Statistica Sinica Preprint No: SS-2018-0456	
Title	Hypothesis Testing in Large-scale Functional Linear Regression
Manuscript ID	SS-2018-0456
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0456
Complete List of Authors	Kaijie Xue and Fang Yao
Corresponding Author	Fang Yao
E-mail	fyao@utstat.toronto.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Statistica Sinica

Hypothesis Testing in Large-scale Functional Linear Regression

Kaijie Xue and Fang Yao

Nankai University and Peking University

Abstract: We explore large-scale functional linear regression in which the scalar response is associated with a potentially ultrahigh number of functional predictors, leading to a more challenging model framework than the classical case. We establish a rigorous procedure for testing a general hypothesis on an arbitrary subset of regression coefficient functions. Specifically, we exploit the techniques developed for post-regularization inferences, and propose a new test for the aforementioned regression based on a decorrelated score function that separates the primary and nuisance parameters in functional spaces. We also devise the corresponding decorrelated Wald and likelihood ratio tests, and establish the exact equivalence among these three tests for the model under consideration. The proposed test is shown to be uniformly convergent to the prescribed significance. We show its finite-sample performance using simulation studies and a data set from the Human Connectome Project that identifies brain regions associated with emotional tasks.

Key words and phrases: Decorrelated score, functional data, high dimensions, functional linear regression, multiplier bootstrap.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

1. Introduction

The classical functional linear regression (FLR) is widely used to model the linear relationship between a scalar response Y and a functional predictor, which is often assumed to be sampled from an $L^2(T)$ random process $X(t)$ defined on a compact interval $T \subseteq \mathbb{R}$. Specifically, given n independent and identically distributed (i.i.d.) pairs $\{Y_i, X_i(\cdot)\}$, the classical FLR is formulated as

$$Y_i = \int_T X_i(t)\beta(t)dt + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where both Y_i and X_i are centered, without loss of generality, that is, $EY_i = 0$ and $EX_i(t) = 0$, for $t \in T$; the unknown regression parameter function $\beta(t)$ is square-integrable, that is, $\beta \in L^2(T)$; and the i.i.d. regression error ϵ_i is independent of X_i with mean zero and finite variance $\sigma^2 < \infty$. This model has been studied extensively in relation to functional data analyses (Ramsay and Dalzell, 1991; Cardot et al., 1999; Fan and Zhang, 2000; Yuan and Cai, 2010, among others), including its theoretical considerations (Hall and Horowitz, 2007; Cai and Yuan, 2012) and statistical inference (Cardot et al., 2003; Lei, 2014; Hilgert et al., 2013; Shang and Cheng, 2015); see Ramsay and Silverman (2005) for an overview and examples. Numerous works have extended the classical FLR. These extensions include the functional response (Faraway, 1997; Cuevas et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

2002; Yao et al., 2005), generalized FLR (Escabias et al., 2004; Müller and Stadtmüller, 2005; Shang and Cheng, 2015), partially FLR (Lian, 2011; Kong et al., 2016), and additive regression (Müller and Yao, 2008; Zhu et al., 2014; Fan et al., 2015), among others.

In modern scientific experiments, the response Y is potentially associated with multiple, or even a large number of functional predictors. For example, Lian (2013) proposed an FLR involving a fixed number of functional predictors. Kong et al. (2016) considered a regularized estimation and variable selection for a partially FLR that contains high-dimensional scalar covariates and a finite number of functional predictors. However, when applying an FLR to large-scale data, the number of potential functional predictors p_n can be much larger than the sample size n, even though the significant predictors of size q_n are usually assumed to be sparse or at a fraction polynomial order of n. Examples can be found in neuroimaging analyses that focus on the relationship between a disease marker and a number of brain regions of interest (ROI) over time. This consideration motivates the following large-scale FLR model:

$$Y_i = \sum_{j=1}^{p_n} \int_T X_{ij}(t)\beta_j(t)dt + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.2}$$

where p_n is allowed to grow exponentially with the sample size n, (without loss

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

of generality) the first q_n important parameter functions $\{\beta_j : j = 1 \ldots, q_n\}$ are assumed to be nonzero, with the rest zero, and the i.i.d. error ϵ_i is independent of $\{X_{ij} : j = 1, \ldots, p_n\}$ with mean zero and variance σ^2. It is common to use a set of pre-fixed (i.e., B-splines, wavelets) or data-driven (i.e., eigenfunctions) bases to represent the underlying process X_j of each predictor $\{X_{ij} : i = 1, \ldots, n\}$. The data-driven bases, such as eigenfunctions, are efficient for representation, but necessarily for regression. However, they have to be estimated from p_n separate functional principal component analysis (FPCA) procedures, which is computationally intensive, especially when $p_n \gg n$. For instance, a singular value decomposition (SVD)-based method usually demands computation of order $O\{p_n(nm^2 + n^2m)\}$, which can be much higher if pre-smoothing is needed. Thus, we adopt a common pre-fixed basis $\{b_k : k \geq 1\}$ that is complete and orthonormal in $L^2(T)$ for all processes X_j, for $j = 1, \ldots, p_n$. As such, we do not further pursue other complicated basis-seeking procedures, such as the functional partial least squares method (Reiss and Ogden, 2007). The proposed method requires computation of order $O(p_n nm)$ and automatically takes smoothing into account.

The main contribution of this study is to develop a rigorous testing procedure for a general hypothesis on an arbitrary subset of regression functions $\{\beta_j : j = 1, \ldots, p_n\}$. The challenge arises from the ultrahigh-dimensionality in p_n, which

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

can be as exponentially large as n, and the intrinsic infinite-dimensionality of each X_j, for $j = 1, \ldots, p_n$. Although the FLR (and its variants) has been well studied, few works have examined their inference procedures. For example, Hilgert et al. (2013) and Lei (2014) considered adaptive tests for a single regression function in a classical FLR, and Shang and Cheng (2015) did so for the generalized FLR. In the current exposition, we adopt a general class of nonconvex penalty functions (Loh and Wainwright, 2015), which include the LASSO penalty (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010) as special cases. Furthermore, the theoretical properties in high-dimensional linear regressions have been studied extensively (Meinshausen and Bühlmann, 2006; van de Geer, 2008; Meinshausen and Yu, 2009; Bickel et al., 2009; Zhang, 2009; Fan and Lv, 2011; Wang et al., 2013, 2014; Fan et al., 2014; Loh and Wainwright, 2015, among many others). Recently, research on inferences in high-dimensional linear regressions has increased, especially for the LASSO-type convex penalty (Tibshirani, 1996). These studies include those of Wasserman and Roeder (2009), Meinshausen and Bühlmann (2010), and Shah and Samworth (2013) on sample splitting and subsampling, Zhang and Zhang (2014) and van de Geer et al. (2014) on bias correction methods, and Lockhart et al. (2014) and Taylor et al. (2014) on conditional inferences on the event that some

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

covariates have been selected, among others.

This study is inspired by the unconditional inference based on a decorrelated score function of Ning and Liu (2017), owing to its generality, and because it does not require data splitting or strong minimal signal conditions. We first exploit a penalized least squares procedure, treating the truncated coefficients of each β_j as a group. In this way, we obtain estimation consistency without needing oracle properties under weaker minimal signal conditions that allow for a wider class of suitable settings. Then, we devise the decorrelated score function in the context of a large-scale FLR that tests a general null hypothesis on any subset of $\{\beta_j : j \leq p_n\}$. Unlike testing a null hypothesis on a single parameter in a high-dimensional linear regression, the limiting distribution for such a general null hypothesis is intractable. Hence, we adopt the multiplier bootstrap to approximate the limiting distribution of the score test statistic under the null hypothesis, and provide theoretical guarantees for all possible levels in a uniform manner. Furthermore, we introduce the counterparts of the score test (i.e., the decorrelated Wald test and decorrelated likelihood ratio test) and establish the exact equivalence of the three tests for the model under consideration.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

2. Regularized estimation by group penalized least squares

Recall that the large-scale FLR defined in (1.2), underlying predictor processes X_j, and the corresponding regression functions β_j are expressed by a complete and orthonormal basis $\{b_k : k \geq 1\}$, leading to an infinite-dimensional representation. Specifically, let the functional predictors and the associated regression functions be expressed as linear combinations of $\{b_k : k \geq 1\}$; that is, $\beta_j = \sum_{k=1}^{\infty} \eta_{jk} b_k$ and $X_{ij} = \sum_{k=1}^{\infty} \theta_{ijk} b_k$, where the coefficients $\theta_{ijk} = \int_T X_{ij}(t) b_k(t) dt$ that coincide with the projections are mean zero random variables with variances $E(\theta_{ijk}^2) = \omega_{jk} > 0$. As a result, model (1.2) can be reformulated as

$$Y_i = \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \theta_{ijk} \eta_{jk} + \epsilon_i. \tag{2.3}$$

To perform an estimation and inference on the regression functions of primary interest, it is not feasible to directly minimize the square loss with respect to the infinite sequences of unknown coefficients η_{jk}. A common practice is to truncate up to the first s_n leading terms allowed to grow with n, where s_n controls the complexity of β_j as a whole function, rather than viewing the basis terms as separate predictors, and balances the bias-variance trade-off in a similar spirit to

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

a classical nonparametric regression. Hence, model (1.2) becomes

$$Y_i = \sum_{j=1}^{p_n} \sum_{k=1}^{s_n} \theta_{ijk}\eta_{jk} + (\epsilon_i + \sum_{j=1}^{p_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk}\eta_{jk}), \qquad i = 1, \ldots, n. \tag{2.4}$$

A similar technique is used by Rice and Silverman (1991), Yao et al. (2005), Hall and Hosseini-Nasab (2006), Cai and Hall (2006), Zhang and Chen (2007), Hall and Horowitz (2007), Fan et al. (2015), and Kong et al. (2016), among others. Ideally, one would use different truncation sizes for each β_k. However, selecting truncations for a large number of functional predictors is computationally infeasible. In practice, we adopt the strategy suggested by Kong et al. (2016) of using a common s_n to perform the regularized estimation. Then, we use an ordinary least squares for the retained predictors and choose different truncations using K-fold cross-validation for, say, $K = 5$. Nonetheless, the use of a common s_n suffices for the methodological development and theoretical analysis.

Remark. To the best of our knowledge, this type of large-scale FLR first appeared in Fan et al. (2015), who considered a penalized procedure for model estimation and selection. However, our primary interest is hypothesis testing. A careful inspection of Condition 1(A) in Fan et al. (2015, Appendix B), which requires $\sum_{k=1}^{\infty} \theta_{ijk}^2 k^4 < C^2$ for a universal constant C, for $i = 1, \ldots, n, j = 1, \ldots, p_n$, reveals that all random processes X_j are bounded in $L^2(T)$, which

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

excludes the Gaussian processes. Furthermore, Condition 2(D) in Fan et al. (2015) assumes that the minimal eigenvalues of $n^{-1}\Theta_j'\Theta_j \geq c_0$. This is bounded from below by a constant c_0 uniformly in $1 \leq j \leq q_n$ (i.e., the important ones), where $\Theta_j = (\theta_{ijk})_{1\leq i\leq n;1\leq k\leq s_n}$ is the $n \times s_n$ design matrix induced by X_j. In fact, this crucial condition is not valid for an infinite-dimensional L^2 process, because the minimal eigenvalues necessarily approach zero when s_n diverges; a typical example is given by the Karhunen-Loève expansion. In contrast, we do not make such assumptions. As such, the predictor processes are genuinely functional in the large-scale FLR (1.2).

In addition to the truncation, it is essential to impose a suitable penalty on each regression function as a whole using a functional version of the group regularization (Yuan and Lin, 2006). To regularize predictors on a comparable scale, we often standardize the scalar predictors in a linear regression (Fan and Li, 2001). For the functional predictors X_j, we choose to account for the variability in the grouped projection coefficients θ_{ijk} in the $n \times s_n$ design matrix $\Theta_j = (\theta_{ijk})_{1\leq i\leq n;1\leq k\leq s_n}$. Hence, $n^{-1/2}||\Theta_j\eta_j||_2$ invokes a group penalty that shrinks the unimportant regression function to zero, where $\|\cdot\|_2$ is the Euclidean or ℓ_2 norm (if an infinite sequence). For technical convenience, we scale up the penalty parameter λ_n by $s_n^{1/2}$, which does not affect the relative weighting of the penalties, given the common group size s_n. Thus, our target is to minimize

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

the penalized square loss function, as follows, denoting $\eta = (\eta_1', \ldots, \eta_{p_n}')'$ with vectors $\eta_j = (\eta_{j1}, \ldots, \eta_{js_n})'$, and $\|\cdot\|_1$ as the ℓ_1 norm:

$$\min_{\eta:||\eta||_1 \le R_n} \Big\{ \overbrace{\underbrace{(2n)^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\theta_{ijk}\eta_{jk})^2}_{L_n(\eta)} + \underbrace{\sum_{j=1}^{p_n}\rho_{\lambda_n s_n^{1/2}}(n^{-1/2}||\Theta_j\eta_j||_2)}_{P_{\lambda_n}(\eta)}}^{Q_n(\eta)} \Big\} \quad (2.5)$$

where $\rho_\lambda(\cdot)$ with the tuning parameter λ belongs to a general class of nonconvex penalty functions satisfying conditions (P1)–(P5) in Appendix A, which includes popular penalties such as the LASSO, SCAD, and MCP (Loh and Wainwright, 2015). The positive constraint R_n should be chosen carefully to make the true value η^* a feasible point, such that $\|\eta^*\|_1 \le R_n$. For instance, it is often the case that $\|\eta^*\|_1 = O(q_n)$, suggesting that $R_n \sim q_n$. Upon solving the optimization problem in (2.5), which is guaranteed to have a global minimum by the Weierstrass extreme value theorem if $\rho_\lambda(\cdot)$ is continuous, the regularized estimator for each β_j is given by $\hat{\beta}_j(t) = \sum_{k=1}^{s_n}\hat{\eta}_{jk}b_k(t)$, where $\hat{\eta}$ is obtained from (2.5). An implementation using a coordinate descent algorithm based on Ravikumar et al. (2008), with a slight modification, is presented in Appendix A. The tuning parameters λ_n and s_n are chosen using K-fold cross-validation (e.g.,$K = 5$). Note that for the purpose of general hypothesis testing, it is sufficient to obtain a consistent estimation of η from (2.5) in both the ℓ_1 and the ℓ_2 sense, as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

stated in Theorem 1, whereas the selection consistency or oracle property is not necessary. Before stating Theorem 1, the main technical conditions (A1)–(A6) are discussed below. Conditions (B1)–(B3) on the relationship between several quantites, such as R_n, s_n, q_n, and λ_n and the penalty function requirements (P1)–(P5) are deferred to Appendix A and B respectively.

Because we consider a large-scale FLR with functional predictors on a comparable scale, it is reasonable to require the second moment of each X_j, $\int_T E(X_j^2)$, to be uniformly bounded from above. Furthermore, the minimal eigenvalue of $\Lambda = \text{diag}\{\Lambda_j : j \leq p_n\}$ decays at a polynomial order of s_n, where $\Lambda_j = \text{diag}\{\omega_{jk}^{1/2} : k \leq s_n\}$; that is,

(A1) $\sup_{j \leq p_n} \sum_{k=1}^{\infty} \omega_{jk} < \infty, \quad \lambda_{\min}(\Lambda) \geq c s_n^{-a/2}$, for some constants $c > 0$ and $a > 1$.

Condition (A1) implies that the variances $\{\omega_{jk} : k \leq s_n\}$ for each j are allowed to be unsorted, with possible ties. This is distinct from Condition 2(D) in Fan et al. (2015), which requires that $\lambda_{\min}(\Lambda)$ be bounded by a constant from below, and is not applicable for functional predictors. For the next assumption on the distributions of several random quantities, we define the subGaussian norm as $\|X\|_{\phi_1} = \sup_{q \geq 1} q^{-1/2}\{E(|X|^q)\}^{1/q}$ for the subGaussian random variable X, and define the sub-exponential norm as $\|X\|_{\phi_2} = \sup_{q \geq 1} q^{-1}\{E(|X|^q)\}^{1/q}$ for the sub-exponential random variable X. We assume the following:

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

(A2) The random quantities ϵ_i, $\omega_{jk}^{-1/2}\theta_{ijk}$, and $(w_l{}'F_i - E_{il})\{E(E_{il}^2)\}^{-1/2}$ are centered subGaussian random variables satisfying $||\epsilon_i||_{\phi_1} \leq c$, $||\omega_{jk}^{-1/2}\theta_{ijk}||_{\phi_1} \leq c$, and $||(w_l{}'F_i - E_{il})\{E(E_{il}^2)\}^{-1/2}||_{\phi_1} \leq c$, respectively, for some positive constant c, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, p_n$, $k = 1, \ldots, \infty$, and $l = 1, \ldots, h_n s_n$.

Together, conditions (A1) and (A2) imply that θ_{ijk} and $(w_l{}'F_i - E_{il})$ are also centered subGaussian satisfying $||\theta_{ijk}||_{\phi_1} \leq c_1$ and $||w_l{}'F_i - E_{il}||_{\phi_1} \leq c_1$, for some positive constant c_1, uniformly in $1 \leq i \leq n$, $1 \leq j \leq p_n$, $1 \leq l \leq h_n s_n$, and $k \geq 1$. Next, we denote the information matrix and the standardized information matrix by $I = E(G_i G_i{}')$ and $\breve{I} = \Lambda^{-1} I \Lambda^{-1}$, respectively, where G_i is the vector containing θ_{ijk} projected from the ith subject. We assume that the eigenvalues of the standardized information matrix satisfy the following:

(A3) $m_0 \leq \lambda_{\min}(\breve{I}) \leq \lambda_{\max}(\breve{I}) \leq m_1 < \infty$, for some constants $m_1 > m_0 > 0$, with $m_0 > 2^{-1} m_1 \mu$, where $\mu > 0$ is a constant such that $\rho_{\lambda,\mu}(t)$ is convex in t; see Appendix A for the general conditions on the nonconvex penalty $\rho_{\lambda,\mu}(t)$.

From (A1) and (A3), we have that $\lambda_{\min}(I) = \lambda_{\min}(\Lambda \breve{I} \Lambda) \geq c s_n^{-a}$, for some constant $c > 0$. As a special case, if the functional predictors are uncorrelated, $\breve{I}$ is reduced to an identity matrix that apparently fulfills (A3). Similarly, we

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

denote the partial information matrix and its standardized version by $I_{\mathcal{H}_n|\mathcal{H}_n^c} = I_{\mathcal{H}_n\mathcal{H}_n} - w' I_{\mathcal{H}_n^c\mathcal{H}_n}$ and $\breve{I}_{\mathcal{H}_n|\mathcal{H}_n^c} = \Lambda_{\mathcal{H}_n}^{-1} I_{\mathcal{H}_n|\mathcal{H}_n^c} \Lambda_{\mathcal{H}_n}^{-1}$, respectively. Then, we impose a mild assumption on the correlation structure between the predictors to be tested and the other nuisance predictors:

(A4) $c_1 \leq \lambda_{\min}(\breve{I}_{\mathcal{H}_n|\mathcal{H}_n^c}) \leq \lambda_{\max}(\breve{I}_{\mathcal{H}_n|\mathcal{H}_n^c}) \leq c_2 < \infty$, for constants $c_2 > c_1 > 0$.

The number of functional predictors p_n can grow exponentially with the sample size:

(A5) $\log p_n \sim n^{\beta}$, for some $\beta \in (0, 9^{-1})$,

where $a_n \sim b_n$ denotes $c_1 \leq \lim_{n\to\infty} |a_n/b_n| \leq c_2$, for some $c_1, c_2 > 0$. We assume that the first q_n nonzero regression functions belong to a Sobolev ball with smoothness governed by a regularity constant δ:

(A6) $\sup_{j\leq q_n} \sum_{k=1}^{\infty} \eta_{jk}^2 k^{2\delta} < c$, for some positive constants δ and c.

Theorem 1. *Under conditions (A1)–(A3), (A5)–(A6), (B1), (B3), and (P1)–(P5), every local minimizer $\hat{\eta}$ of $Q_n(\eta)$ obtained from (2.5) satisfies that*

1) $||\hat{\eta} - \eta||_2 \leq c_0 \lambda_n s_n^{a/2+1/2} q_n^{1/2}$, *with probability tending to one, for some constant $c_0 > 0$,*

2) $||\hat{\eta} - \eta||_1 \leq c_1 \lambda_n s_n^{a/2+1} q_n$, *with probability tending to one, for some constant $c_1 > 0$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Note that the upper bounds in 1) and 2) depend on the truncation size s_n, which behaves like a tuning parameter in a nonparametric regression, and reflects the variability of $\hat{\eta}$. From Theorem 1, the consistency of the estimated regression curves $\hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t)$ follows

$$\sup_{j \le p_n} ||\hat{\beta}_j - \beta_j||_{L^2} \le \sup_{j \le p_n} ||\hat{\eta}_j - \eta_j||_2 + s_n^{-\delta} \sup_{j \le q_n} \Big(\sum_{k=s_n+1}^{\infty} \eta_{jk}^2 k^{2\delta} \Big)^{1/2}$$
$$= O(\lambda_n s_n^{a/2+1/2} q_n^{1/2} + s_n^{-\delta}), \tag{2.6}$$

with probability tending to one, where a and δ govern the smoothness of the functional processes and the regression functions, respectively. Note that (B1) in Appendix B incorporates $\delta > a + 1 > 2$, which indicates that the regression curves are relatively smoother than the functional processes, and that q_n is relatively small in the sample size, reflecting the sparseness of the model. In particular, because $||\eta^*||_1 = \sum_{j=1}^{q_n} \sum_{k=1}^{s_n} |\eta_{jk}^*| = O(q_n)$ under (A6) and (B1), it is feasible to assume $R_n \sim q_n$. In addition, (B3) implies that $\max\{(\log p_n/n)^{1/2}, q_n s_n^{-\delta}\} \le \lambda_n \le R_n^{-1}$. By simple calculation, we can minimize (2.6) using $s_n^* = \big\{2\delta(a+1)^{-1}\lambda_n^{-1} q_n^{-1/2}\big\}^{2/(a+1+2\delta)}$, yielding $\sup_{j \le p_n} ||\hat{\beta}_j - \beta_j||_{L^2} = O\big\{(\lambda_n^2 q_n)^{\delta/(a+1+2\delta)}\big\}$. Note that the estimation consistency in Theorem 1 is sufficient to guarantee the consistency of the testing procedure in following sections. That is, we do not have to further refine the convergence rate, which is another advantage of our

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

proposal.

3. Bootstrapped score test for a general hypothesis in a large-scale FLR

Our goal is to test a class of hypotheses that is of full generality in a large-scale FLR framework. Denote $\mathcal{P}_n = \{1, \ldots, p_n\}$ as the index set of all functional predictors, let $\mathcal{H}_n \subseteq \mathcal{P}_n$ be an arbitrary nonempty subset of $\mathcal{P}_n$ with cardinality $|\mathcal{H}_n| = h_n \leq p_n$, and denote the complement of $\mathcal{H}_n$ as $\mathcal{H}_n^c = \mathcal{P}_n \setminus \mathcal{H}_n$. Then, the hypothesis can be expressed as

$$H_0 : \|\beta_j\|_{L^2} = 0 \text{ for all } j \in \mathcal{H}_n \text{ v.s. } H_a : \|\beta_j\|_{L^2} > 0 \text{ for some } j \in \mathcal{H}_n, \quad (3.7)$$

noting that the cardinality h_n can be as large as p_n, allowing for a hypothesis of any size on $\{\beta_j : j = 1, \ldots, p_n\}$.

To test the general null hypothesis in (3.7), we use a combination of consistently estimated regression functions and a new type of score function. As illustrated in Ning and Liu (2017), the motivation for considering a decorrelated score is the high-dimensionality of the nuisance parameter space $\mathcal{H}_n^c = \mathcal{P}_n \setminus \mathcal{H}_n$, which makes the limiting distribution of the estimated nuisance parameter constrained by the null hypothesis intractable (Fu and Knight, 2000). Hence, the key is to decorrelate the score function of the primary parameter in $\mathcal{H}_n$ from that of the nuisance parameter in $\mathcal{H}_n^c$ in order to control the variability induced by the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

high dimensionality. This decorrelation operation is a natural extension of the profile score to the high-dimensional case, and leads to a test that is asymptotically equivalent to the classical Rao score test in the low-dimensional case (Cox and Hinkley, 1979; Ning and Liu, 2017).

We first introduce some notation for the score decorrelation in the proposed large-scale FLR. Recall that ω_{jk} is the variance of the i.i.d. projection coefficient $\{\theta_{ijk} = \int_T X_{ij}(t)b_k(t)dt : i = 1, \ldots, n\}$. Denote $\Lambda_j = \text{diag}\{\omega_{j1}^{1/2}, \ldots, \omega_{js_n}^{1/2}\}$, for $j \leq p_n$, as the block diagonal matrix $\Lambda_{\mathcal{H}_n} = \text{diag}\{\Lambda_j : j \in \mathcal{H}_n\}$; similarly $\Lambda_{\mathcal{P}_n} \equiv \Lambda$. Let $\Theta = (G_1', \ldots, G_n')' = (\Theta_{\mathcal{H}_n}, \Theta_{\mathcal{H}_n^c})$, $\Theta_{\mathcal{H}_n} = (E_1', \ldots, E_n')'$, and $\Theta_{\mathcal{H}_n^c} = (F_1', \ldots, F_n')'$, where G_i, E_i, and F_i are vectors containing the coefficients θ_{ijk} from the corresponding functional predictors for the ith subject. Here, $\Theta_{\mathcal{H}_n}$ is formed by concatenating $\{\Theta_j : j \in \mathcal{H}_n\}$ in a row, as is $\Theta_{\mathcal{H}_n^c}$, where Θ_j is an $n \times s_n$ design matrix with θ_{ijk} as its ikth entry. In addition, denote $\eta = (\eta_{\mathcal{H}_n}', \eta_{\mathcal{H}_n^c}')'$ and $Y = (Y_1, \ldots, Y_n)'$, where $\eta_{\mathcal{H}_n}$ stacks $\{\eta_j : j \in \mathcal{H}_n\}$ in a column, as in the case of $\eta_{\mathcal{H}_n^c}$. Here, we view the least squares $L_n(\eta) = L_n(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n}) = (2n)^{-1}(Y - \Theta\eta)'(Y - \Theta\eta)$ as the negative likelihood function of η without introducing extra notation. Furthermore, denoting $I_{\mathcal{H}_n^c\mathcal{H}_n} = E(F_i E_i{}')$ and $I_{\mathcal{H}_n^c\mathcal{H}_n^c} = E(F_i F_i{}')$, we define

$$w = I_{\mathcal{H}_n^c\mathcal{H}_n^c}^{-1} I_{\mathcal{H}_n^c\mathcal{H}_n} = (w_1, \ldots, w_{h_n s_n}) \in \mathbb{R}^{(p_n - h_n)s_n \times h_n s_n}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

For the decorrelation, we define a new score function with respect to the primary parameter $\eta_{\mathcal{H}_n}$, denoted by $S(\eta)$, in the context of our large-scale FLR, as follows:

$$\begin{aligned} S(\eta) &= S(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = n^{-1}\Lambda_{\mathcal{H}_n}^{-1}(w'\Theta'_{\mathcal{H}_n^c} - \Theta'_{\mathcal{H}_n})(Y - \Theta_{\mathcal{H}_n}\eta_{\mathcal{H}_n} - \Theta_{\mathcal{H}_n^c}\eta_{\mathcal{H}_n^c}) \\ &= n^{-1}\sum_{i=1}^{n}\Lambda_{\mathcal{H}_n}^{-1}(w'F_i - E_i)(Y_i - E_i'\eta_{\mathcal{H}_n} - F_i'\eta_{\mathcal{H}_n^c}). \end{aligned} \tag{3.8}$$

It is easy to verify that this new score function with respect to the primary parameter $\eta_{\mathcal{H}_n}$ is uncorrelated with the traditional score function with respect to the nuisance parameter $\eta_{\mathcal{H}_n^c}$; that is, $E\{S(\eta)\nabla_{\eta_{\mathcal{H}_n^c}} L_n(\eta)\} = 0$ (Ning and Liu, 2017), where ∇_γ denotes the gradient vector taken with respect to γ.

Given the consistent estimation of the regression coefficients and the decorrelated score function, we are ready to construct the proposed score test for the general hypothesis in (3.7) in a large-scale FLR. Note that the decorrelated score function $S(\eta)$ defined in (3.8) cannot be calculated directly from the observed data, owing to the unknown quantities $w = I_{\mathcal{H}_n^c\mathcal{H}_n^c}^{-1} I_{\mathcal{H}_n^c\mathcal{H}_n}$ and $\Lambda_{\mathcal{H}_n}$. It is straightforward to estimate $\Lambda_{\mathcal{H}_n}$ by substituting in $\hat{\omega}_{jk} = n^{-1}\sum_{i=1}^{n}\theta_{ijk}^2$, denoted by $\hat{\Lambda}_{\mathcal{H}_n}$; the process is similar for $\hat{\Lambda}$ and $\hat{\Lambda}_{\mathcal{H}_n^c}$. To estimate w, a natural choice is the moment estimator $\hat{w} = \hat{I}_{\mathcal{H}_n^c\mathcal{H}_n^c}^{-1}\hat{I}_{\mathcal{H}_n^c\mathcal{H}_n}$, where $\hat{I}_{\mathcal{H}_n^c\mathcal{H}_n^c} = n^{-1}\Theta_{\mathcal{H}_n^c}{}'\Theta_{\mathcal{H}_n^c}$ for $I_{\mathcal{H}_n^c\mathcal{H}_n} = E(F_iE_i{}')$, and $\hat{I}_{\mathcal{H}_n^c\mathcal{H}_n} = n^{-1}\Theta_{\mathcal{H}_n^c}{}'\Theta_{\mathcal{H}_n}$ for $I_{\mathcal{H}_n^c\mathcal{H}_n^c} = E(F_iF_i{}')$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

However, this estimator may not exist, because the matrix $\hat{I}_{\mathcal{H}_n^c\mathcal{H}_n^c}$ can be singular in high-dimensional settings. We follow the suggestion by Ning and Liu (2017) to adopt the Dantzig selector (Candes and Tao, 2007) to estimate the $(p_n - h_n)s_n \times h_n s_n$ unknown matrix w by column. Alternative procedures can also be used (not pursued here for brevity). Specifically, for each $l = 1, \ldots, h_n s_n$, we solve

$$\hat{w}_l \in \underset{w_l}{\operatorname{argmin}} ||w_l||_1 \quad \text{s.t.} \quad ||n^{-1}\sum_{i=1}^{n} E_{il}F_i{}' - w_l' n^{-1}\sum_{i=1}^{n} F_i F_i{}'||_\infty \leq \tau_n, \quad (3.9)$$

where τ_n is a common tuning parameter chosen using K-fold cross-validation, giving the resulting estimator $\hat{w}$. Therefore, we have the estimated decorrelated score function

$$\begin{aligned}\hat{S}(\eta) &= \hat{S}(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = n^{-1}\hat{\Lambda}_{\mathcal{H}_n}^{-1}(\hat{w}'\Theta_{\mathcal{H}_n^c}' - \Theta_{\mathcal{H}_n}')(Y - \Theta_{\mathcal{H}_n}\eta_{\mathcal{H}_n} - \Theta_{\mathcal{H}_n^c}\hat{\eta}_{\mathcal{H}_n^c}) \\ &= n^{-1}\textstyle\sum_{i=1}^{n}\hat{\Lambda}_{\mathcal{H}_n}^{-1}(\hat{w}'F_i - E_i)(Y_i - E_i'\eta_{\mathcal{H}_n} - F_i'\eta_{\mathcal{H}_n^c}), \qquad (3.10)\end{aligned}$$

where $\hat{\Lambda}_{\mathcal{H}_n}$ is invertible by Lemma 3 in the Supplementary Material. Then, we substitute in the estimator $\hat{\eta}$ obtained from minimizing (2.5) to construct the decorrelated score test statistic under the null hypothesis $H_0 : \|\beta_j\|_{L^2} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

$0,$ for all $j \in \mathcal{H}_n$, leading to

$$\hat{T}^* = n^{1/2}\hat{S}(0, \hat{\eta}_{\mathcal{H}_n^c}) = n^{-1/2}\sum_{i=1}^{n}\hat{S}_i, \;\; \hat{S}_i = \hat{\Lambda}_{\mathcal{H}_n}^{-1}(\hat{w}'F_i - E_i)(Y_i - F_i'\hat{\eta}_{\mathcal{H}_n^c}) \quad (3.11)$$

Note that the null hypothesis in (3.7) is of full generality with the dimension $h_n s_n$, where s_n grows with n (often at a fractional polynomial order) to approximate the infinite-dimensional functional spaces, and h_n can be as large as p_n. Unlike testing a finite-dimensional null hypothesis, it is difficult to find a tractable limiting distribution, even when testing a single functional predictor, $h_n = 1$. Hence, we use its infinity norm $||\hat{T}^*||_\infty = \max\{|\hat{T}_l^*| : l = 1, \ldots, h_n s_n\}$ to test against the null hypothesis in (3.7), and adopt a computationally efficient and theoretically guaranteed bootstrap method to approximate the limiting distribution of $||\hat{T}^*||_\infty$. Because a standard bootstrap is expensive as a result of repeatedly estimating η and w, we consider the multiplier bootstrap method proposed by Chernozhukov et al. (2014). Specifically, denote $\hat{T}_e^* = n^{-1/2}\sum_{i=1}^{n} e_i\hat{S}_i$, where $\{e_1, \ldots, e_n\}$ is a set of i.i.d. standard normal random variables independent of the data. Then, define

$$c_B(\alpha) = \inf\{t \in \mathbb{R} : P_e(||\hat{T}_e^*||_\infty \leq t) \geq 1 - \alpha\} \quad (3.12)$$

as the $100(1-\alpha)$th percentile of $||\hat{T}_e^*||_\infty$, where $P_e(\cdot)$ denotes the probability

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

with respect to $\{e_1, \ldots, e_n\}$. Based on this critical value, we reject the null hypothesis at the significance level α provided that $||\hat{T}^*||_\infty \geq c_B(\alpha)$. Furthermore, note that the vector $\hat{T}^*$ in $||\hat{T}^*||_\infty$ is nearly standardized, owing to the transformation $\hat{\Lambda}_{\mathcal{H}_n}^{-1}$ in (3.11). This is sensible because the multiplier bootstrap method indeed requires that the test statistics have comparative scaling. Theorem 2 states that under the null hypothesis and some mild conditions, the Kolmogorov distance between the distributions of $||\hat{T}^*||_\infty$ and $||\hat{T}_e^*||_\infty$ converges to zero as the sample size grows. This provides theoretical guarantees for the decorrelated score test based on the multiplier bootstrap method uniformly over all $\alpha \in (0, 1)$.

Theorem 2. *Under conditions (A1)–(A6) in Section 2 and (B1)–(B3) and (P1)–(P5) in Appendices A and B, respectively, and using the local minimizer $\hat{\eta}$ from Theorem 1, then under $H_0 : \|\beta_j\|_{L^2} = 0$, for all $j \in \mathcal{H}_n$, the Kolmogorov distance between the distributions of $||\hat{T}^*||_\infty$ and $||\hat{T}_e^*||_\infty$ satisfies*

$$\lim_{n\to\infty} \sup_{t\geq 0} \left| P(||\hat{T}^*||_\infty \leq t) - P_e(||\hat{T}_e^*||_\infty \leq t) \right| = 0$$

and, consequently, $\lim_{n\to\infty} \sup_{\alpha\in(0,1)} \left| P\{||\hat{T}^*||_\infty > c_B(\alpha)\} - \alpha \right| = 0.$

4. Exact equivalence to decorrelated Wald and likelihood ratio tests

Based on the decorrelation used in the score function in (3.8), we can construct the counterparts of other classical tests, such as the Wald and likelihood ratio

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

tests, for high-dimensional models (e.g., the Cox proportional hazard model) in which these tests can be shown asymptotically equivalent (Fang et al., 2016). In this section, we introduce the decorrelated Wald and likelihood ratio tests that can be shown to be exactly (not asymptotically) equivalent in the context of a large-scale FLR.

For the decorrelated Wald test, we adopt a one-step procedure based on the estimated decorrelated score function in (3.10) to find an estimator $\acute{\eta}_{\mathcal{H}_n}$ of $\eta_{\mathcal{H}_n}$, as follows:

$$\acute{\eta}_{\mathcal{H}_n} = \hat{\eta}_{\mathcal{H}_n} - \{\partial \hat{S}(\hat{\eta}_{\mathcal{H}_n}, \hat{\eta}_{\mathcal{H}_n^c})/\partial \eta_{\mathcal{H}_n}\}^{-1} \hat{S}(\hat{\eta}_{\mathcal{H}_n}, \hat{\eta}_{\mathcal{H}_n^c}). \tag{4.13}$$

Then, the decorrelated Wald test statistic is given by

$$\hat{W}^* = n^{1/2} \hat{\Lambda}_{\mathcal{H}_n}^{-1} \hat{I}_{\mathcal{H}_n|\mathcal{H}_n^c} \acute{\eta}_{\mathcal{H}_n}, \tag{4.14}$$

where $\hat{I}_{\mathcal{H}_n|\mathcal{H}_n^c} = \hat{I}_{\mathcal{H}_n\mathcal{H}_n} - \hat{w}' \hat{I}_{\mathcal{H}_n^c\mathcal{H}_n}$. Consequently, the decorrelated Wald test is such that we reject the null hypothesis in (3.7) at the significance level α if $||\hat{W}^*||_\infty \geq c_B(\alpha)$, where $c_B(\alpha)$ is the critical value defined in (3.12).

To define the decorrelated likelihood ratio test statistic, we begin with some assumptions and notation. Without loss of generality, assume that the index set $\mathcal{H}_n$ for the null hypothesis corresponds to the first h_n functional predictors (i.e., $\mathcal{H}_n = \{1, \ldots, h_n\}$), and rewrite the loss function $L_n(\eta)$ as $L_n(\eta) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

$L_n(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = L_n(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c})$, where $\eta_{\mathcal{H}_n}\backslash\eta_{jk}$ represents the vector that excludes η_{jk}. We introduce the following negative decorrelated partial likelihood function $L_{jk}(\eta)$ for each η_{jk}, for $j = 1, \ldots, h_n$ and $k = 1, \ldots, s_n$:

$$\begin{aligned} L_{jk}(\eta) &= L_{jk}(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = L_{jk}(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c}) \\ &= L_n(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c} - \eta_{jk} w_{(j-1)s_n+k}) \\ &= \frac{1}{2n}||Y - \Theta_{\mathcal{H}_n}\eta_{\mathcal{H}_n} - \Theta_{\mathcal{H}_n^c}(\eta_{\mathcal{H}_n^c} - \eta_{jk} w_{(j-1)s_n+k})||_2^2, \quad (4.15) \end{aligned}$$

where $w_{(j-1)s_n+k}$ is the $\{(j-1)s_n+k\}$th column of the matrix $w = I_{\mathcal{H}_n^c\mathcal{H}_n^c}^{-1} I_{\mathcal{H}_n^c\mathcal{H}_n}$. Note that $E\{\partial L_{jk}(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c})/\partial\eta_{jk}\nabla_{\eta_{\mathcal{H}_n^c}} L_n(\eta)\} = 0$ uniformly in $j = 1, \ldots, h_n$ and $k = 1, \ldots, s_n$. The estimated version of $L_{jk}(\eta)$ is

$$\begin{aligned} \hat{L}_{jk}(\eta) &= \hat{L}_{jk}(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = \hat{L}_{jk}(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c}) \\ &= L_n(\eta_{jk}, \eta_{\mathcal{H}_n}\backslash\eta_{jk}, \eta_{\mathcal{H}_n^c} - \eta_{jk} \hat{w}_{(j-1)s_n+k}) \\ &= \frac{1}{2n}||Y - \Theta_{\mathcal{H}_n}\eta_{\mathcal{H}_n} - \Theta_{\mathcal{H}_n^c}(\eta_{\mathcal{H}_n^c} - \eta_{jk} \hat{w}_{(j-1)s_n+k})||_2^2, \quad (4.16) \end{aligned}$$

where $\hat{w}_{(j-1)s_n+k}$ is obtained from (3.9). To implement this test, we also need an estimator $\grave{\eta}_{jk}$ for each η_{jk} that approximately minimizes $\hat{L}_{jk}(\eta_{jk}, 0, \hat{\eta}_{\mathcal{H}_n^c})$ with respect to η_{jk}. Unlike Fang et al. (2016), who used $\acute{\eta}_{\mathcal{H}_n}$ from the decorrelated Wald test, we again employ a one-step estimator $\grave{\eta}_{jk}$ based on the fact

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

that $\partial \hat{L}_{jk}(\grave{\eta}_{jk}, 0, \hat{\eta}_{\mathcal{H}_n^c})/\partial \eta_{jk}$ is close to zero; that is,

$$\grave{\eta}_{jk} = -\{\partial^2 \hat{L}_{jk}(0, 0, \hat{\eta}_{\mathcal{H}_n^c})/\partial \eta_{jk}^2\}^{-1}\{\partial \hat{L}_{jk}(0, 0, \hat{\eta}_{\mathcal{H}_n^c})/\partial \eta_{jk}\}. \tag{4.17}$$

Denote $\hat{\Upsilon}$ as an $h_n s_n \times 1$ vector with $\{(j-1)s_n + k\}$th element equal to $2n\{\hat{L}_{jk}(0, 0, \hat{\eta}_{\mathcal{H}_n^c}) - \hat{L}_{jk}(\grave{\eta}_{jk}, 0, \hat{\eta}_{\mathcal{H}_n^c})\}$. Then, the decorrelated likelihood ratio test statistic is given by

$$\hat{L}^* = \hat{\Lambda}_{\mathcal{H}_n}^{-2} \text{diag}\{(\Theta_{\mathcal{H}_n^c}\hat{w} - \Theta_{\mathcal{H}_n})'(\Theta_{\mathcal{H}_n^c}\hat{w} - \Theta_{\mathcal{H}_n})/n\}\hat{\Upsilon}, \tag{4.18}$$

with the same critical value $c_B(\alpha)$ as that in (3.12) for a level-α test. The exact equivalence between the three proposed tests for the large-scale FLR is established in Theorem 3, where $\hat{W}^*$ and $\hat{L}^*$ denote the decorrelated Wald and likelihood ratio statistics, as in (4.14) and (4.18), respectively.

Theorem 3. *Under conditions (A1)–(A6) in Section 2 and (B1)–(B3) and (P1)–(P5) in Appendices A and B, respectively, and using the local minimizer $\hat{\eta}$ from Theorem 1, then under $H_0 : \|\beta_j\|_{L^2} = 0$ for all $j \in \mathcal{H}_n$, one has $||\hat{T}^*||_\infty = ||\hat{W}^*||_\infty = ||\hat{L}^*||_\infty^{1/2}$.*

We conclude this section by pointing out that the exact (not asymptotic) equivalence between these three tests under the general null hypothesis in (3.7)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

occurs because we use one-step estimators in the Wald and likelihood ratio statistics and in the linear structure of the FLR model. Hence, it suffices to focus on, for instance, the decorrelated score test only.

5. Simulation Studies

The simulated data $\{y_i, i = 1, \ldots, n\}$ are generated from the following model:

$$y_i = \sum_{j=1}^{p_n} \int_0^1 \beta_j(t) x_{ij}(t) dt = \sum_{j=1}^{p_n} \sum_k \eta_{jk} \theta_{ijk} + \epsilon_i,$$

with $n = 100$ subjects and $p_n = 200$ functional predictors, where the errors $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. from $N(0, \sigma^2)$. The functional predictors have mean zero and a covariance function derived from the Fourier basis $\phi_1 = 1$, $\phi_{2\ell} = 2^{1/2} \cos\{\ell\pi(2t-1)\}$, for $\ell = 1, \ldots, 25$, and $\phi_{2\ell-1} = 2^{1/2} \sin\{(\ell-1)\pi(2t-1)\}$, for $\ell = 2, \ldots, 25$, $t \in T = [0, 1]$. The underlying regression function is $\beta_j(t) = \sum_{k=1}^{50} \eta_{jk} \phi_k(t)$, for $j \leq q_n = 3$, where $\eta_{jk} = c_j(1.2 - 0.2k)$ for $k \leq 4$, and $\eta_{jk} = 0.4 c_j (k-3)^{-4}$ for $5 \leq k \leq 50$, with constants $\{c_j : j \leq q_n\}$ chosen for different settings, and the other $\beta_j(t) = 0$, for all $t \in T$. To generate $X_{ij}(t)$, for $j = 1, \ldots, p_n$, define $V_{ij}(t) = \sum_{k=1}^{50} \tilde{\theta}_{ijk} \phi_k(t)$, where $\{\tilde{\theta}_{ijk}\}$ follows an independently distributed $N(0, k^{-2})$ for different i and j. The p_n functional

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

predictors are then defined using the autoregressive relationship,

$$X_{ij}(t) = \sum_{j'=1}^{p_n} \rho^{|j-j'|} V_{ij'}(t) = \sum_{k=1}^{50} \sum_{j'=1}^{p_n} \rho^{|j-j'|} \tilde{\theta}_{ij'k} \phi_k(t) = \sum_{k=1}^{50} \theta_{ijk} \phi_k(t),$$

where $\theta_{ijk} = \sum_{j'=1}^{p_n} \rho^{|j-j'|} \tilde{\theta}_{ij'k}$, and the constant $\rho \in (0, 1)$ controls the correlations between the functional predictors; here, we present the case of $\rho = 0.3$. For the observed measurements, we take discrete realizations of $\{X_{ij}(\cdot), j = 1, \ldots, p_n\}$ at 100 equally spaced times $\{t_{ijl}, l = 1, \ldots, 100\} \in T$. Next, we use an orthonormal cubic spline basis to fit the model, where the tuning parameters s_n and λ_n are chosen using five-fold cross-validation and the algorithm with the SCAD penalty (see Appendix A). Then, we construct the decorrelated score test statistic and its associated $\alpha = 5\%$ empirical quantile using a wild bootstrap with $N = 10000$ bootstrap samples. Table 1 summarizes the empirical sizes and powers under the null and several alternative hypotheses in different settings specified by $\{c_j : j \leq q_n\}$, based on the rejection proportion over 500 Monte Carlo replicates. The computation takes between two and three minutes, on average, for each case in one Monte Carlo run.

From Table 1, the rejection proportions of the first 11 null hypotheses increase quickly as the signal of β_1 increases with $\beta_j = 0$, for $j \geq 2$, which is expected for a power function curve. In addition, the rejection proportion un-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Table 1: Simulation results for different settings of the regression curves $\{\beta_j : j \leq q_n\}$ specified by $\{c_j : j \leq q_n\}$, under various hypotheses measured over 500 Monte Carlo replicates, where $n = 100, p_n = 200, q_n = 3, \rho = 0.3$, and $\sigma^2 = 1$. Shown are the empirical rejection proportions with corresponding standard errors in parentheses. In particular, the rejection rates in the first 11 rows depict the pattern of the power function under $H_0 : \|\beta_1\|_{L^2} = 0$ with ascending signal strength in β_1, while $\|\beta_j\|_{L^2} = 0$ for $j \geq 2$. This is followed by testing different hypotheses H_0, when the underlying $\|\beta_j\|_{L^2} \neq 0$, for $j = 1, 2, 3$. Note that in all settings, $\|\beta_j\|_{L^2} = 0$, for $j \geq 4$.

Setting of $\{\beta_j : j \leq 3\}$	$H_0 : \|\beta_j\|_{L^2} = 0, j \in \mathcal{H}_n$	Rejection proportion
$c_1 = 0, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.046 (.009)
$c_1 = .1, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.086 (.013)
$c_1 = .2, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.300 (.021)
$c_1 = .3, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.574 (.022)
$c_1 = .4, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.752 (.019)
$c_1 = .5, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.894 (.014)
$c_1 = .6, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.948 (.100)
$c_1 = .7, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.974 (.007)
$c_1 = .8, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.988 (.005)
$c_1 = .9, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	.996 (.003)
$c_1 = 1, c_2 = 0, c_3 = 0$	$\mathcal{H}_n = \{1\}$	1.00 (.000)
$c_1 = 1, c_2 = 1, c_3 = 1$	$\mathcal{H}_n = \{1\}$	.986 (.005)
$c_1 = 1, c_2 = 1, c_3 = 1$	$\mathcal{H}_n = \{1, \ldots, 5\}$	1.00 (.000)
$c_1 = 1, c_2 = 1, c_3 = 1$	$\mathcal{H}_n = \{1, \ldots, 20\}$	1.00 (.000)
$c_1 = 1, c_2 = 1, c_3 = 1$	$\mathcal{H}_n = \{5, \ldots, 20\}$	.050 (.010)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

der the first null hypothesis is, as expected, close to the prespecified significance level $\alpha = 5\%$. Among the last four null hypotheses, which include larger sets of regression parameter functions, the proposed test has a rejection rate close to the significance level $\alpha = 5\%$ when the nonzero β_j all reside in the alternative parameter space (i.e., the last null hypothesis), and possesses good power for testing the other three null hypotheses. We repeated the experiments using different settings of n, p_n, q_n, ρ, and σ^2 , finding similar patterns with descending power for larger values of p_n and σ^2 and ascending power for larger n. The influence of ρ on the power is not as noticeable as that of p_n and σ^2, whereas the influence of q_n is mainly associated with the hypothesis of interest. These similar results are not reported here.

6. Real-Data Example

We analyze a data set on 848 individuals from the Human Connectome Project (HCP); see `http://www.humanconnectome.org/` for more information on the HCP. The response of interest is a continuous score called *Emotion Task Shape Acc*, calculated from emotion-processing fMRI tasks. These tasks are related to the brain processing of negative emotions such as fear or anger; a detailed description is available at `https://www.humanconnectome.org/storage/app/media/documentation/s500/hcps500meg2releas`

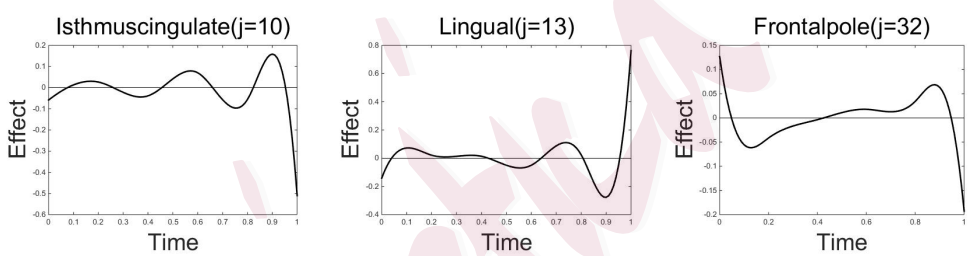
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

`ereferencemanual.pdf`. There are 35 regions of the brain (e.g., lingual, paracentral, isthmuscingulate etc.). Here, we are interested in identifying those regions that have a significant effect when processing negative emotional tasks. Thus, we have $p_n = 35$ functional predictors, where the fMRI readings for each functional predictor are recorded at 176 equally spaced time points, rescaled to a unit interval. Previous studies have shown that three regions, the isthmuscingulate (Rockstroh and Elbert, 2010), lingual (Goldin et al., 2008), and frontalpole (Musha et al., 1997), are responsible for negative emotions. Thus, it is of keen interest to pick out these crucial regions from the study.

We adopt an orthonormal cubic B-spline basis, and fit the large-scale FLR with the number of inside knots $k_n = s_n - 4$ and the tuning parameter λ_n chosen using five-fold cross-validation. As a result of the regularized estimation, 17 of the 35 regression functions are retained in the model. To perform hypothesis testing on the importance of these regions, we first conduct a marginal test for each individual region using the proposed decorrelated score test statistic. From the results, we reject the null hypotheses for isthmuscingulate $(j = 10, p = .0028)$, lingual $(j = 13, p = .0007)$, and frontalpole $(j = 32, p = .0346)$ at a significance level of 0.05. Based on the marginal significance, we carry out an overall test for $H_0 : \|\beta_j\|_{L^2} = 0$, for all $j \notin \{10, 13, 32\}$, and fail to reject this null hypothesis at level 0.05 with a p-value of 0.2725. This indicates that the other

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

functional predictors are not statistically important. Therefore, it is reasonable to retain these three regions in our model (i.e., isthmuscingulate $(j = 10)$, lingual $(j = 13)$, and frontalpole $(j = 32)$). To further justify their significance, we re-fit the FLR model using these three predictors only, and conduct a marginal test for each of the three regions. We find that all three marginal tests are rejected at level 0.05, with p-values of 0.0138, 0.0174, and 0.0203, respectively. This indicates that a model with these three regions may not be reduced further. In terms of computation, the proposed method takes around eight minutes.



(a) (b) (c)

Figure 1: The estimated regression coefficient functions obtained from the FLR model containing three functional predictors corresponding to the isthmuscingulate $(j = 10)$, lingual $(j = 13)$, and frontalpole $(j = 32)$ regions, respectively.

The estimated regression parameter functions for the three regions are displayed in Figure 1. From the left panel, it appears that the negative emotion is periodically associated with the isthmuscingulate region over the entire duration, and becomes more influential over time. This is consistent with the finding of Rockstroh and Elbert (2010) that the isthmuscingulate region is responsible for

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

negative emotions such as fear. Figure 1(b) shows that the effect of the lingual region appears neutral before $t = 0.7$ on the re-scaled unit time scale, but becomes stronger thereafter, supporting the finding of Goldin et al. (2008) of an association between the lingual region and negative emotion. In Figure 1(c), the effect of the frontalpole region varies from negative to positive on the response. This pattern agrees with the finding of Musha et al. (1997) that the frontalpole region is associated with emotions in the change of mood from happiness to sadness. Note that caution is required when interpreting the regression functions, especially for the estimates near the beginning and end times, owing to a boundary effect.

Appendix

A. Nonconvex penalty and algorithm

Without loss of generality, we assume that the data are centered so that we have $n^{-1}\sum_{i=1}^{n} Y_i = 0$ and $n^{-1}\sum_{i=1}^{n} \theta_{ijk} = 0$, for any $j = 1, \ldots, p_n$, $k = 1, \ldots, s_n$. In addition, for each $j = 1, \ldots, p_n$, we denote $\hat{f}_j = \Theta_j \hat{\eta}_j$, where $\hat{\eta}_j$ is an estimator of η_j, and $U_j = \Theta_j(\Theta_j{}'\Theta_j)^{-1}\Theta_j{}'$. The optimization of (2.5) can be achieved by adopting the coordinate descent method similar to those used in Ravikumar et al. (2008) and Fan et al. (2015) with slight modification, where $\rho_{\lambda_n}(\cdot)$ is replaced by $\rho_{\lambda_n s_n^{1/2}}(\cdot)$. For completeness, we restate below a general class of non-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

convex penalty functions ρ_λ satisfying the technical conditions (P1)–(P5) as in Loh and Wainwright (2015).

(P1) ρ_λ is an even function, and $\rho_\lambda(0) = 0$.

(P2) For $t \geq 0$, $\rho_\lambda(t)$ is nondecreasing in t.

(P3) $g_\lambda(t) = \rho_\lambda(t)/t$ is nonincreasing in t, for $t > 0$.

(P4) $\rho_\lambda(t)$ is differentiable except at $t = 0$, $\lim_{t\to 0^+} \rho'_\lambda(t) = \lambda L$, for some positive constant L.

(P5) $\rho_{\lambda,\mu}(t)$ is convex in t, for some positive constant μ, where $\rho_{\lambda,\mu}(t) = \rho_\lambda(t) + 2^{-1}\mu t^2$.

It is known that most nonconvex regularizers, e.g., LASSO, SCAD and MCP, meet those conditions, and Lemma 1 in the online supplement studies the properties of those penalty functions. Then, we provide a fitting algorithm for the large-scale FLR by slightly modifying that of Ravikumar et al. (2008).

(i) Start with the initial estimator $\hat{f}_j = 0$, for each $j = 1, \ldots, p_n$.

(ii) Caculate the residual $R_j = Y - \sum_{k\neq j} \hat{f}_k$, while fixing the values of $\{\hat{f}_k : k \neq j\}$.

(iii) Caculate the $\hat{P}_j = U_j R_j$.

(iv) Let $\hat{f}_j = \max\left\{1 - \rho'_{\lambda_n s_n^{1/2}}(n^{-1/2}||\hat{f}_j||_2)n^{1/2}/||\hat{P}_j||_2, 0\right\}\hat{P}_j$.

(v) Let $\hat{f}_j = \hat{f}_j - n^{-1}1_n{}'\hat{f}_j 1_n$, where 1_n denotes the $n \times 1$ vector of ones.

(vi) Repeat (ii) to (v) for $j = 1, \ldots, p_n$ and iterate until convergence to obtain the final estimates $\hat{f}_j$, for $j = 1, \ldots, p_n$.

(vii) Compute $\hat{\eta}_j = (\Theta_j{}'\Theta_j)^{-1}\Theta_j{}'\hat{f}_j$ by using the final estimates $\hat{f}_j$ from step (vi) to get the final estimates $\hat{\eta}_j$, for $j = 1, \ldots, p_n$.

B. Conditions on the large-scale FLR model

Next we quantify the relationship among the parameters q_n, s_n, R_n and the sample size n, which is needed for establishing the estimation consistency in Theo-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

rem 1. Recall that q_n is the number of significant predictors, R_n is a parameter such that $||\eta^*||_1 \leq R_n$ where η^* represents the true value of η. Then we assume

(B1) $\max(n^{\beta} q_n^2 s_n^{a+1-\delta}, n^{2\beta} q_n^2 s_n^{a/2+1-\delta}, n^{5\beta/2-1/2} R_n q_n s_n^{a/2+1}, n^{3\beta/2-1/2} R_n q_n s_n^{a+1},$
$n^{\beta+1/2} q_n s_n^{-\delta} \log s_n) = o(1).$

In particular, since $||\eta^*||_1 = \sum_{j=1}^{q_n} \sum_{k=1}^{s_n} |\eta_{jk}^*| = O(q_n)$ under (A6) and (B1), it is feasible to assume $R_n \sim q_n$ in practice. We provide two concrete examples to illustrate (B1) as follows:

• If $R_n \sim q_n \sim c$ for some constant $c > 0$, then (B1) is reduced to

$$\max(n^{\beta} s_n^{a+1-\delta}, n^{2\beta} s_n^{a/2+1-\delta}, n^{5\beta/2-1/2} s_n^{a/2+1}, n^{3\beta/2-1/2} s_n^{a+1}, n^{\beta+1/2} s_n^{-\delta} \log s_n) = o(1). \tag{B.19}$$

It is easy to check that there exists s_n satisfying (B.19) if $\min\{(2\delta-a-2)/(4\beta), (\delta-a-1)/\beta, (2\delta-2)/(2\beta+1)\} > \max\{(a+2)/(1-5\beta), (2a+2)/(1-3\beta)\}$.

• If $R_n \sim q_n \sim s_n$, then (B1) is reduced to

$$\max(n^{\beta} s_n^{a+3-\delta}, n^{2\beta} s_n^{a/2+3-\delta}, n^{5\beta/2-1/2} s_n^{a/2+3}, n^{3\beta/2-1/2} s_n^{a+3}, n^{\beta+1/2} s_n^{1-\delta} \log s_n) = o(1), \tag{B.20}$$

and s_n satisfies (B.20), if $\min\{(2\delta-a-6)/(4\beta), (\delta-a-3)/\beta, (2\delta-2)/(2\beta+1)\} > \max\{(a+6)/(1-5\beta), (2a+6)/(1-3\beta)\}$.

Next, we denote $\rho_n = \sup_{l \leq h_n s_n} ||w_l||_0$, where $||w_l||_0$ is the number of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

nonzero elements in the lth column of w. Now we quantify the relationship between ρ_n and various other parameters, which is needed in Theorem 2.

(B2) $\max\{n^{3\beta/2}\rho_n q_n s_n^{3a/2-\delta}\log s_n, n^{5\beta/2-1/2}\rho_n q_n^2 s_n^{2a+1-\delta}, n^{2\beta-1/2}(\log n)^{1/2}\rho_n s_n^{3a/2}, n^{3\beta-1}R_n\rho_n q_n s_n^{2a+1}\} = o(1)$.

Note that the order of ρ_n is determined by the relative orders of parameters q_n, s_n and R_n. For instance, if the predictors in $\mathcal{H}_n$ are uncorrelated with nuisance predictors, (B2) holds trivially. If $\rho_n \sim c$ and $R_n \sim q_n \sim s_n$, then (B1) entails (B2). We also impose some conditions on the tuning parameters λ_n and τ_n in the regularizers in (2.5) and in the Dantizig method (3.9), respectively.

(B3) $n^{\beta}q_n s_n^{a+1} = o(\lambda_n^{-1})$, $n^{2\beta}q_n s_n^{a/2+1} = o(\lambda_n^{-1})$, $n^{5\beta/2-1/2}\rho_n q_n s_n^{2a+1} = o(\lambda_n^{-1})$, $n^{\beta/2-1/2}R_n = o(\lambda_n)$, $q_n s_n^{-\delta} = o(\lambda_n)$, $\tau_n \sim \{\log(p_n s_n)/n\}^{1/2}$.

In particular, if $R_n \sim q_n$, then (B3) implies that $\max\{(\log p_n/n)^{1/2}, q_n s_n^{-\delta}\} \leq \lambda_n \leq R_n^{-1}$, which is consistent with the assumption (6) of Theorem 1 in Loh and Wainwright (2015). By combining (A5), (B1) with (B3), one has $\tau_n \sim n^{\beta/2-1/2}$.

Supplementary Material

The auxiliary lemmas used to show the main theorems, as well as the proofs of those lemmas and theorems, are deferred to the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Acknowledgments

Fang Yao's research was partially supported by the National Natural Science Foundation of China Grants 11931001 and 11871080, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education. Kaijie Xue's research was partially supported by the National Natural Science Foundation of China Grant 11871080, the National Natural Science Foundation of China Grant 11901313, Fundamental Research Funds for the Central Universities, Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, and Key Laboratory of Pure Mathematics and Combinatorics, Ministry of Education. The authors would like to thank the associate editor and the two referees for their insightful comments.

References

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*, 1705–1732.

Cai, T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*, 2159–2179.

Cai, T. and M. Yuan (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association 107*, 1201–1216.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

Cardot, H., F. Ferraty, A. Mas, and P. Sarda (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics. Theory and Applications 30*, 241–255.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters 45*, 11–22.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics 42*, 1564–1597.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical Statistics.* CRC Press.

Cuevas, A., M. Febrero, and R. Fraiman (2002). Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics 30*, 285–300.

Escabias, M., A. M. Aguilera, and M. J. Valderrama (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics 16*, 365–384.

Fan, J. and R. Li (2001, December). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*, 5467–5484.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *The*

Annals of Statistics 42, 819–849.

Fan, J. and J.-T. Zhang (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B 62*, 303–322.

Fan, Y., G. M. James, and P. Radchenko (2015). Functional additive regression. *The Annals of Statistics 43*, 2296–2325.

Fang, E. X., Y. Ning, and H. Liu (2016). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society, Series B*. DOI: 10.1111/rssb.12224.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics 39*, 254–261.

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*, 1356–1378.

Goldin, P. R., K. McRae, W. Ramel, and J. J. Gross (2008). The neural bases of emotion regulation: Reappraisal and suppression of negative emotion. *Biological Psychiatry 63*, 577–586.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*, 70–91.

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

Hilgert, N., A. Mas, and N. Verzelen (2013). Minimax adaptive tests for the functional linear model. *The Annals of Statistics 41*, 838–869.

Kong, D., K. Xue, F. Yao, and H. H. Zhang (2016). Partially functional linear regression in high dimensions.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Biometrika 103, 147–159.

Lei, J. (2014). Adaptive global testing for functional linear models. *Journal of the American Statistical Association 109*, 624–634.

Lian, H. (2011). Functional partial linear model. *Journal of Nonparametric Statistics 23*, 115–128.

Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statistica Sinica 23*, 51–74.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. *The Annals of statistics 42*, 413–468.

Loh, P.-L. and M. J. Wainwright (2015). Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research 71*, 559–616.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society, Series B 72*, 417–473.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics 37*, 246–270.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *The Annals of Statistics 33*, 774–805.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Associa-*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

tion 103, 1534–1544.

Musha, T., Y. Terasaki, H. A. Haque, and G. A. Ivanitsky (1997). Feature extraction from eegs associated with emotions. *Artificial Life and Robotics 1*, 15–19.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics 45*, 158–195.

Ramsay, J. O. and C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B 53*, 539–572.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second ed.). Springer Series in Statistics. New York: Springer.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2008). Sparse additive models. *Journal of the Royal Statistical Society, Series B 71*, 1009–1030.

Reiss, P. T. and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association 102*, 984–996.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B 53*(1), 233–243.

Rockstroh, B. and T. Elbert (2010). Traces of fear in the neural web — magnetoencephalographic responding to arousing pictorial stimuli. *International Journal of Psychophysiology 78*, 14–16.

Shah, R. D. and R. J. Samworth (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, Series B 75*, 55–80.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

Shang, Z. and G. Cheng (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics 43*, 1742–1773.

Taylor, J., R. Lockhart, R. J. Tibshirani, and R. Tibshirani (2014). Post-selection adaptive inference for least angle regression and the lasso. arXiv:1401.3889.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*, 1166–1202.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*, 614–645.

Wang, L., Y. Kim, and R. Li (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics 41*, 2505–2536.

Wang, Z., H. Liu, and T. Zhang (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics 42*, 2164–2201.

Wasserman, L. and K. Roeder (2009). High-dimensional variable selection. *The Annals of Statistics 37*, 2178–2201.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional linear regression analysis for longitudinal dataegression analysis for longitudinal data. *The Annals of Statistics 33*, 2873–2903.

Yuan, M. and T. T. Cai (2010). A reproducing kernel Hilbert space approach to functional linear regression.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0456

The Annals of Statistics 38, 3412–3444.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*, 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B 76*, 217–242.

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

Zhang, T. (2009). Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics 37*, 2109–2144.

Zhu, H., F. Yao, and H. H. Zhang (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society, Series B 76*, 581–603.

Kaijie Xue, School of Statistics and Data Science, Nankai University, Tianjin 300071, China

E-mail: kaijie@nankai.edu.cn

Fang Yao, corresponding author, Department of Probability and Statistics, School of Mathematical Sciences, Center for Statistical Science, Peking University

E-mail: fyao@math.pku.edu.cn