both densely and sparsely observed functional data can be carried out in the proposed pseudo LRT framework. Under the assumptions required by Propositions 3.2 and 3.3 respectively, and under the additional assumptions (C1) and (C3) of Proposition 2.1, the asymptotic null distribution of the pseudo LRT with the covariance estimator $\widehat{\mathbf{\Sigma}}$ is the same as if the true covariance were used and is given by (5).

4 Simulation study

In this section we investigate the finite sample Type I error rates and power of the pseudo LRT. Each simulated data set has n subjects. The data, $Y_i(t)$, for subject i, $i = 1, \ldots, n$, and timepoint t, $t \in \mathcal{T} = [0, 1]$, are generated from model (7) with scores ξ_{ik} that have mean zero and variance $E[\xi_{ik}^2] = \sigma_k^2$, where $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.25$, and $\sigma_k^2 = 0$ for all $k \geq 4$. Also $\theta_{2k-1}(t) = \sqrt{2}\cos(2k\pi t)$ and $\theta_{2k}(t) = \sqrt{2}\sin(2k\pi t)$ for all $k \geq 1$. The interest is in testing the hypothesis H_0: $\mu(t) = 0, \forall t \in [0, 1]$, versus H_A: $\mu(t) \neq 0$ for some t. We varied μ in a family of functions parameterized by a scalar parameter $\rho \geq 0$ that controls the departure from H_0, with $\rho = 0$ corresponding to H_0. This family consists of increasing and symmetric functions $\mu_\rho(t) = \rho/\{1 + e^{10(0.5-t)}\} - \rho/2$. We used two noise variances: $\sigma_\epsilon^2 = 0.125$ (small) and $\sigma_\epsilon^2 = 2$ (large). All results are based on 1000 simulations.

4.1 Dense functional data

In this scenario, each subject is observed at m equally spaced time points $t_j = (j - 1/2)/m$, for $j = 1, \ldots, m$. We consider two types of generating distributions for the scores, ξ_{ik}: in one setting they are generated from a Normal distribution, $N(0, \sigma_k^2)$, while in another

setting they are generated from a mixture distribution of two Normals $N(-\sqrt{\sigma_k^2/2}, \sigma_k^2/2)$ and $N(\sqrt{\sigma_k^2/2}, \sigma_k^2/2)$ with equal probability. We model the mean function using linear splines with K knots. The choice of K is not important, as long as it is large enough to ensure the desired flexibility (Ruppert, 2002). We selected the number of knots, based on the simple default rule of thumb $K = \max\{20, \min(0.25 \times \text{number of unique } t_j, 35)\}$ inspired from Ruppert et al. (2003). The pseudo LRT requires estimation of the covariance function, $\mathbf{\Sigma}$, or, equivalently, $\mathbf{\Sigma}_0$; see Section 3.2. This step is crucial as the accuracy of the covariance estimator has a sizeable impact on the performance of the pseudo LRT.

Let $\widetilde{G}(t_j, t_{j'})$ be the sample covariance estimator of $\text{cov}\{Y_i(t_j), Y_i(t_{j'})\}$, and let $\widehat{G}(\cdot, \cdot)$ be obtained by smoothing $\{\widetilde{G}(t_j, t_{j'}) : t_j \neq t_{j'}\}$ using a bivariate thin-plate spline smoother. We used the `R` package `mgcv` (Wood, 2006), with the smoothing parameter selected by restricted maximum likelihood (REML). The noise variance is estimated by $\widehat{\sigma}_\epsilon^2 = \int_0^1 \{\widetilde{G}(t,t) - \widehat{G}(t,t)\}_+ dt$; if this estimate is not positive then it is replaced by a small positive number. Denote by $\widehat{\sigma}_k^2$ and $\widehat{\theta}_k$ the kth eigenvalue and eigenfunction of the covariance $\widehat{G}$, for $k \geq 1$. The smoothing-based covariance estimator, $\widehat{\Sigma}_0$, is determined using expression (8), where M, the number of eigenvalues/eigenfunction is selected using the cumulative percentage criterion (see for example Di et al., 2009). In our simulation study, we used M corresponding to 99% explained variance. Once $\widehat{\mathbf{\Sigma}}_0$ is obtained, the data are "pre-whitened" by multiplication with $\widehat{\Sigma}_0^{-1/2}$. Then, the pseudo LRT is applied to the transformed data. The p-value of the test is automatically obtained from the function `exactLRT` (based on 10^5 replications) of the `R` package `RLRsim` (Scheipl, et al., 2008), which implements Algorithm A1, given in Section 2.

Table 1 shows the Type I error rates of the pseudo LRT corresponding to nominal levels $\alpha = 0.20, 0.10, 0.05$ and 0.01, and for various sample sizes ranging between $n = 50$ and $n = 200$ and m ranging between 80 and 400. Table 1 shows that the pseudo LRT using a smooth estimator of the covariance has Type I error rates that are close to the nominal level, for all significance levels. The results also indicate that the performance of the pseudo LRT is robust in regard to violations of the Gaussian assumption on the scores; see the lines corresponding to 'non-normal' for the distribution of the scores. This is corroborated by further investigation for the case when the scores ξ_{ik} are generated using scaled t_5 (heavy tailed) or centered and scaled χ_5^2 (skewed) distributions; see Table 1 in the Web Supplement.

Figure 1 shows the power functions for testing the null hypothesis $H_0 : \mu \equiv 0$. The results are only little affected by the magnitude of noise, and for brevity we only present the case of low noise level. The solid lines correspond to pseudo LRT with smooth covariance estimator,

Table 1: *Type I error rates, based on* 1000 *simulations, of the pseudo LRT for testing* $H_0 : \mu \equiv 0$ *in the context of dense functional data generated by model (7) with* $\sigma_\epsilon^2 = 0.125$*, for various* n *and* m*, and when the scores* ξ_{ik} *are generated from a Normal distribution (normal) or mixture distribution of two Normals (non-normal). In the pseudo LRT, the mean function is modeled using linear splines.*

(n, m)	scores distribution	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
(50, 100)	normal	0.216	0.111	0.057	0.021
(50, 400)	normal	0.236	0.124	0.068	0.016
(100, 100)	normal	0.209	0.115	0.054	0.009
(100, 400)	normal	0.220	0.112	0.059	0.013
(200, 80)	normal	0.217	0.099	0.054	0.012
(50, 100)	non-normal	0.209	0.126	0.060	0.012
(50, 400)	non-normal	0.223	0.129	0.076	0.010
(100, 100)	non-normal	0.222	0.112	0.053	0.010
(100, 400)	non-normal	0.215	0.127	0.062	0.016
(200, 80)	non-normal	0.199	0.103	0.052	0.009

the dashed lines correspond to the LRT test with known covariance matrix, and the dotted lines correspond to the global L^2-norm-based test of Zhang and Chen (2007), henceforth denoted ZC test. The performance of the pseudo LRT with the smooth covariance estimator is very close to its counterpart based on the true covariance; hence the pronounced overlap between the solid and dashed lines of the Figure 1. Overall, the results indicate that the pseudo LRT has excellent power properties, and furthermore that the power slightly improves as the number of measurements per subject m increases. Intuitively, this should be expected as a larger number of sampling curves per curve, m, corresponds more available information about the process, and thus about the mean function. By comparison, the power of the L^2 norm-based test is very low and it barely changes with m. In further simulations not reported here in the interest of space, the only situation we found where the ZC test becomes competitive for the pseudo LRT is when the deviation of the mean function from the function specified by the null hypothesis is confined to the space spanned by the eigenfunctions of the covariance function of the curves. In fact, the asymptotic theory in Zhang and Chen's

Theorem 7 suggests that this would be the case where their test is most powerful.

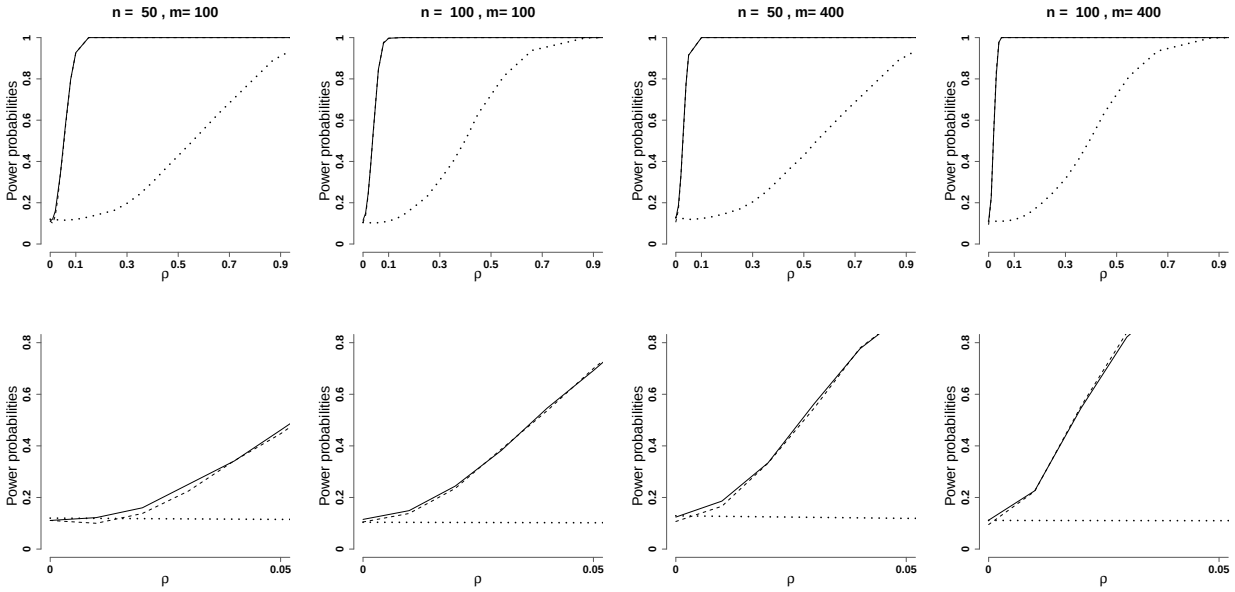

Figure 1: The power functions for testing $H_0 : \mu \equiv 0$ for dense functional data generated from model (7) with true mean function parameterized by ρ, for low noise variance $\sigma_\epsilon^2 = 0.125$. Top panels: power probabilities for different sample sizes n and number of measurements per curve m. Bottom panels: power probabilities for the same scenarios as the top panels, for $\rho \in [0, 0.05]$ to show detail in the low power region. Results are for the pseudo LRT based on the true covariance (dashed line), the smooth covariance estimator (solid line), ZC's L^2 norm-based test (dotted line) and for a nominal level $\alpha = 0.10$.

4.2 Sparse functional data

We now consider the case when each subject is observed at m_i time points $t_{ij} \in [0, 1]$, $j = 1, \ldots, m_i$, generated uniformly from the set $\{t_j = (j-1/2)/m : j = 1, \ldots, m\}$, where $m = 75$. There are $n = 250$ subjects and an equal number $m_i = 10$ time points per subject. The main difference from the dense sampling case is the calculation of the covariance estimator. For sparse data we start with a raw undersmooth covariance estimator based on the pooled data. Specifically, we first center the data $\{Y_i(t_{ij}) - \widetilde{\mu}(t_{ij})\}$, using a pooled undersmooth estimator of the mean function, $\widetilde{\mu}(t_j)$, and then construct the sample covariance of the centered data, using complete pairs of observations. At the second step, the raw estimator is smoothed using the `R` package `mgcv` (Wood, 2006). The smoothing parameter is selected via a modified generalized cross validation (GCV) or the un-biased risk estimator (UBRE) using $\gamma > 1$ to increases the amount of smoothing (Wood, 2006). The data $\{Y_i(t_{ij}) - \widetilde{\mu}(t_{ij})\}$ are correlated

which causes undersmoothing, but using $\gamma > 1$ counteracts this effect. Reported results are based on $\gamma = 1.5$, a choice which was observed to yield good covariance estimators in simulations for various sample sizes. Further investigation of the choice of γ would be useful but is beyond the scope of this paper.

Table 2: *Type I error rates based on* 1000 *simulations when testing* $H_0 : \mu \equiv 0$ *with sparse functional data,* $n = 250$ *subjects and* $m_i = 10$ *observations per subject. Pseudo LRT with the true covariance (true) and a smoothing-based estimator of the covariance (smooth) are compared. The mean function is modeled using linear splines with* $K = 20$ *knots.*

σ^2_ϵ	method	cov. choice	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
0.125	LRT	true	0.213	0.109	0.055	0.009
	pLRT	smooth	0.210	0.113	0.062	0.017
2	LRT	true	0.207	0.092	0.051	0.011
	pLRT	smooth	0.196	0.089	0.043	0.012

Table 2 illustrates the size performance of the pseudo LRT for sparse data, indicating results similar to the ones obtained for the dense sampling scenario. Figure 2 (bottom panels) shows the power functions for testing $H_0 : \mu \equiv 0$ when the true mean function is from the family described earlier. For the large noise scenario, $\sigma^2_\epsilon = 2$, the results of the pseudo LRT with the smooth covariance estimator are very close to the counterparts based on the true covariance. This is expected, as the noise is relatively easier to estimate, and thus when the noise is a large part of the total random variation, then a better estimate of the covariance function is obtained. On the other hand, having large noise affects the power negatively. When the noise has a small magnitude, the power when the covariance estimator is used is still very good and relatively close to the power when the true covariance function is used. Results for ZC are not shown because their approach requires densely sampled data.

5 Two samples of functional data

As with scalar or multivariate data, functional data are often collected from two or more populations, and we are interested in hypotheses about the differences between the popula-

tion means. Here we consider only the case of two samples both for simplicity and because the example in Section 6 has two samples.

Again as with scalar or multivariate data, the samples can be independent or paired. The experimental cardiology study discussed in Cuevas *et al.* (2004), where calcium overload was measured at a frequency of 10s for one hour in two independent groups (control and treatment), is an example of independent samples of functional data. In the matched case-control study considered in Section 6, Electroencephalogram (EEG) data collected at a frequency of 125Hz for over 4 hours for an apneic group and a matched healthy control group; the matching procedure induces dependence between cases and controls. For other examples of dependent samples of functional data see, for example, Morris and Carroll (2006), Di *et al.* (2009), and Staicu *et al.* (2010).

We discuss global testing of the null hypothesis of equality of the mean functions in two samples of curves. Results are presented separately for independent and dependent functional data. Testing for the structure of the mean difference in two independent samples of curves can be done by straightforwardly extending the ideas presented in Section 3.2. In the interest of space, the details are described in the Web Supplement. Here we focus on the case when the two sets of curves are dependent, and furthermore when in each set, the curves are sparsely sampled.

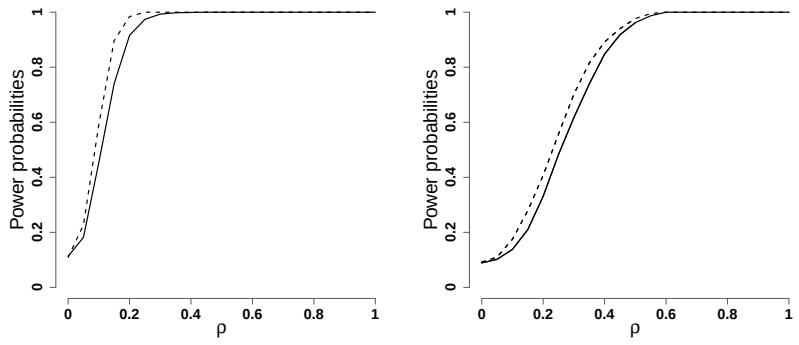

Figure 2: The power functions for testing $H_0 : \mu \equiv 0$ for sparse functional data generated from model (7) with true mean function parameterized by ρ, for two noise magnitudes $\sigma_\epsilon^2 = 0.125$ (left panel) and $\sigma_\epsilon^2 = 2$ (right panel). The results are for the pseudo LRT based on the true covariance (dashed line) and the smooth covariance estimator (solid line) and for a nominal level $\alpha = 0.10$.

5.1 Dependent samples of functional data

We use the functional ANOVA framework introduced by Di *et al.* (2009) and discuss inference for the population-level curves. Let $Y_{idj} = Y_{id}(t_{idj})$ be response for cluster i and group d at time point t_{idj}. For example, in the application in Section 6, the clusters are the matched pairs and the groups are subjects with SDB and controls. Let Y_{idj} be modeled as

$$Y_{id}(t_{idj}) = \mu(t) + \mu_d(t_{idj}) + V_i(t_{idj}) + W_{id}(t_{idj}) + \epsilon_{idj}, \tag{9}$$

where $\mu(t)$ is the overall mean function, $\mu_d(t)$ is the group-specific mean function, $V_i(t)$ is the cluster-specific deviation at time point t, $W_{id}(t)$ is the cluster-group deviation at t, ϵ_{idj} is the measurement error and $t_{idj} \in \mathcal{T}$ for $i = 1, \ldots, n$, $d = 1, 2$, and $j = 1, \ldots, m_{id}$. For identifiability we assume that $\mu_1 + \mu_2 \equiv 0$. It is assumed that level 1 (subject) random functions, V_i, and level 2 (subject-group) random functions, W_{id}, are uncorrelated mean zero stochastic processes with covariance functions $\Gamma_1(\cdot, \cdot)$ and $\Gamma_2(\cdot, \cdot)$ respectively (Di *et al.*, 2009). Furthermore, it is assumed that ϵ_{idj}'s are independent and identically distributed with mean zero and variance $E[\epsilon_{idj}] = \sigma_\epsilon^2$ and independent of V_i's and W_{id}'s. As in Section 3.2, let the basis expansions of $\Gamma_1(\cdot, \cdot)$ and $\Gamma_2(\cdot, \cdot)$ be: $\Gamma_1(t, t') = \sum_{k \geq 1} \sigma_{1,k}^2 \theta_{1,k}(t)\theta_{1,k}(t')$, and $\Gamma_2(t, t') = \sum_{l \geq 1} \sigma_{2,l}^2 \theta_{2,l}(t)\theta_{2,l}(t')$. Here $\sigma_{1,1}^2 > \sigma_{1,2}^2 > \ldots$ are the level 1 ordered eigenvalues and $\sigma_{2,1}^2 > \sigma_{2,2}^2 > \ldots$ are the level 2 ordered eigenvalues. Then V_i and W_{id} can be approximated by the KL expansion: $V_i(t) = \sum_{k \geq 1} \xi_{ik}\theta_{1,k}(t)$, $W_{id}(t) = \sum_{l \geq 1} \zeta_{idl}\theta_{2,l}(t)$, where ξ_{ik} and ζ_{idl} are principal component scores withe mean zero and variance equal to $\sigma_{1,k}^2$ and $\sigma_{2,l}^2$. As before it is assumed that the covariance functions have finite non-zero eigenvalues and in addition that ξ_{ik}, ζ_{idl} and ϵ_{idj} are mutually independent and they are jointly Gaussian distributed.

The main objective is to test that the group mean functions are equal, or equivalently that $\mu_1 \equiv 0$. Irrespective of the sampling design (dense or sparse), we assume that the set of *pooled* time points, $\{t_{idj} : i, j\}$ is dense in $\mathcal{T}$ for each d. Our methodology requires that the same sampling scheme is maintained for the two samples of curves, e.g., the curves are either sparsely observed in both samples or densely observed in both samples. (One could extend the theory to the case of one sample being densely observed and the other sparse, but data of this type would be rare so we did not attempt such an extension.) We use quasi-residuals, $\widetilde{Y}_{idj} = Y_{id}(t_{idj}) - \widetilde{\mu}(t_{idj})$, where $\widetilde{\mu} = (\widetilde{\mu}_1 + \widetilde{\mu}_2)/2$ is the average of the estimated mean functions, $\widetilde{\mu}_d$ for $d = 1, 2$, which are obtained using the pooled data in each group. Because of the identifiability constraint, the estimated $\widetilde{\mu}(\cdot)$ can be viewed as a smooth estimate of the overall mean function $\mu(\cdot)$. We assume that the overall mean function is estimated

well enough (Kulasekera, 1995); this is the case when kernel or spline smoothing techniques (Nadaraya, 1964; Watson, 1964, Fan and Gijbels, 1996; Ruppert, 1997; etc.) are used to estimate the group mean functions and the sample sizes are sufficiently large. Then $\widetilde{Y}_{idj}$ can be modeled similarly to (9), but without $\mu(\cdot)$. Thus, we assume that $\mu \equiv 0$ and that the null hypothesis is $\mu_1 \equiv 0$. The pseudo LRT methodology differs according to the sampling design. Here we focus on the setting of sparse sampled curves; the Web Supplement details the methods for the dense sampled curves.

Assume that the sampling design is irregular and sparse. As pointed out in Crainiceanu *et al.* (2012), taking pairwise differences is no longer realistic. Nevertheless, we assume that $\mu_1(t)$ can be approximated by pth degree truncated polynomials: $\mu_1(t) = x_t\boldsymbol{\beta} + z_t\mathbf{b}$. Let $\mathbf{X}_{id}$ denote the $m_{id} \times (p+1)$ dimensional matrix with the jth row equal to $x_{t_{idj}}$, and let $\widetilde{\mathbf{X}}_i = [\mathbf{X}_{i1}^T \mid -\mathbf{X}_{i2}^T]^T$, and analogously define the $m_{id} \times K$ matrices $\mathbf{Z}_{id}$'s for $d = 1, 2$ and construct $\widetilde{\mathbf{Z}}_i = [\mathbf{Z}_{i1}^T \mid -\mathbf{Z}_{i2}^T]^T$ respectively.

Denote by $\widetilde{\mathbf{Y}}_i$ the m_i-dimensional vector obtained by stacking first $\widetilde{Y}_{i1j}$'s over $j = 1, \ldots, m_{i1}$, and then $\widetilde{Y}_{i2j}$'s over $j = 1, \ldots, m_{i2}$, where $m_i = m_{i1} + m_{i2}$. It follows that, the $m_i \times m_i$-dimensional covariance matrix of $\widetilde{\mathbf{Y}}_i$, denoted by $\boldsymbol{\Sigma}_i$ can be partitioned as

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{i,11} & \boldsymbol{\Sigma}_{i,12} \\ \boldsymbol{\Sigma}_{i,21} & \boldsymbol{\Sigma}_{i,22} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{\Sigma}_{i,dd}$ is $m_{id} \times m_{id}$-dimensional matrix with the (j, j') element equal to $\Gamma_1(t_{idj}, t_{idj'}) + \Gamma_2(t_{idj}, t_{idj'}) + \sigma_\epsilon^2 1(j = j')$, and $\boldsymbol{\Sigma}_{i,dd'}$ is $m_{id} \times m_{id'}$-dimensional matrix with the (j, j') element equal to $\Gamma_1(t_{i1j}, t_{i2j'})$ for $d, d' = 1, 2$, $d \neq d'$. We can rewrite the model $\widetilde{\mathbf{Y}}_i$ using a LMM framework as $\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{X}}_i\boldsymbol{\beta} + \widetilde{\mathbf{Z}}_i\mathbf{b} + \mathbf{e}_i$, where $\mathbf{e}_i$ is m_i-dimensional vector, independent, with mean zero, and covariance matrix given by $\boldsymbol{\Sigma}_i$ described above. The hypothesis $\mu_1 \equiv 0$ can be tested as in Section 3.2. The required covariance estimators $\widehat{\boldsymbol{\Sigma}}_i$ are obtained by replacing $\boldsymbol{\Sigma}_{i,dd'}$ with $\widehat{\boldsymbol{\Sigma}}_{i,dd'}$ respectively for $1 \leq d, d' \leq 2$, which in turn are based on estimators of eigenfunctions, eigenvalues at each of the two levels, and the noise variance. For example, Di, Crainiceanu and Jank (2011) developed estimation methods for Γ_1, Γ_2 and σ_ϵ^2, $\{\sigma_{1,k}^2, \theta_{1,k}(t)\}_k$, and $\{\sigma_{2,l}^2, \theta_{2,l}(t)\}_l$. The next proposition presents conditions for these estimators, under which the assumption (C3), of Proposition 2.1 holds. Thus, under the additional assumptions (C1) and (C3) the asymptotic null distribution of pLRT statistic is given by (5).

PROPOSITION 5.1. *Assume the following conditions for model (9) hold:*

(M1') The number of measurements per subject per visit is finite, i.e., $\sup_i m_{id} < \infty$ *for* $d =$

$1, 2$. *Furthermore it is assumed that, for each subject* i*, the corresponding observation points* $\{t_{idj} : j = 1, \ldots, m_{id}\}$ *are generated uniformly and without replacement from a set* $\{t_1, \ldots, t_m\}$*, where* $t_k = (k - 1/2)/m$*, for* $k = 1, \ldots, m$ *and* m *diverges with* n

(M2') *If* $\widehat{\theta}_{d,k}(t)$*,* $\widehat{\sigma}^2_{d,k}$*, and* $\widehat{\sigma}^2_\epsilon$ *denote the estimators of the group-specific eigenfunctions, eigenvalues, and of the noise variance correspondingly, then* $\sup_{t\in\mathcal{T}} |\widehat{\theta}_{1,k}(t) - \theta_{1,k}(t)| = O_p(n^{-\alpha})$*,* $\widehat{\sigma}^2_{1,k} - \sigma^2_{1,k} = O_p(n^{-\alpha})$*,* $\sup_{t\in\mathcal{T}} |\widehat{\theta}_{2,l}(t) - \theta_{2,l}(t)| = O_p(n^{-\alpha})$*,* $\widehat{\sigma}^2_{2,l} - \sigma^2_{2,l} = O_p(n^{-\alpha})$*, for all* k*,* l*, and* $\widehat{\sigma}^2_\epsilon - \sigma^2_\epsilon = O_p(n^{-\alpha})$*.*

(M3') *We have* $m \sim n^\delta$ *where* $0 < \delta < 2\alpha$*.*

Then condition (C2) holds for the estimator $\widehat{\mathbf{\Sigma}} = \text{diag}\{\widehat{\mathbf{\Sigma}}_1, \ldots, \widehat{\mathbf{\Sigma}}_n\}$ *of* $\mathbf{\Sigma}$*, whose* i*th block* $\mathbf{\Sigma}_i$ *is defined in (10).*

The main idea in proving this result is to use the close form expression of the inverse of a partition matrix. The result can be derived using similar techniques as in the proof of the Proposition 3.3, but they involve more tedious algebra. Conditions (M1')–(M3') are analogous to (F1')–(F3') and are concerned with the sampling design, the regularity of the true covariance functions, and the accuracy of the different covariance components estimation. We conclude that the sampling design assumptions can be relaxed at the cost of accurate estimation of the level 1 covariance function, Γ_1.

6 The Sleep Heart Health Study

The Sleep Heart Health Study (SHHS) is a large-scale comprehensive multi-site study of sleep and its impacts on health outcomes. Detailed descriptions of this study can be found in Quan et al. (1997), Crainiceanu et al. (2009), and Di et al. (2009). The principal goal of the study is to evaluate the association between sleep measures and cardiovascular and non-cardiovascular health outcomes. In this paper, we focus on comparing the brain activity as measured by sleep electroencephalograms (EEG) between subjects with and without sleep-disordered breathing (SDB). The SHHS collected in-home polysomnogram (PSG) data on thousands of subjects at two visits. For each subject and visit, two-channel Electroencephalograph (EEG) data were recorded at a frequency of 125Hz (125 observations/second). Here we focus on a particular characteristic of the spectrum of the EEG data, the proportion of δ-power, which

is a summary measure of the spectral representation of the EEG signal. For more details on its definition and interpretations, see Borbely and Achermann (1999), Crainiceanu *et al.* (2009) and Di *et al.* (2009). In our study we use percent δ-power calculated in 30-second intervals. Figure 3 shows the sleep EEG percent δ-power in adjacent 30-second intervals for the first 4 hours after sleep onset, corresponding to 3 matched pairs of subjects; missing observations indicate wake periods. In each panel the percent δ-power is depicted in black lines for the SDB subjects and in gray lines for the corresponding matched controls.

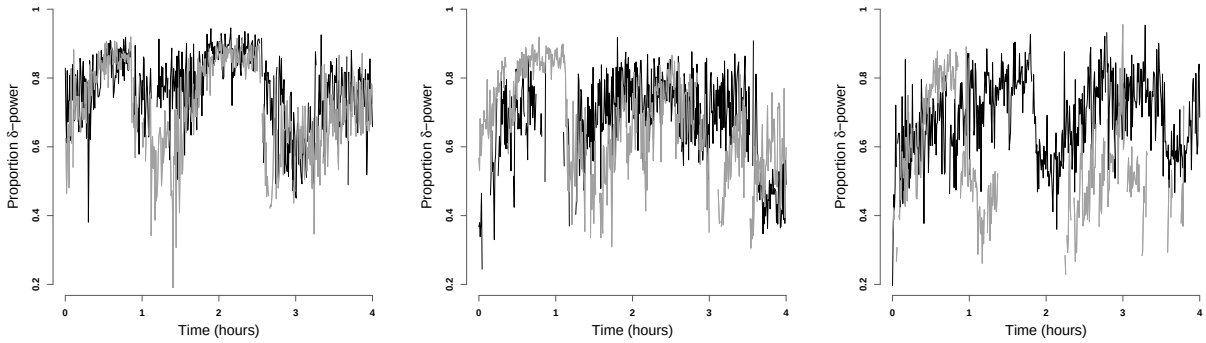

Figure 3: Sleep EEG percent δ power for the first 4 hours after sleep onset, corresponding to 3 matched pairs of controls (gray) and SDB (black).

Our interest is to compare the proportion of δ-power between the severe SDB subjects and healthy individuals, i.e., subjects without SDB, while controlling for various demographic factors. Subjects with severe SDB are identified as those with respiratory disturbance index (RDI) greater than 30 events/hour, while subjects without SDB are identified as those with an RDI smaller than 5 events/hour. Propensity score matching (Swihart, *et al.* 2012) was used to balance the groups and minimize confounding. SDB subjects were matched with no-SDB subjects on age, BMI, race, and sex to obtain a total of 51 matched pairs. In this study missing data patterns are subject-specific, with the proportion of missingness varying dramatically across subjects. Thus, simply taking the within-group differences would be inefficient. We use pseudo LRT for dependent samples of sparse functional data, as described in Section 5.1, to test for the equality of the proportion of δ-power in the two groups.

To be specific, let $\{Y_{iA}(t), Y_{iC}(t)\}$ be the proportion of δ-power measured at the tth 30 seconds interval from sleep onset, where $t = t_1, \ldots, t_T = t_{480}$, for the ith pair of matched subjects, where A refers to the SDB and C refers to the control. For each subject some of the

observations might be missing. Following Crainiceanu, et al. (2012), we model each set of curves $Y_{id}(t)$ by (9) for $d = A, C$. We are interested in testing the hypothesis H_0: $\mu_{AC} \equiv 0$, where $\mu_{AC}(t) = \mu_A(t) - \mu_C(t)$ is the difference mean function. As a preliminary step we obtain initial estimators of the group mean functions, for each of the SDB and control groups, say $\widetilde{\mu}_A(t)$ and $\widetilde{\mu}_C(t)$. We use penalized spline smoothing of all pairs $\{t, Y_{id}(t)\}$. Pseudo-residuals are calculated as $\widetilde{Y}_{id}(t) = Y_{id}(t) - \{\widetilde{\mu}_A(t) + \widetilde{\mu}_C(t)\}/2$. It is assumed that $\widetilde{Y}_{id}(t)$ can be modeled as (9), where the mean functions are $\mu_{AC}(t)$, for $d = A$ and is $-\mu_{AC}(t)$ for $d = C$ respectively. Linear splines with $K = 35$ knots are used to model the difference mean function, μ_{AC}. Pseudo LRT is applied to the pseudo-residuals, with an estimated covariance $\widehat{\mathbf{\Sigma}}$ based on the methods in Di, et al (2011).

The pseudo LRT statistic for the null hypothesis that $\mu_{AC} \equiv 0$ is 27.74, which corresponds to a p-value $< 10^{-5}$. This indicates strong evidence against the null hypothesis of no differences between the proportion δ-power in the SDB and control group. We also tested the null hypothesis on a constant difference, that is, $\mu_{AC} \equiv a$ for some constant a; the pseudo LRT statistic is 25.63 with a p-value nearly 0. Thus, there is strong evidence that the two mean functions differ by more than a constant shift. Using a pointwise confidence intervals approach, Crainiceanu, *et al.* (2012) found that differences between the apneic and control group were not significant, indicating that their local test is less powerful than pseudo LRT when testing for global differences. The global pseudo LRT does find strong evidence against the null of no difference, but cannot pinpoint where these differences are located. We suggest using the pseudo LRT introduced in this paper to test for difference and, if differences are significant by the pseudo LRT, then locating them with the methods described in Crainiceanu, *et al.* (2012). This combination of methods allows a more nuanced analysis and either method alone could provide.

7 Discussion

This paper develops a pseudo LRT procedure for testing the structure of the mean function and derives its asymptotic distribution when data exhibit complex correlation structure. In simulations pseudo LRT maintained its nominal level very well when a smooth estimator of the covariance was used and exhibited excellent power performance. Pseudo LRT was applied to test for the equality of mean curves in the context of two dependent or independent samples of curves. The close relation between the LRT and restricted LRT (RLRT) seems to imply that one should expect similar theoretical properties of the pseudo RLTR, obtained

by substituting the true covariance by a consistent estimator, when data exhibit complex correlation structure. Recent empirical investigation by Wiencierz, Greven, and Küchenhoff (2011) shows promising results.

8 Supplementary material

Supplementary material available online at Scandinavian Journal of Statistics includes 1) proofs of all theoretical results and 2) details of the hypothesis testing for the mean difference in two independent samples of curves, as well as two dependent samples of densely sampled curves. The code developed for the simulations can be found at

`http://www4.stat.ncsu.edu/\~staicu/software/pLRT_Rcode.zip`

References

Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Comput. Statist. and Data Analysis* **51**, 4793 – 4813.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Borbely, A. A. and Achermann, P. (1999). Sleep homeostasis and models of sleep regulation. *J. Biological Rhythms* **14**, 557–568.

Bosq, D. (2000) *Linear Processes in Function Spaces: Theory and Applications*. Lectures notes in statistics, Springer Verlag

Cai, T., Liu, W. and Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc* **494**, 594–607.

Chen, Y. and Liang, K. Y. (2010). On the asymptotic behavior of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* **97**, 603–602.

Crainiceanu, C. M. (2003). Nonparametric Likelihood Ratio Testing. *PhD Dissertation.* Cornell University.

Crainiceanu, C. M., Caffo, B. , Di, C., and Punjabi, N. M. (2009). Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *J. Amer. Statist. Assoc.* **104**, 541–555.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood Ratio Tests in Linear Mixed Models with One Variance Component. *Jour. Royal. Statist. Series B* **66**, 165–185.

Crainiceanu, C. M., Ruppert, D., Claeskens, G. and Wand, M. P. (2005). Exact Likelihood Ratio Tests for Penalized Splines. *Biometrika* **92**, 91–103.

Crainiceanu, C. M., Staicu, A.-M., Ray, S., and Punjabi, N. M. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes, *Statistics in Medicine*, to appear.

Cuevas, A. , Febrero, M. and Fraiman, R. (2004). An anova test for functional data. *Comput. Statist. and Data Analysis* **47**, 111–122.

Di, C., Crainiceanu, C. M., Caffo, B. and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Statist.* **3**, 458–488.

Di, C., Crainiceanu, C. M. and Jank, W. (2011). Multilevel sparse functional principal component analysis. *Under review.*

Diggle, P., Heagerty, P.J., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd Ed. Oxford University Press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications* Chapman and Hall, London.

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

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *J. Amer. Statist. Assoc.* **103**, 1520–1533.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Ann. Statist.* **29**, 153–193.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix Computations.* The Johns Hopkins University Press.

Harville, D. A. (1997). *Matrix Algebra From A Statistician's Perspective.* Springer, New York.

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

Hall, P., Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Statist. Methodol.* **68** 109–126.

James, G., Hastie, T. G., and Sugar, C. A. (2001). Principal Component Models for Sparse Functional Data. *Biometrika* **87**, 587–602.

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statis. Sci.* **21**, 155–166.

Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.* **37** 3–79.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *J. Amer. Statist. Assoc.* **90**, 1085–1093.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963–974.

Li, Y. and Hsing, T. (2010). Uniform Convergence Rates for Nonparametric Regression and Principal Component Analysis in Functional/Longitudinal Data. *Ann. Statist.* **38**, 3321–3351.

Liang, K. Y. and Self, S. G. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. R. Statist. Soc. B.* **58**, 785–796.

Loève, M. (1945). Fonctions aléatoires de second ordre. *C. R. Acad. Sci.* 220–469.

Morris, J. S. and Carroll , R. J. (2006). Wavelet-based functional mixed models. *J. Roy. Statist. Soc. Ser. B*, **68**, 179–199.

Nadaraya, E.A. (1964). On estimating regression. *Theor. Probab. Appl.* **9**, 141–142.

Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika* **58**, 545–554.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M. and Wahl, P.

W. (1997). The sleep heart health study: Design, rationale, and methods. *Sleep* **20**, 1077–1085.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis (2nd Ed.)*. New York: Springer-Verlag.

Rice, J. A. (2004). Functional and Longitudinal Data Analysis: Perspectives on smoothing. *Statist. Sinica* **14**, 631–647.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11**, 735–757.

Ruppert, D., Wand, W. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Scheipl, F., Greven, S. and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Statist. and Data Analysis*, **52**, 3283–3299.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood and likelihood rati otests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 605–610.

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

Staicu, A.-M., Lahiri, S. N. and Carroll, R. J. (2012). Tests of significance for spatially correlated multilevel functional data. *Under review.*

Swihart, B.J., Caffo, B., Crainiceanu, C. M., and Punjabi, N.M. (2012). Modeling multilevel sleep transitional data via poisson log-linear multilevel models. *Working paper*, Johns Hopkins University.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya Ser. A* **26**, 359– 372.

Wiencierz, A., Greven, S. and Küchenhoff, H. (2011). Restricted Likelihood Ratio Testing in Linear Mixed Models with General Error Covariance Structure. *Electron. J. Statist* **5** 1718–1734.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Woodbury, M. A. (1950). Inverting modified matrices, *Memorandum Report*, Princeton University.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577–590.

Zhang, J. T. and Chen, J. W. (2007). Statistical Inference for Functional Data. *Ann. Statist.* **35**, 1052–1079.

Supplementary Material for:

'Likelihood Ratio Tests for Dependent Data with Applications to Longitudinal and Functional Data Analysis'

Ana-Maria Staicu, Yingxing Li, Ciprian M. Crainiceanu, and David Ruppert

This Web Supplement contains two sections. Section A.1 discusses the proof of Propositions 2.1, 3.1, 3.2, and 3.3. Section A.2 presents hypothesis testing for the structure of the mean difference in two independent sets of curves irrespective of their sampling design, as well as in two dependent sets of curves densely sampled and the proof of Proposition 5.1, in the setting of sparsely sampled curves.

A.1 One sample of functional data. Proofs

Proof of Proposition 2.1: To be consistent with the published literature on this topis we use $\lambda = \sigma_b^2$ hereafter. Recall the definition of the pseudo log-likelihood $\log \widehat{L}_{\widehat{Y}}(\boldsymbol{\beta}, \lambda)$ in Section 2. Solving the first order condition for $\boldsymbol{\beta}$, we get the maximum pseudo profile likelihood estimation $\widehat{\boldsymbol{\beta}}(\lambda) = (\widehat{X}^T \widehat{H}_\lambda^{-1} \widehat{X})^{-1} \widehat{X}^T \widehat{H}_\lambda^{-1} \widehat{\mathbf{Y}}$. Let $\log \widehat{L}_{\widehat{\mathbf{Y}}}(\lambda)$ be the pseudo profile likelihood when $\boldsymbol{\beta}$ is maximized out, i.e.,

$$2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(\lambda) = 2 \log \widehat{L}_{\widehat{\mathbf{Y}}}\{\widehat{\boldsymbol{\beta}}(\lambda), \lambda\} = -[\log |\widehat{H}_\lambda| + \{\widehat{\mathbf{Y}} - \widehat{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T \widehat{H}_\lambda^{-1} \{\widehat{\mathbf{Y}} - \widehat{X}\widehat{\boldsymbol{\beta}}(\lambda)\}].$$

Let $\log \widehat{L}^{0,N}$ be the maximum pseudo log-likelihood under the null hypothesis (3). Then we can decompose $pLRT_N$ into two parts, i.e.,

$$pLRT_N = 2 \sup_{\lambda \geq 0} \{\log \widehat{L}_{\widehat{\mathbf{Y}}}(\lambda) - \log \widehat{L}_{\widehat{\mathbf{Y}}}(0)\} + 2\{\log \widehat{L}_{\widehat{\mathbf{Y}}}(0) - \log \widehat{L}^{0,N}\}, \tag{A.1}$$

where the first part corresponds to testing for $\lambda = 0$ and the second part corresponds to testing for the fixed effects $\beta_\ell = 0$ for $\ell \in Q$. In what follows we take each part at a time.

First part of (A.1). We first rewrite this part in a more convenient way and then prove that it converges weakly to $LRT_\infty(\lambda')$, which is defined by (5). Recall that $\widehat{H}_\lambda = I_N + \lambda \widehat{Z}\widehat{Z}^T$.

Define $\widehat{\xi}_{k,N}$ as the kth eigenvalues of $N^{-\varrho}\widehat{Z}^T\widehat{Z}$ for $k = 1, \ldots, K$. Since $\widehat{Z}\widehat{Z}^T$ has the same nonzero eigenvalues as $\widehat{Z}^T\widehat{Z}$, we have, $\log|\widehat{H}_\lambda| = \sum_{k=1}^K \log(1 + \lambda N^{\varrho}\widehat{\xi}_{k,N})$.

According to Patterson and Thompson (1971), there exists an $N \times (N - p - 1)$ matrix $\widehat{W}$ such that $\widehat{W}\widehat{W}^T = I_N - \widehat{X}(\widehat{X}^T\widehat{X})^{-1}\widehat{X}^T$, $\widehat{W}^T\widehat{W} = I_{N-p-1}$. We have $-2\log\widehat{L}_{\widehat{\mathbf{Y}}}(0) = \widehat{\mathbf{Y}}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{Y}}$. By an application of the Woodbury matrix identity (Woodbury, 1950); see also Harville (1997, p. 424) we can write

$$\begin{aligned} 2\log\widehat{L}_{\widehat{\mathbf{Y}}}(\lambda) - 2\log\widehat{L}_{\widehat{\mathbf{Y}}}(0) &= \lambda\widehat{\mathbf{Y}}^T\widehat{W}\widehat{W}^T\widehat{Z}(I_K + \lambda\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{Z})^{-1}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{Y}} \\ &\quad - \sum_{k=1}^K \log(1 + \lambda N^{-\varrho}\widehat{\xi}_{k,N}). \end{aligned} \tag{A.2}$$

Define $\widehat{\zeta}_{k,N}$ as the kth eigenvalue of $N^{-\varrho}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{Z}$ and let $\widehat{U}_{\widehat{Z}\widehat{W}}$ be the $K \times K$ matrix whose kth column is the eigenvector associated with $\widehat{\zeta}_{k,N}$. Note that $\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{Z} = \widehat{U}_{\widehat{Z}\widehat{W}}\mathrm{diag}(N^{\varrho}\widehat{\zeta}_{1,N}, \ldots, N^{\varrho}\widehat{\zeta}_{K,N})\widehat{U}^T_{\widehat{Z}\widehat{W}}$ and thus

$$2\log\widehat{L}_{\widehat{\mathbf{Y}}}(\lambda) - 2\log\widehat{L}_{\widehat{\mathbf{Y}}}(0) = \sum_{k=1}^K \frac{\lambda N^{\varrho}\widehat{w}^2_{k,N}}{1 + \lambda N^{\varrho}\widehat{\zeta}_{k,N}} - \sum_{k=1}^K \log(1 + \lambda N^{\varrho}\widehat{\xi}_{k,N}), \tag{A.3}$$

where $\widehat{w}_{k,N}$ is the kth component of the column vector $\widehat{\mathbf{w}}_N = N^{-\varrho/2}\widehat{U}^T_{\widehat{Z}\widehat{W}}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{Y}}$.

For simplicity of exposition we define

$$\widehat{f}_N(\lambda') = \sum_{k=1}^K \frac{\lambda'\widehat{w}^2_{k,N}}{1 + \lambda'\widehat{\zeta}_{k,N}} - \sum_{k=1}^K \log(1 + \lambda'\widehat{\xi}_{k,N}) \tag{A.4}$$

and write $2\sup_{\lambda\geq 0}\{\log\widehat{L}_{\widehat{\mathbf{Y}}}(\lambda) - \log\widehat{L}_{\widehat{\mathbf{Y}}}(0)\} = \sup_{\lambda'\geq 0}\widehat{f}_N(\lambda')$. We will show that $\sup_{\lambda'\geq 0}\widehat{f}_N(\lambda') \Rightarrow \sup_{\lambda'\geq 0} LRT_\infty(\lambda')$ in two steps:

(S1) $\widehat{f}_N(\lambda')$ converges weakly to $LRT_\infty(\lambda')$ on the space of $C[0, M]$, for $M < \infty$;

(S2) a continuous mapping theorem type result holds for $\sup_{\lambda'\geq 0}\widehat{f}_N(\lambda')$.

We show (S1) in two parts: 1) first prove that $\widehat{f}_N(\lambda') \xrightarrow{D} LRT_\infty(\lambda')$ and 2) then show that $\widehat{f}_N(\lambda')$ is a tight sequence (Billingsley 1968, p54). The definition of $LRT_\infty(\lambda')$ is similarly to that of $\widehat{f}_N(\lambda')$ except that $\widehat{\zeta}_{k,N}$'s, $\widehat{\xi}_{k,N}$'s and $\widehat{w}_{k,N}$'s are replaced by ζ_k's, ξ_k's and w_k's. For the first part, it is sufficient to prove that $\widehat{w}_{k,N} \xrightarrow{D} w_k$, $\widehat{\zeta}_{k,N} \xrightarrow{P} \zeta_k$, and $\widehat{\xi}_{k,N} \xrightarrow{P} \xi_k$ for all k; Lemma A.1.1, discusses these results next. The weak convergence of $\widehat{f}_N(\lambda')$ to $LRT_\infty(\lambda')$, or in general, the finite dimensional convergence $\{\widehat{f}_N(\lambda'_1), \ldots, \widehat{f}_N(\lambda'_L)\}$ to $\{LRT_\infty(\lambda'_1), \ldots, LRT_\infty(\lambda'_L)\}$ follows by an application of the continuous mapping theorem.

LEMMA A.1.1. *Let $\widehat{\xi}_{k,N}$, $\widehat{\zeta}_{k,N}$, and $\widehat{w}_N$ be defined as above. Assume conditions (C2) and (C3) of Proposition 2.1 are true. Then:*
(a) For each $k = 1, \ldots, K$, as $N \to \infty$ we have $\widehat{\xi}_{k,N} \xrightarrow{P} \xi_k$, $\widehat{\zeta}_{k,N} \xrightarrow{P} \zeta_k$.
(b) If in addition condition (C1) is true, then $\widehat{w}_N \xrightarrow{D} w$, where $\mathbf{w} = (w_1, \ldots, w_K)$ is defined by Proposition 2.1.

Proof: To show the results of Lemma A.1.1 we need to introduce additional notation. Let $\tilde{\mathbf{Y}}$, $\tilde{X}$, $\tilde{Z}$ and $\tilde{W}$ be defined similarly to $\widehat{\mathbf{Y}}$ $\widehat{X}$, $\widehat{Z}$ and $\widehat{W}$ but with the $\widehat{\Sigma}$ replaced by Σ, and similarly define $\tilde{\xi}_{k,N}$, $\tilde{\zeta}_{k,N}$, and $\tilde{\mathbf{w}}_N$ corresponding to $\widehat{\xi}_{k,N}$, $\widehat{\zeta}_{k,N}$, and $\widehat{\mathbf{w}}_N$; for example $\tilde{\xi}_{k,N}$ and $\tilde{\zeta}_{k,N}$ are exactly the quantities introduced by the Proposition 2.1 with the same notation.
(a) We will show that $\widehat{\xi}_k - \tilde{\xi}_k = o_p(1)$ and $\widehat{\zeta}_k - \tilde{\zeta}_k = o_p(1)$; then the result (a) follows by applying Slutsky's theorem and assuming the condition (C3). By using Theorem 8.1-6 of Golub and Van Loan (1983) it is sufficient to prove that

$$||N^{-\varrho}\widehat{Z}^T\widehat{Z} - N^{-\varrho}\tilde{Z}^T\tilde{Z}|| = o_p(1) \text{ and} \tag{A.5}$$

$$||N^{-\varrho}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{Z} - N^{-\varrho}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Z}|| = o_p(1), \tag{A.6}$$

where $||A|| = \sqrt{\sum_i \sum_j a_{ij}^2}$ is the Frobenius norm of some matrix $A = (a_{ij})_{i,j}$. These results follow from employing condition (C2), namely that $\mathbf{a}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{a} - \mathbf{a}^T\mathbf{\Sigma}^{-1}\mathbf{a} = o_p(1)$ for non-random unit vector a, and from applications of norm inequalities as well as continuous mapping theorem.
(b) Next, we prove the convergence in distribution of $\widehat{\mathbf{w}}_N$. The idea is first to show that $\tilde{\mathbf{w}}_N \xrightarrow{D} \mathbf{w}$ under the null hypothesis and then to show that $\|\widehat{\mathbf{w}}_N - \tilde{\mathbf{w}}_N\| = o_p(1)$.

Under the null hypothesis, we have $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\mathbf{e}}$ and thus $\tilde{\mathbf{w}}_N = N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{\mathbf{e}}$, since $\tilde{W}^T\tilde{\mathbf{Y}} = \tilde{W}^T\tilde{\mathbf{e}}$. Because $\tilde{\mathbf{e}} = \Sigma^{-1/2}\mathbf{e}$ it follows that $\tilde{\mathbf{e}} \sim N(\mathbf{0}_{N\times 1}, I_N)$ and thus $\tilde{\mathbf{w}}_N$ has mean-zero multivariate normal distribution, since it is a linear combination of independent normal variables. The result, $\tilde{\mathbf{w}}_N \xrightarrow{D} \mathbf{w}$, is concluded by assuming condition (C3), using Cramér-Wold device and an application of the (Lévy's) continuity theorem.

We prove next that $\|\widehat{\mathbf{w}}_N - \tilde{\mathbf{w}}_N\| = o_p(1)$. Recall that $\widehat{\mathbf{w}}_N - \tilde{\mathbf{w}}_N = N^{-\varrho/2}\widehat{U}^T_{\widehat{Z}\widehat{W}}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{e}} - N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{\mathbf{e}}$ and thus we have:

$$\begin{aligned} ||\widehat{\mathbf{w}}_N - \tilde{\mathbf{w}}_N|| \;\leq\; & ||\widehat{U}^T_{\widehat{Z}\widehat{W}} - \tilde{U}^T_{\tilde{Z}\tilde{W}}||||N^{-\varrho/2}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{e}}|| \\ & + ||\tilde{U}^T_{\tilde{Z}\tilde{W}}||||N^{-\varrho/2}(\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{e}} - \tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{\mathbf{e}})|| \end{aligned} \tag{A.7}$$

Using norm, matrix manipulation, and furthermore employing condition (C2), namely that $\mathbf{a}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{e} - \mathbf{a}^T\mathbf{\Sigma}^{-1}\mathbf{e} = o_p(1)$ for non-random unit vector a, one can show that $\|\tilde{U}_{\tilde{Z}\tilde{W}}\| = O_p(1)$, $||\widehat{U}_{\widehat{Z}\widehat{W}} - \tilde{U}_{\tilde{Z}\tilde{W}}|| = o_p(1)$, $||N^{-\varrho/2}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{e}}|| = O_p(1)$, and $||N^{-\varrho/2}\widehat{Z}^T\widehat{W}\widehat{W}^T\widehat{\mathbf{e}} - N^{-\varrho/2}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{\mathbf{e}}|| = o_p(1)$.This concludes the proof of Lemma A.1.1. #

Next, we prove the second part of (S1). Using Theorem 8.3 of Billingsley (1968), it suffices to show that for every ε' and $\eta' > 0$, there exists $\delta_0 > 0$ and N_0 such that for $N \geq N_0$,

$$\frac{1}{\delta_0}P\{\sup_{t\leq t'\leq t+\delta_0}|\widehat{f}_N(t') - \widehat{f}_N(t)| \geq \varepsilon'\} \leq \eta'. \tag{A.8}$$

It is noteworthy to point out that for every $\delta > 0$, and every $0 \leq t \leq t' \leq t + \delta$ we have

$$|\widehat{f}_N(t) - \widehat{f}_N(t')| \;\leq\; \sum_{k=1}^{K}|t-t|\widehat{w}_{k,N}^2 + \sum_{k=1}^{K}\log\left\{1 + \frac{(t'-t)\widehat{\xi}_{k,N}}{1+t\widehat{\xi}_{k,N}}\right\} \leq \sum_{k=1}^{K}\delta\widehat{w}_{k,N}^2 + \sum_{k=1}^{K}\delta\widehat{\xi}_{k,N},$$

since $\widehat{\xi}_{k,N}$'s, $\widehat{\zeta}_{k,N}$'s are nonnegative, it holds true $\log(1+x) < x$ for $x > 0$.
Let ε' and η' be arbitrary but fixed positive values. Then for every $\delta > 0$ we have

$$P\left\{\sup_{t\leq t'\leq t+\delta}|f_N(t') - f_N(t)| \geq \varepsilon'\right\} \;\leq\; \sum_{k=1}^{K}P\{\widehat{w}_{k,N}^2 \geq \varepsilon'/(2K\delta)\} + \sum_{k=1}^{K}P\{\widehat{\xi}_{k,N} \geq \varepsilon'/(2K\delta)\} \tag{A.9}$$

which follows from an application of the Bonferroni inequality, $P(\sum_{i=1}^{2K}A_i \geq a) \leq \sum_{i=1}^{2K}P\{A_i \geq a/(2K)\}$, along with the observation that if $\sum_{i=1}^{2K}A_i \geq a$ holds, for variables $\{A_i : i = 1,\ldots,2K\}$ then we must have that $A_i \geq a/(2K)$, for some i. It is sufficient to show that there exists $\delta_0 = \delta_0(\varepsilon',\eta') > 0$ and $N_0 = N_0(\varepsilon',\eta') \geq 1$ such that the the right hand expression of the above inequality, with δ replaced by δ_0, is bounded by $\delta_0\eta'$ for all $N \geq N_0$.
For the first term of (A.9) let $F_{k,N}(t)$ and $F_k(t)$ be the cumulative distribution functions of $\widehat{w}_{k,N}$ and w_k respectively, for $k = 1,\ldots,K$. Because $\widehat{\mathbf{w}}_{k,N} \Rightarrow \mathbf{w}_k$, it follows that $|F_{k,N}(t) - F_k(t)| \to 0$ for all t. Then it is not hard to show that for every $\delta < \widetilde{\delta}^0$ there is $N*_\delta > 0$ such that $\sum_{k=1}^{K}P\{\widehat{w}_{k,N}^2 \geq \varepsilon'/(2K\delta)\} < \delta\eta'/2$ for all $N > N_\delta^*$.
Consider now the second sum of (A.9). For every summand k we have:

$$P\{\widehat{\xi}_{k,N} \geq \varepsilon'/(2K\delta)\} \;\leq\; P\{\widehat{\xi}_{k,N} > \varepsilon'/(2K\delta),\ |\widehat{\xi}_{k,N} - \xi_k| \leq \varepsilon'\} + P(|\widehat{\xi}_{k,N} - \xi_k| > \varepsilon')$$

which is less or equal than $P(|\widehat{\xi}_{k,N} - \xi_k| > \varepsilon')$, for any $\delta < \varepsilon'/\{2K(\xi_k + \varepsilon')\}$; since for this choice $|\widehat{\xi}_{k,N} - \xi_k| > \varepsilon'$ and the first term equals zero. Because $\widehat{\xi}_{k,N} \Rightarrow \xi_k$ and ξ_k is constant, we have that $\widehat{\xi}_{k,N} \to \xi_k$ in probability. It follows that for every $\delta < \varepsilon'/[2K\{\varepsilon' + \max_k(\xi_k)\}]$ we have that $\sum_{k=1}^{K}P\{\widehat{\xi}_{k,N} \geq \varepsilon'/(2K\delta)\} \leq \delta\eta'/2$ for all $N > N_\delta^{**}$).

Combining the two findings, one can find suitable δ_0 and N_0 so that expression (A.9) holds. This concludes the tightness proof of the sequence $\widehat{f}_N(\lambda')$.

To show (S2) we apply the continuous mapping theorem as in Crainiceanu (2003) and use weak convergence result of (S1) for all $M < \infty$; to save space we omit the details.

Second part of (A.1). Next we will show that there exists independent standard normal variables $\nu_1, \ldots, \nu_{p-q+1}$ such that

$$2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(0) - 2 \log \widehat{L}^{0,N} \Rightarrow \sum_{i=1}^{p-q+1} \nu_i^2, \tag{A.10}$$

We discuss the case when $p - q + 1 > 0$; if $p - q + 1 = 0$ then (A.10) is trivial.

Before we simplify the left hand side of equation (A.10), we introduce the following definition. Partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T | \boldsymbol{\beta}_{(2)}^T)^T$, where $\boldsymbol{\beta}_{(2)}$ contains all β_ℓ for $\ell \in Q$. Similarly, partition $X = (X_{(1)}|X_{(2)})$ according to the partition of $\boldsymbol{\beta}$. We define $\widehat{X}_{(i)} = \widehat{\Sigma}^{-1/2} X_{(i)}$. For any matrix A with linearly independent columns, denote by $S_A = A(A^T A)^{-1} A^T$ the projection matrix onto the space spanned by the columns of A. In the special case when $\#Q = p + 1$, $X_{(2)} = X$ and $X_{(1)}$ does not exist, we use the convention that $S_{\widehat{X}_{(1)}} = \mathbf{0}_{N \times N}$.

Under the null hypothesis we have that $\widehat{\mathbf{Y}} = \widehat{X}_{(1)} \boldsymbol{\beta}_{(1)} + \widehat{\mathbf{e}}$. Then $2 \log \widehat{L}^{0,N} = -\widehat{\mathbf{Y}}^T (I_N - S_{\widehat{X}_{(1)}}) \widehat{\mathbf{Y}} = -\widehat{\mathbf{e}}^T (I_N - S_{\widehat{X}_{(1)}}) \widehat{\mathbf{e}}$, and $2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(0) = -\widehat{\mathbf{e}}^T (I_N - S_{\widehat{X}}) \widehat{\mathbf{e}}$. It follows $2 \log \widehat{L}_{\widehat{\mathbf{Y}}}(0) - 2 \log \widehat{L}^{0,N} = \widehat{\mathbf{e}}^T (S_{\widehat{X}} - S_{\widehat{X}_{(1)}}) \widehat{\mathbf{e}}$. where $S_{\widehat{X}} - S_{\widehat{X}_{(1)}}$ is a projection matrix. There exists a $N \times (p - q + 1)$ matrix such that $\widehat{W}_0 \widehat{W}_0^T = S_{\widehat{X}} - S_{\widehat{X}_{(1)}}$ and $\widehat{W}_0^T \widehat{W}_0 = I_{p-q+1}$. Denote $\widehat{\boldsymbol{\varpi}} = \widehat{W}_0^T \widehat{\mathbf{e}}$. Applying the same arguments as for the Lemma A.1.1, part (b) we can conclude that $\widehat{\varpi}_i$'s are asymptotically standard normal independent variables assuming (C1), that the null hypothesis is true. It implies that equation (A.10) holds, and in turn that (5) holds.

The proof is now concluded, as independence between $\sup LRT_\infty(\lambda)$ and $\sum_{i=1}^{p-q+1} \nu_i^2$ can be established using the same techniques as in the proof of Theorem 3 of Crainiceanu (2003). Hence Proposition 2.1 holds. #

Proof of Proposition 3.1: First the following regularity conditions are imposed:

(A1) The true parameter of the correlation structure $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ lies in the interior of a compact set.

(A2) For any $\boldsymbol{\theta} \in \Theta$, the functions $\partial \varphi(t, t'; \boldsymbol{\theta}) / \partial \theta_\ell$, and $\partial^2 \varphi(t, t'; \boldsymbol{\theta}) / \partial \theta_\ell \partial \theta'_\ell$ are bounded bivariate functions of t, t'.

(A3) For any $\boldsymbol{\theta} \in \Theta$, the eigenvalues of the correlation matrix $C_i(\boldsymbol{\theta})$ of a generic subject i are between $0 < \varrho_0 < \varrho_1 < \infty$.

Let $C(\boldsymbol{\theta}) = \text{diag}\{C_1(\boldsymbol{\theta}) \ldots, C_n(\boldsymbol{\theta})\}$ be the true correlation matrix, and denote by $C(\widehat{\boldsymbol{\theta}})$ its estimate. Condition (C2) entails two parts: (a) $\widehat{\sigma}_e^{-2}\mathbf{a}^T C(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{a} - \sigma_e^{-2}\mathbf{a}^T C(\boldsymbol{\theta})^{-1}\mathbf{a} = o_p(1)$, and (b) $\widehat{\sigma}_e^{-2}\mathbf{a}^T C(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{e} - \sigma_e^{-2}\mathbf{a}^T C(\boldsymbol{\theta})^{-1}\mathbf{e} = o_p(1)$ where $\mathbf{a} \in R^N$, $\|\mathbf{a}\| = 1$, $\mathbf{e}$ is the N-dimensional vector of $\mathbf{e}_i$ and $N = \sum_{i=1}^n m_i$.

To prove part (a) it is sufficient to show that: (a1) $\mathbf{a}^T C^{-1}(\boldsymbol{\theta})\mathbf{a} = O(1)$ and (a2) $\mathbf{a}^T\{C^{-1}(\boldsymbol{\theta}) - C^{-1}(\widehat{\boldsymbol{\theta}})\}\mathbf{a} = o_p(1)$. The result then follows by using $\widehat{\sigma}_e^2 - \sigma_e^2 = o_p(1)$, the observation $\widehat{\sigma}_e^{-2} = O_p(1)$ and an application of the triangle inequality. Showing (a1) is easy by using the regularity condition (A3). Consider now (a2).Let $\varphi_{\boldsymbol{\theta}}(t, t'; \boldsymbol{\theta}) = \partial\varphi(t, t'; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, for some time points t, t' and let $\partial C_i(\boldsymbol{\theta})/\partial\theta_l$ be the matrix with components $\varphi_{\theta_l}(t_{ij}, t_{ij'}; \boldsymbol{\theta})$ for $j, j' = 1, \ldots, m_i$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$. Fix $l = 1, \ldots, d$, and denote by $D_{i,l}(\boldsymbol{\theta}) = C_i^{-1}(\boldsymbol{\theta})\{\partial C_i(\boldsymbol{\theta})/\partial\theta_l\}C_i^{-1}(\boldsymbol{\theta})$, which is $\partial C_i^{-1}(\boldsymbol{\theta})/\partial\theta_l$. Using the regularity conditions (A1)-(A3) is easy to see that $\|C_i^{-1}(\widehat{\boldsymbol{\theta}}) - C_i^{-1}(\boldsymbol{\theta})\| = O_p(n^{-1/2})$, where $\|A\|$ is the Frobenius norm of matrix A, defined above. Moreover one can show that $\|C_i^{-1}(\widehat{\boldsymbol{\theta}}) - C_i^{-1}(\boldsymbol{\theta})\| \leq \sum_{l=1}^d |\widehat{\theta}_l - \theta_l|\, M_l$, for some M_l such that $\sup_{\boldsymbol{\theta}} \|D_{i,l}(\boldsymbol{\theta})\| < M_l$ for all $i = 1, \ldots, n$ and $l = 1, \ldots, d$. It follows that $\|C^{-1}(\boldsymbol{\theta}) - C^{-1}(\widehat{\boldsymbol{\theta}})\|_2 = \max_i \|C_i^{-1}(\boldsymbol{\theta}) - C_i^{-1}(\widehat{\boldsymbol{\theta}})\|_2 = o_p(1)$.

To prove part (b) it is sufficient to show that: (b1) $\mathbf{a}^T C^{-1}(\boldsymbol{\theta})\mathbf{e} = O_p(1)$ and (b2) $\mathbf{a}^T\{C^{-1}(\boldsymbol{\theta}) - C^{-1}(\widehat{\boldsymbol{\theta}})\}\mathbf{e} = o_p(1)$. To show (b1) it suffices to show that $\text{var}\{\mathbf{a}^T C^{-1}(\boldsymbol{\theta})\mathbf{e}\} = O(1)$. This is easy to check since $\|C_i^{-1}(\boldsymbol{\theta})\|_2 < \infty$ for all i and $\|\mathbf{a}\| = 1$. Consider now part (b2). Let $f_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{a}_i^T C_i^{-1}(\boldsymbol{\theta})\mathbf{e}_i$. Using the first order Taylor expansion of the function $f_{\mathbf{a},\mathbf{e}}(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$ we obtain;

$$f_{\mathbf{a},\mathbf{e}}(\widehat{\boldsymbol{\theta}}) = f_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T f'_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) + o_p(1), \tag{A.11}$$

where $f'_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) = \partial f_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. We show that the vector $n^{-1/2} f'_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) = o_p(1)$, by proving that each of its components, $n^{-1/2}\partial f_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta})/\partial\theta_l = o_p(1)$, is $o_p(1)$. Using condition (A3) we have

$$\text{var}\{\sum_{i=1}^n \mathbf{a}_i^T D_{i,l}(\boldsymbol{\theta})\mathbf{e}_i\} = \sum_{i=1}^n \sigma_e^2 \mathbf{a}_i^T D_{i,l}(\boldsymbol{\theta}) C_i(\boldsymbol{\theta}) D_{i,l}^T(\boldsymbol{\theta})\mathbf{a}_i$$

which is finite, since $m_i < \infty$ and $\|C_i(\boldsymbol{\theta})\| < \infty$. The result (b2) follows easily since $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T f'_{\mathbf{a},\mathbf{e}}(\boldsymbol{\theta}) = o_p(1)$, by employing the assumption (L3) of the proposition. #

Proof of Proposition 3.2: For illustration simplicity we assume that m satisfies assumption (F2) and thus $\widetilde{m} = m$, and $\widetilde{t}_l = t_l$. Under the dense design, $\mathbf{X}_i$ and $\mathbf{Z}_i$ do not depend on

i, and $\mathbf{\Sigma} = \mathbf{I}_n \otimes \mathbf{\Sigma}_0$, where the (j, j')th element of $\mathbf{\Sigma}_0$ is $\Gamma(t_j, t_{j'}) + \sigma_\epsilon^2 1(t_j = t_{j'})$. It suffices to show that (C2') holds for any m-dimensional vector, $\mathbf{a}$, $||\mathbf{a}|| = 1$, and $\mathbf{e}_0 = n^{-1/2} \sum_{i=1}^n \mathbf{e}_i$, i.e.

$$\|\widehat{\mathbf{\Sigma}}_0^{-1} - \mathbf{\Sigma}_0^{-1}\|_2 = o_p(1) \tag{A.12}$$

$$\mathbf{a}^T \widehat{\mathbf{\Sigma}}_0^{-1} \mathbf{e}_0 - \mathbf{a}^T \mathbf{\Sigma}_0^{-1} \mathbf{e}_0 = o_p(1). \tag{A.13}$$

Consider equation (A.12). Let Γ^0 be the $m \times m$ covariance matrix obtained from the covariance function $\Gamma(t, t')$ evaluated over the set of observed points $\{t_1, \ldots, t_m\}^2$. Furthermore denote by $\mathbf{G} = (g_{jk})_{1 \le j \le m, 1 \le k \le M}$ the $m \times M$ matrix of eigenvectors and by $\mathbf{\Lambda} = \text{diag}\{\sigma_1^2, \ldots, \sigma_M^2\}$ the $M \times M$ diagonal matrix of associated eigenvalues corresponding to Γ^0. It follows that $g_{jk} \approx \theta_k(t_j)/\sqrt{m}$ for $k = 1, \ldots, M$ and $j = 1, \ldots, m$ since $\sum_{j=1}^m \theta_k^2(t_j)/m$ is the Riemann approximation of the unit-valued integral $\int_{\mathcal{T}} \theta_k(t)^2 \, dt = 1$. In the new notation we write $\mathbf{\Sigma}_0 = \sigma_\epsilon^2 \mathbf{I}_m + m\mathbf{G}\mathbf{\Lambda}\mathbf{G}^T$ and by using the Woodbury matrix identity it follows that $\mathbf{\Sigma}_0^{-1} = \sigma_\epsilon^{-2}\mathbf{I}_m - \sigma_\epsilon^{-2}\mathbf{G}\text{diag}\{\sigma_k^2/(\sigma_k^2 + \sigma_\epsilon^2/m)\}_k \mathbf{G}^T$.

In a similar way, define $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{\Lambda}}$ the estimated quantities corresponding to $\mathbf{G}$ and $\mathbf{\Lambda}$, respectively, for $k = 1, \ldots, M$ and $\widehat{\mathbf{\Sigma}}_0 = \widehat{\sigma}_\epsilon^2 \mathbf{I}_m + m\widehat{\mathbf{G}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{G}}^T$. Following the assumption $\|\widehat{\theta}_k - \theta_k\| = O_p(n^{-\alpha})$ we have $\|\widehat{\mathbf{g}}_k - \mathbf{g}_k\| = \|\widehat{\theta}_k - \theta_k\| = O_p(n^{-\alpha})$, where $\widehat{\mathbf{g}}_k$ and $\mathbf{g}_k$ are the kth columns of $\widehat{\mathbf{G}}$ and $\mathbf{G}$ respectively.

We show (A.12) by using the Woodbury matrix identity for $\widehat{\mathbf{\Sigma}}_0^{-1}$ and $\mathbf{\Sigma}_0^{-1}$ and triangle inequality:

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}_0^{-1} - \mathbf{\Sigma}_0^{-1}\|_2 \;\le\; & \left\| \widehat{\sigma}_\epsilon^{-2}\widehat{\mathbf{G}}\text{diag}\{\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_k^2 + \widehat{\sigma}_\epsilon^2/m}\}_k \widehat{\mathbf{G}}^T - \sigma_\epsilon^{-2}\mathbf{G}\text{diag}\{\frac{\sigma_k^2}{\sigma_k^2 + \sigma_\epsilon^2/m}\}_k \mathbf{G}T \right\|_2 \\ & + |\widehat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2}|, \end{aligned}$$

since (a) $\widehat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2} = o_p(1)$, (b) $\|\widehat{\mathbf{G}}\text{diag}\{\widehat{\sigma}_k^2/(\widehat{\sigma}_k^2 + \widehat{\sigma}_\epsilon^2/m)\}_k \widehat{\mathbf{G}}^T\|_2 = \max_k\{\widehat{\sigma}_k^2/(\widehat{\sigma}_k^2 + \widehat{\sigma}_\epsilon^2/m)\}_k = \widehat{\sigma}_1^2/(\widehat{\sigma}_1^2 + \widehat{\sigma}_\epsilon^2/m) = O_p(1)$, (c) $||\widehat{\mathbf{G}} - \mathbf{G}|| = O_p(n^{-\alpha})$, (d) $||\text{diag}\{\widehat{\sigma}_k^2/(\widehat{\sigma}_k^2 + \widehat{\sigma}_\epsilon^2/m) - \sigma_k^2/(\sigma_k^2 + \sigma_\epsilon^2/m)\}_k|| = O_p(n^{-\alpha}m^{-1})$, (e) $\sum_{k=1}^M \sigma_k^2 < \infty$ as well as the unitary invariance of 2-norms and the norm relationship $\|\cdot\|_2 \le \|\cdot\|$.

We turn now to (A.13). Again we use Woodbury matrix identity and reduce (A.13) to

$$\mathbf{a}^T \left[\widehat{\mathbf{G}}\text{diag}\{\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_k^2 + \widehat{\sigma}_\epsilon^2/m}\}_k \widehat{\mathbf{G}}^T - \mathbf{G}\text{diag}\{\frac{\sigma_k^2}{\sigma_k^2 + \sigma_\epsilon^2/m}\}_k \mathbf{G}^T \right] \mathbf{e} = o_p(1),$$

under the assumption that $\widehat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2} = o_p(n^{-\alpha})$, since $\mathbf{a}^T\mathbf{e} = O_p(m^{1/2})$. Moreover, if the

number of eigenvalues M is finite, and $(1-\sigma_k^{-2}\sigma_\epsilon^2/m) = 1+O(m^{-1})$ it suffices to show that

$$\{\widehat{\sigma}_k^2/(\widehat{\sigma}_k^2+\widehat{\sigma}_\epsilon^2/m) - \sigma_k^2/(\sigma_k^2+\sigma_\epsilon^2/m)\}a^T\widehat{\mathbf{g}}_k\widehat{\mathbf{g}}_k^T\mathbf{e} = o_p(1) \tag{A.14}$$

$$a^T(\widehat{\mathbf{g}}_k - \mathbf{g}_k)\mathbf{g}_k^T\mathbf{e} = O_p(n^{-\alpha}m^{1/2}) \tag{A.15}$$

$$a^T\mathbf{g}_k(\widehat{\mathbf{g}}_k - \mathbf{g}_k)^T\mathbf{e} = O_p(n^{-\alpha}m^{1/2}). \tag{A.16}$$

Relation (A.14) follows from application of Chebyshev's inequality and the following facts: $\widehat{\sigma}_k^2/(\widehat{\sigma}_k^2+\widehat{\sigma}_\epsilon^2/m) - \sigma_k^2/(\sigma_k^2+\sigma_\epsilon^2/m) = O_p(n^{-\alpha}m^{-1})$, $\|\widehat{\mathbf{g}}_k\| = 1$, $\|a\| = 1$ and $\|\mathbf{e}\| = O_p(m^{1/2})$.

Relation(A.15) is obvious since $\mathbf{g}_k^T\mathbf{e} = O_p(m^{1/2})$ and $|a^T(\widehat{\mathbf{g}}_k - \mathbf{g}_k)| = O_p(n^{-\alpha})$. To show (A.16) it is sufficient to show $(\widehat{\mathbf{g}}_k - \mathbf{g}_k)^T\mathbf{e} = O_p(n^{-\alpha}m^{-1/2})$ which follows from an application of Chebyshev's inequality.

The result now follows using the assumption that $O_p(n^{-\alpha}m^{1/2}) = o_p(1)$. This concludes our proof. #

Proof of Proposition 3.3: Under the sparse design, $\mathbf{\Sigma}$ is a block diagonal matrix, with the ith block equal to the $m_i \times m_i$-dimensional matrix $\mathbf{\Sigma}_i = \sigma_\epsilon^2(\mathbf{I}_{m_i} + \mathbf{K}_i)$ where $\mathbf{K}_i = \mathbf{G}_i\Lambda\mathbf{G}_i^T$, $\mathbf{G}_i$ is $m_i \times M$ dimensional matrix with the (l,k)th element equal to $\theta_k(t_{il})$, and $\mathbf{\Lambda}$ is $M \times M$ block diagonal matrix whose (k,k)th component is $\sigma_k^2/\sigma_\epsilon^2$. Note that $\mathbf{G}_i$ and $\mathbf{\Lambda}$ are defined differently from the corresponding ones defined in the proof of Proposition 3.2. Similarly, we can define $\widehat{\mathbf{\Sigma}}_i = \widehat{\sigma}_\epsilon^2(\mathbf{I}_{m_i} + \widehat{\mathbf{K}}_i)$, where $\widehat{\mathbf{K}}_i = \widehat{\mathbf{G}}_i\widehat{\mathbf{\Lambda}}\widehat{\mathbf{G}}_i^T$. Partition $\mathbf{a}$ and $\mathbf{e}$ in condition (C2) into n vectors, $\mathbf{a}_i$ and $\mathbf{e}_i$ of length m_i. To prove condition (C2), it suffices to show

$$\|\text{diag}\{\widehat{\mathbf{\Sigma}}_i^{-1} - \mathbf{\Sigma}_i^{-1}\}_i\|_2 = o_p(1), \tag{A.17}$$

$$\sum_{i=1}^n \mathbf{a}_i^T\widehat{\mathbf{\Sigma}}_i^{-1}\mathbf{e}_i - \sum_{i=1}^n \mathbf{a}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{e}_i = o_p(1). \tag{A.18}$$

Consider first equation (A.17). We make use of the well known inequalities: $\|\text{diag}\{A_i\}_i\|_2 \leq \max_i \|A_i\|_2$, and $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$ for matrices $A = (a_{jj'})_{j,j'}$, A_i's, where $\|A\|_1 = \max_{j'}\sum_j |a_{jj'}|$ and $\|A\|_\infty = \max_j \sum_{j'} |a_{jj'}|$ are the 1 and ∞ matrix norm induced. The result follows by noting that

$$\|\widehat{\mathbf{\Sigma}}_i^{-1} - \mathbf{\Sigma}_i^{-1}\|_2 \leq |\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2| + \sigma_\epsilon^2\|(\mathbf{I}_{m_i} + \widehat{\mathbf{K}}_i)^{-1}\|_2 \, \|\widehat{\mathbf{K}}_i - \mathbf{K}_i\|_2 \, \|(\mathbf{I}_{m_i} + \mathbf{K}_i)^{-1}\|_2$$

and that $\max_i \|\widehat{\mathbf{K}}_i - \mathbf{K}_i\|_\infty = O_p(n^{-\alpha})$, $\max_i \|\widehat{\mathbf{K}}_i - \mathbf{K}_i\|_1 = O_p(n^{-\alpha})$. In addition $\|(\mathbf{I}_{m_i} + \widehat{\mathbf{K}}_i)^{-1}\| < 1$ and $\|(\mathbf{I}_{m_i} + \mathbf{K}_i)^{-1}\| < 1$ because both $\mathbf{K}_i$ and $\widehat{\mathbf{K}}_i$ are positive definite matrices, and thus have positive eigenvalues.

Next we prove the validity of (A.18). Using Woodbury matrix identity, we rewrite each term of the left hand side of (A.18) as

$$\begin{aligned}\sum_{i=1}^{n}\mathbf{a}_i^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{e}_i &= \widehat{\sigma}_\epsilon^2\sum_{i=1}^{n}\mathbf{a}_i^T\mathbf{e}_i - \widehat{\sigma}_\epsilon^2\sum_{i=1}^{n}\mathbf{a}_i^T\widehat{\mathbf{G}}_i(\widehat{\boldsymbol{\Lambda}}^{-1}+\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)^{-1}\widehat{\mathbf{G}}_i^T\mathbf{e}_i;\\ \sum_{i=1}^{n}\mathbf{a}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{e}_i &= \sigma_\epsilon^2\sum_{i=1}^{n}\mathbf{a}_i^T\mathbf{e}_i - \sigma_\epsilon^2\sum_{i=1}^{n}\mathbf{a}_i^T\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i,\end{aligned}$$

where $\mathbf{e}_i$ are independent m_i-dimensional vectors with distribution $N(0,\boldsymbol{\Sigma}_i)$. By an application of the continuity theorem for $S_n = \sum_{i=1}^n \mathbf{a}_i^T\mathbf{e}_i = \sum_{i=1}^n |c_i|\epsilon_i$, where $c_i^2 = \mathbf{a}_i^T\boldsymbol{\Sigma}_i\mathbf{a}_i$ and ϵ_i are independent standard normal variables we find $S_n = O_p(1)$, since $\|\boldsymbol{\Sigma}_i\| < \sigma_\epsilon^2(\max_i m_i)\{\mathrm{tr}(\boldsymbol{\Lambda})+1\}$ and $\sum_{i=1}^n\|\mathbf{a}_i\|^2 = 1$. Thus $(\widehat{\sigma}_\epsilon^2-\sigma_\epsilon^2)\sum_{i=1}^n\mathbf{a}_i^T\mathbf{e}_i = o_p(1)$. It implies that to prove (A.18) it is sufficient to show that

$$\sum_{i=1}^{n}\mathbf{a}_i^T\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i = O_p(1) \tag{A.19}$$

$$\sum_{i=1}^{n}\mathbf{a}_i^T\{\widehat{\mathbf{G}}_i(\widehat{\Lambda}^{-1}+\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)^{-1}\widehat{\mathbf{G}}_i^T - \mathbf{G}_i(\Lambda^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\}\mathbf{e}_i = o_p(1). \tag{A.20}$$

Consider equation (A.19). Because $\mathbf{e}_i$ are multivariate normal with variance $\boldsymbol{\Sigma}_i$, set c_i such that $c_i^2 = \mathbf{a}_i^T\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\boldsymbol{\Sigma}_i\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{a}_i$ and rewrite (A.19) as $S_n = \sum_{i=1}^n|c_i|\epsilon_i$, for independent standard normal variables ϵ_i. It suffices to prove that $\sum_{i=1}^n c_i^2 < \infty$, the result then follows from an application of the continuity theorem. For this we show that there exists $L<\infty$ such that $\|\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\boldsymbol{\Sigma}_i\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\| \le L$ for all i. This is not hard to show, because $\|\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\boldsymbol{\Sigma}_i\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\| \le \|\mathbf{G}_i(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\|^2\|\boldsymbol{\Sigma}_i\|$, and moreover $\|\mathbf{G}_i\|^2 = \mathrm{tr}(\mathbf{G}_i^T\mathbf{G}_i) < M$, $\|(\boldsymbol{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\| \le M^{1/2}\Lambda_{11}+M(\max_i m_i)\Lambda_{11}^2$ and $\|\boldsymbol{\Sigma}_i\| \le \sigma_\epsilon^2(\max_i m_i)\{\mathrm{tr}(\boldsymbol{\Lambda})+1\}$. Then $L = \sigma_\epsilon^2M^3\Lambda_{11}^2\{1+M^{1/2}\Lambda_{11}(\max_i m_i)\}^2(\max_i m_i)\{\mathrm{tr}(\boldsymbol{\Lambda})+1\}$ is finite and satisfies our inequality.

We show next (A.20). Because the eigenvalues σ_k^2 and the eigenfunctions $\theta_k(t)$ are consistently estimated, this reduces to showing that the dominant terms on the left hand side

of equation (A.20) are $o_p(1)$. Equivalently, we need to show that

$$\sum_{i=1}^{n} \mathbf{a}_i^T(\widehat{\mathbf{G}}_i - \mathbf{G}_i)(\mathbf{\Lambda}^{-1} + \mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i = o_p(1) \tag{A.21}$$

$$\sum_{i=1}^{n} \mathbf{a}_i^T\mathbf{G}_i(\mathbf{\Lambda}^{-1} + \mathbf{G}_i^T\mathbf{G}_i)^{-1}(\widehat{\mathbf{G}}_i - \mathbf{G}_i)^T\mathbf{e}_i = o_p(1) \tag{A.22}$$

$$\sum_{i=1}^{n} \mathbf{a}_i^T\mathbf{G}_i\{(\widehat{\mathbf{\Lambda}}^{-1} + \widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)^{-1} - (\mathbf{\Lambda}^{-1} + \mathbf{G}_i^T\mathbf{G}_i)^{-1}\}\mathbf{G}_i^T\mathbf{e}_i = o_p(1). \tag{A.23}$$

Consider now equation (A.21). The key idea is to use assumption (F1'), that for each subject i, the observation time points t_{ij} are generated uniformly from $(t_1, \cdots, t_m)$. Note that if $t_{il} = t_{i'l'} = t_j$, then $\widehat{g}_{ik,l} = \widehat{g}_{ik,l} = \widehat{\theta}_k(t_j)$, where $\widehat{g}_{ik,l}$ and $g_{ik,l}$ are the lth element of $\widehat{g}_{ik}$ and g_{ik}. The left hand-side of expression (A.21) can be written as

$$\begin{aligned}
& \sum_{i=1}^{n} \mathbf{a}_i^T(\widehat{\mathbf{G}}_i - \mathbf{G}_i)(\mathbf{\Lambda}^{-1} + \mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i \\
= & \sum_{i=1}^{n}\sum_{l=1}^{m_i}\sum_{k=1}^{M} a_{il}(\widehat{g}_{ik,l} - g_{ik,l})e'_{ik} \\
= & \sum_{k=1}^{M}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}e'_{ik}(\widehat{g}_{ik,l} - g_{ik,l})1(t_{il} = t_j) \\
= & \sum_{k=1}^{M}\sum_{j=1}^{m}\{\widehat{\theta}_k(t_j) - \theta_k(t_j)\}\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}e'_{ik}1(t_{il} = t_j),
\end{aligned} \tag{A.24}$$

where e'_{ik} is the kth element of $\mathbf{e}'_i = (\mathbf{\Lambda}^{-1} + \mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i$ and $1(t_{il} = t_j)$ is equal to 1 if $t_{il} = t_j$ and 0 otherwise. Set $B_{n,j} = \sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}e'_{ik}1(t_{il} = t_j)$. Because M is finite, it suffices to show that, for each k

$$E[\sum_{j=1}^{m}\{\widehat{\theta}_k(t_j) - \theta_k(t_j)\}B_{n,j}]^2 = o(1); \tag{A.25}$$

the result that expression $(A.24)$ is $o_p(1)$ follows then from an application of Bonferroni and Chebychev's inequalities. Simple algebra calculations points out that

$$E[\sum_{j=1}^{m}\{\widehat{\theta}_k(t_j) - \theta_k(t_j)\}B_{n,j}]^2 \leq n^{-2\alpha}E(\sum_{j=1}^{m}|B_{n,j}|)^2 \leq n^{-2\alpha}m\sum_{j=1}^{m}E(B_{n,j}^2),$$

using (F2'), that $\sup_{t\in\mathcal{T}}|\widehat{\theta}_k(t) - \theta_k(t)| = O_p(n^{-\alpha})$; thus to show (A.25) it suffices to show that $n^{-2\alpha}m\sum_{j=1}^{m}E(B_{n,j}^2) = o(1)$ as $n \to \infty$. This follows from $O(n^{-2\alpha}m) = o(1)$ and

$$E(B_{n,j}^2) = E\{\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}^2(e'_{ik})^2 1(t_{il} = t_j)\} = m^{-1}E\{\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}^2(e'_{ik})^2\} = O(m^{-1}), \tag{A.26}$$

using the independence between e'_{ik}'s and t_{ik}'s, and the fact that $E\{1(t_{il}=t_j)\}=m^{-1}$. For the last equality of (A.26) we used the following observations: 1)$||\mathbf{a}||=O(1)$, 2) $E(e'^2_{ik})<tr\{\mathrm{cov}(\mathbf{e}'_i)\}$, and 3) $\|\mathrm{cov}(\mathbf{e}'_i)\|<\infty$, where $\mathrm{cov}(\mathbf{e}'_i)=(\mathbf{\Lambda}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{\Sigma}_i\mathbf{G}_i(\mathbf{\Lambda}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}$.

Next we show (A.22) holds. Following a similar rationale, we rewrite equation (A.22) as

$$\begin{aligned}\sum_{i=1}^n \mathbf{a}_i^T\mathbf{G}_i(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}(\widehat{\mathbf{G}}_i-\mathbf{G}_i)^T\mathbf{e}_i &= \sum_{i=1}^n\sum_{l=1}^{m_i}\sum_{k=1}^M a'_{ik}(\widehat{g}_{ik,l}-g_{ik,l})^T e_{il}\\ &= \sum_{k=1}^M\sum_{j=1}^m\{\widehat{\theta}_k(t_j)-\theta_k(t_j)\}\sum_{i=1}^n\sum_{l=1}^{m_i}a'_{ik}e_{il}1(t_{il}=t_j),\end{aligned}$$

where a'_{ik} is the kth element of $\mathbf{a}'_i=\mathbf{a}_i^T\mathbf{G}_i(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}$. Set $C_{n,j}=\sum_{i=1}^n\sum_{l=1}^{m_i}a'_{ik}e_{il}1(t_{il}=t_j)$, and denote by $\mathbf{a}'$ the vector obtained by stacking $\mathbf{a}'_i$ over $i=1,\ldots,n$. we have that $||\mathbf{a}'||=O(1)$. Using similar arguments as above, we obtain $EC_{n,j}^2=O(m^{-1})$ for all j and furthermore conclude that $E[\sum_{j=1}^m\{\widehat{\theta}_k(t_j)-\theta_k(t_j)\}C_{n,j}]^2=o(1)$ and thus equation (A.22) holds.

Finally, we show (A.23) holds. Direct calculations show that

$$\begin{aligned}&\sum_{i=1}^n\mathbf{a}_i^T\mathbf{G}_i\{(\widehat{\mathbf{\Lambda}}^{-1}+\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)^{-1}-(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\}\mathbf{G}_i^T\mathbf{e}_i\\ =\ &\sum_{i=1}^n\mathbf{a}_i^T\mathbf{G}_i(\widehat{\mathbf{\Lambda}}^{-1}+\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)^{-1}(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i-\widehat{\mathbf{\Lambda}}^{-1}-\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i)(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i.\end{aligned}$$

Using again the consistency of the eigenvalues and eigenfunctions, it suffices to show that

$$\sum_{i=1}^n\mathbf{a}_i^T\mathbf{G}_i(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}(\widehat{\mathbf{\Lambda}}^{-1}+\widehat{\mathbf{G}}_i^T\widehat{\mathbf{G}}_i-\mathbf{\Lambda}^{-1}-\mathbf{G}_i^T\mathbf{G}_i)(\mathbf{\Lambda}^{-1}+\mathbf{G}_i^T\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{e}_i=o_p(1). \quad (A.27)$$

Use the notation of $\mathbf{a}'_i$ and $\mathbf{e}'_i$ above. Simple algebra points out that (A.27) follows from the following claims: 1) $\sum_{i=1}^n(\mathbf{a}'_i)^T(\widehat{\mathbf{G}}_i^T-\mathbf{G}_i^T)\mathbf{G}_i\mathbf{e}'_i=o_p(1)$, 2) $\sum_{i=1}^n(\mathbf{a}'_i)^T\mathbf{G}_i^T(\widehat{\mathbf{G}}_i-\mathbf{G}_i)\mathbf{e}'_i=o_p(1)$, and 3) $\sum_{i=1}^n(\mathbf{a}'_i)^T(\widehat{\mathbf{\Lambda}}^{-1}-\mathbf{\Lambda}^{-1})\mathbf{e}'_i=o_p(1)$. We can use roughly the same ideas as earlier to justify 1) and 2). Claim 3) follows from simpler arguments, as we now show. We notice that 3) can be re-written as $\sum_{k=1}^M(\widehat{\sigma}_k^2/\sigma_\epsilon^2-\sigma_k^2/\sigma_\epsilon^2)\sum_{i=1}^n\sum_{l=1}^{m_i}a'_{il}e'_{il}1(\Lambda_{ll}=\sigma_k^2/\sigma_\epsilon^2)$, which is $o_p(1)$, since for every k we have $(\widehat{\Lambda}_{kk}-\Lambda_{kk})=O_p(n^{-\alpha})$ and $\sum_{i=1}^n a'_{ik}e'_{ik}=O_p(1)$.

It follows that equation (A.20) holds and furthermore that condition (C2) is satisfied. #

A.2 Two samples of functional data

AA.1 Independent samples of functional data

Let $Y_{idj} = Y_{id}(t_{idj})$ be the response at time point t_{idj} corresponding to the ith subject within the dth sample, for $d = 1, 2$, $i = 1, \ldots, n_d$, and $j = 1, \ldots, m_{id}$. As in Section 3.2 it is assumed that $t_{idj} \in \mathcal{T}$ for some bounded and closed interval $\mathcal{T}$. For simplicity we consider $n_1 = n_2$, but our results can be extended easily to the case when $n_1/n_2 \to a$ for $0 < a < \infty$. It is assumed that, for each $d = 1, 2$, the response $Y_{id}(t_{idj})$ can be modeled similarly to (7) as:

$$Y_{id}(t_{idj}) = \mu(t_{idj}) + \mu_d(t_{idj}) + \sum_{k \geq 1} \xi_{d,ik}\theta_{d,k}(t) + \epsilon_{idj}, \tag{A.28}$$

where $\mu(\cdot)$ is the overall mean function, $\mu_d(\cdot)$ is the group specific mean deviation, and $\{\theta_{d,k}(t) : k \geq 1\}$ is the group specific orthogonal basis. For identifiability we assume that $\mu_1 + \mu_2 \equiv 0$. Moreover, the $\xi_{d,ik}$ are uncorrelated for all i, k and d, with mean zero and variance $E[\xi^2_{d,ik}] = \sigma^2_{d,k}$, and ϵ_{idj} are assumed independent and identically distributed with mean zero and variance $E[\epsilon^2_{dj}] = \sigma^2_{d,\epsilon}$. Denote by $\Gamma_d(\cdot, \cdot)$ the group d specific covariance function and consider its expansion in terms of orthogonal eigenfunctions, $\Gamma_d(t, t') = \sum_{k \geq 1} \sigma^2_{d,k}\theta_{d,k}(t)\theta_{d,k}(t')$, where $\theta_{d,k}$ are eigenfunctions and $\sigma^2_{d,1} > \sigma^2_{d,2} > \ldots$ are ordered eigenvalues for $d = 1, 2$. We assume that for each group the covariance function admits a finite number of non-zero eigenvalues. Our theoretical arguments are based on the additional assumption that $\{\xi_{d,ik}\}_k$ and ϵ^2_{dj} are jointly Gaussian distributed.

The main objective is to test that the group mean functions are equal, or equivalently that $\mu_1 \equiv 0$. Irrespective of the sampling design (dense or sparse), we assume that the set of *pooled* time points, $\{t_{idj} : i, j\}$ is dense in $\mathcal{T}$ for each d. Our methodology requires that the same sampling scheme is maintained for the two samples of curves, e.g., the curves are not densely observed in one sample and sparsely observed in the other sample. (One could extend the theory to the case of one sample being densely observed and the other sparse, but data of this type would be rare so we did not attempt such an extension.) We use quasi-residuals, $\widetilde{Y}_{idj} = Y_{id}(t_{idj}) - \bar{\mu}(t_{idj})$, where $\bar{\mu} = (\widetilde{\mu}_1 + \widetilde{\mu}_2)/2$ is the average of the estimated mean functions, $\widetilde{\mu}_d$ for $d = 1, 2$, which are obtained using the pooled data in each group. Because of the identifiability constraint, the estimated $\bar{\mu}$ can be viewed as a smooth estimate of the overall mean function μ. We assume that the overall mean function is estimated well enough (Kulasekera, 1995), so that $\widetilde{Y}_{idj}$ can be modeled similarly to (A.28), but without μ. Thus, we assume that $\mu \equiv 0$ and that the null hypothesis is $\mu_1 \equiv 0$. We

model $\mu_1(t)$ by pth truncated power polynomials: $\mu_1(t) = x_t\boldsymbol{\beta} + z_t\mathbf{b}$, where $\mathbf{b}$ is $N(\mathbf{0}, \sigma_b^2 I_K)$. Let $\mathbf{X}_{id}$ denote the $m_{id} \times (p+1)$ dimensional matrix with the jth row equal to $x_{t_{idj}}$, and let $\widetilde{\mathbf{X}}_i = [\mathbf{X}_{i1}^T \mid -\mathbf{X}_{i2}^T]^T$, and analogously define the $m_{id} \times K$ matrices $\mathbf{Z}_{id}$'s for $d = 1, 2$ and construct $\widetilde{\mathbf{Z}}_i = [\mathbf{Z}_{i1}^T \mid -\mathbf{Z}_{i2}^T]^T$ respectively. Here the vertical bar separates submatrices.

Denote by $\widetilde{\mathbf{Y}}_i$ the m_i-dimensional vector obtained by stacking first $\widetilde{Y}_{i1j}$'s over $j = 1, \ldots, m_{i1}$, and then $\widetilde{Y}_{i2j}$'s over $j = 1, \ldots, m_{i2}$, where $m_i = m_{i1} + m_{i2}$. It follows that, the $m_i \times m_i$-dimensional covariance matrix of $\widetilde{\mathbf{Y}}_i$, denoted by $\boldsymbol{\Sigma}_i = \text{diag}\{\boldsymbol{\Sigma}_{i,1}, \boldsymbol{\Sigma}_{i,2}\}$ is a block diagonal $m_i \times m_i$ dimensional matrix, where $\boldsymbol{\Sigma}_{i,d}$ is $m_{id} \times m_{id}$-dimensional matrix with the (j, j') element equal to $\Gamma_d(t_{idj}, t_{idj'}) + \sigma_{d,\epsilon}^2 1(j = j')$ for $d = 1, 2$. We can rewrite $\widetilde{\mathbf{Y}}_i$ using a LMM framework as $\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{X}}_i\boldsymbol{\beta} + \widetilde{\mathbf{Z}}_i\mathbf{b} + \mathbf{e}_i$, where $\mathbf{e}_i$ is m_i-dimensional vector, independent over i, with mean zero, and covariance matrix given by $\boldsymbol{\Sigma}_i$ described above.

Thus the hypothesis $\mu_1 \equiv 0$ is equivalent to $H_0 : \boldsymbol{\beta} = \mathbf{0}$ and $\sigma_b^2 = 0$ in (3); the pseudo LRT can be applied as discussed in Section 3.2, where the estimator $\widehat{\boldsymbol{\Sigma}}$ replaces $\boldsymbol{\Sigma}_{i,1}$ and $\boldsymbol{\Sigma}_{i,2}$ with their estimators, $\widehat{\boldsymbol{\Sigma}}_{i,1}$ and $\widehat{\boldsymbol{\Sigma}}_{i,2}$. The estimators $\widehat{\boldsymbol{\Sigma}}_{i,1}$ and $\widehat{\boldsymbol{\Sigma}}_{i,2}$ can be obtained as discussed in Section 3.2. Therefore, presuming that the data are densely sampled, condition (C2) of the Proposition 2.1 is met, under the assumption that (F1)–(F2) hold for each of the two samples. Likewise, in the sparse sampling design, (C2) is met under the assumption that (F1')–(F3') hold for each of the two samples. It follows that under these assumptions and the additional assumptions (C1) and (C3) of Proposition 2.1, the asymptotic null distribution of the pseudo LRT for testing the equality of the group mean function is given by (5).

AA.2 Dependent samples of functional data, dense design

Assume now two dependent sets of curves, and furthermore consider that in each set, the curves are densely sampled on a common grid of points $t_{idj} = t_j$ and $m_{id} = m$ for all i, d, j. Denote by $\{t_1, \ldots, t_m\}$ the common grid of points at which every curve is measured, and denote by $\widetilde{Y}_i(t_j) = Y_{i2}(t_j) - Y_{i1}(t_j)$ the ith pairwise difference. Using $\widetilde{Y}_i(t_j)$ reduces the data to a one-sample problem and allows us to apply the theory in Section 3. Note that $\widetilde{Y}_i(t_j)$ has a similar KL expansion as (7), $\widetilde{Y}_i(t_j) = -2\mu_1(t_j) + \sum_{l\geq 1} \widetilde{\zeta}_{il}\theta_{2,l}(t_j) + \widetilde{\epsilon}_i(t_j)$, where $\widetilde{\zeta}_{il}$'s can be viewed as principal component scores, are uncorrelated and have mean zero and variance equal to $2\sigma_{2,l}^2$. Also, $\widetilde{\epsilon}_i(t_j)$'s are independent and identically distributed as $N(0, 2\sigma_\epsilon^2)$. To assess the hypothesis that $\mu_1 = 0$, one can apply the pseudo LRT, as discussed in Section 3.2. In particular the conditions required by Proposition 3.2 involve only the estimation of the covariance function Γ_2 and of the noise variance σ_ϵ^2.

AA.3 Dependent samples of functional data, sparse design

Proof of Proposition 5.1: Under the sparse design, $\mathbf{\Sigma}$ is a block diagonal matrix, whose ith block is the $(m_{i1}+m_{i2})\times(m_{i1}+m_{i2})$ dimensional matrix

$$\mathbf{\Sigma}_i = \begin{pmatrix} \mathbf{\Sigma}_{i,11} & \mathbf{\Sigma}_{i,12} \\ \mathbf{\Sigma}_{i,12}^T & \mathbf{\Sigma}_{i,22} \end{pmatrix}, \tag{A.29}$$

where $\mathbf{\Sigma}_{i,dd} = \sigma_\epsilon^2(\mathbf{I}_{m_{id}} + \mathbf{G}_{1,id}\mathbf{\Lambda}_1\mathbf{G}_{1,id}^T + \mathbf{G}_{2,id}\mathbf{\Lambda}_2\mathbf{G}_{2,id}^T)$ and $\mathbf{\Sigma}_{i,12} = \sigma_\epsilon^2\mathbf{G}_{1,i1}\mathbf{\Lambda}_1\mathbf{G}_{1,i2}^T$. Here $\mathbf{G}_{1,id}$ and $\mathbf{G}_{2,id}$ are $m_{id}\times M_1$ and $m_{id}\times M_2$ dimensional matrices respectively, with the (j,k)th element and the (j,l)th element equal to the eigenfunctions $\theta_{1,k}(t_{idj})$, and $\theta_{2,l}(t_{idj})$ respectively, and $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are the $M_1\times M_1$ and $M_2\times M_2$ block diagonal matrices whose (k,k)th and (l,l)th components are $\sigma_{1,k}^2/\sigma_\epsilon^2$ and $\sigma_{2,l}^2/\sigma_\epsilon^2$ respectively. Similarly, define $\widehat{\mathbf{\Sigma}}_i$ by replacing σ_ϵ^2, $\mathbf{G}_{\iota,id}$, $\mathbf{\Lambda}_\iota$ with their respective estimators, for $\iota=1,2$. Partition $\mathbf{a}$ and $\mathbf{e}$ into n vectors, with $\mathbf{a}_i=(\mathbf{a}_{i1}^T,\mathbf{a}_{i2}^T)^T$ and $\mathbf{e}_i=(\mathbf{e}_{i1}^T,\mathbf{e}_{i2}^T)^T$ of length $m_{i1}+m_{i2}$. We want to prove condition (C2),

$$\sum_{i=1}^n \mathbf{a}_i^T\widehat{\mathbf{\Sigma}}_i^{-1}\mathbf{a}_i - \sum_{i=1}^n \mathbf{a}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{a}_i = o_p(1), \tag{A.30}$$

$$\sum_{i=1}^n \mathbf{a}_i^T\widehat{\mathbf{\Sigma}}_i^{-1}\mathbf{e}_i - \sum_{i=1}^n \mathbf{a}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{e}_i = o_p(1). \tag{A.31}$$

The equality (A.30) follows from assumption (M3'), which implies that $\max_i||\widehat{\mathbf{G}}_{\iota,id}-\mathbf{G}_{\iota,id}|| = o_p(1)$, $\max_i||\widehat{\mathbf{\Lambda}}_{\iota,ii}-\mathbf{\Lambda}_{\iota,ii}|| = o_p(1)$, and $\widehat{\sigma}_\epsilon^2-\sigma_\epsilon^2 = o_p(1)$ and from the continuous mapping theorem, which implies that $\max_i||\widehat{\mathbf{\Sigma}}_i^{-1}-\mathbf{\Sigma}_i^{-1}|| = o_p(1)$. The proof is concluded since $\sum_{i=1}^n||\mathbf{a}_i||^2 = O(1)$.

We turn to equality (A.31). We re-write $\mathbf{\Sigma}_i^{-1}$ using inverse of the partition matrix as follows:

$$\mathbf{\Sigma}_i^{-1} = \begin{pmatrix} \mathbf{V}_{i1}^{-1} & -\mathbf{\Sigma}_{i,11}^{-1}\mathbf{\Sigma}_{i,12}\mathbf{V}_{i2}^{-1} \\ -\mathbf{\Sigma}_{i,22}^{-1}\mathbf{\Sigma}_{i,12}^T\mathbf{V}_{i1}^{-1} & \mathbf{V}_{i2}^{-1} \end{pmatrix} = \begin{pmatrix} [\mathbf{\Sigma}_i^{-1}]_{11} & [\mathbf{\Sigma}_i^{-1}]_{12} \\ [\mathbf{\Sigma}_i^{-1}]_{21} & [\mathbf{\Sigma}_i^{-1}]_{22} \end{pmatrix},$$

where $\mathbf{V}_{i1} = \mathbf{\Sigma}_{i,11}-\mathbf{\Sigma}_{i,12}\mathbf{\Sigma}_{i,22}^{-1}\mathbf{\Sigma}_{i,12}^T$, and $\mathbf{V}_{i2} = \mathbf{\Sigma}_{i,22}-\mathbf{\Sigma}_{i,12}^T\mathbf{\Sigma}_{i,11}^{-1}\mathbf{\Sigma}_{i,12}$. Similarly, we can define $\widehat{\mathbf{\Sigma}}_i^{-1}$ by replacing the quantities with their estimates. The left hand side of equation (A.31) can be decomposed into $v_{11}+v_{12}+v_{21}+v_{22}$, where $v_{sl} = \sum_{i=1}^n \mathbf{a}_{is}^T([\widehat{\mathbf{\Sigma}}_i^{-1}]_{sl}-[\mathbf{\Sigma}_i^{-1}]_{sl})\mathbf{e}_{il}$. It is sufficient to show that $v_{sl}=o_p(1)$ for $1\le s,l\le 2$. These results can be derived using similar techniques as in the proof of the Proposition 3.3, but they involve more tedious algebra. In the interest of space the details are omitted here. It follows that our proof is concluded. #

Table 1: *Type I error rates, based on* 1000 *simulations, of the pseudo LRT for testing* $H_0 : \mu \equiv 0$ *in the context of dense functional data generated by model (7) with* $\sigma_\epsilon^2 = 0.125$*, for various* n *and* m*, and when the scores* ξ_{ik} *are generated from a scaled* t_5*- distribution with* 5 *degrees of freedom (*t_5*) or centered and scaled* χ_5^2*- distribution with* 5 *degrees of freedom (*χ_5*) (non-normal). In the pseudo LRT, the mean function is modeled using linear splines.*

$(n,\, m)$	scores distribution	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
	t_5 (heavy tailed)				
(50, 100)		0.224	0.107	0.048	0.013
(50, 400)		0.237	0.134	0.064	0.015
(100, 100)		0.215	0.104	0.043	0.009
(100, 400)		0.214	0.113	0.071	0.018
(200, 80)		0.206	0.094	0.053	0.013
	χ_5^2 (right skewed)				
(50, 100)		0.213	0.109	0.061	0.017
(50, 400)		0.206	0.091	0.053	0.013
(100, 100)		0.207	0.119	0.057	0.011
(100, 400)		0.201	0.085	0.047	0.012
(200, 80)		0.193	0.095	0.052	0.005