Statistica Sinica **34** (2024), 679-697
doi:https://doi.org/10.5705/ss.202022.0077

INFERENCE ON LARGE-SCALE PARTIALLY FUNCTIONAL LINEAR MODEL WITH HETEROGENEOUS ERRORS

Kaijie Xue and Fang Yao*

Shanghai University of International Business and Econimics and Peking University

Abstract: We investigate a partially functional linear model by focusing on the heterogenous error scenario in which the scalar response is associated with an ultra-large number of both functional predictors and scalar covariates. Moreover, the model does not require the standard condition on eigenvalue decay for functional predictors, leading to a more challenging and general framework. The target is to establish a rigorous inferential procedure for hypothesis testing on an arbitrary subset of both regression functions and scalar coefficients. Specifically, we devise a confidence region for post-regularization inference using a pseudo score function that is not decorrelated owing to the heterogenous errors. The proposed test does not require estimation consistency of the functional part, and is shown to be uniformly convergent to the prescribed significance. We investigate the finite-sample performance of the proposed model using simulation studies and an application to functional magnetic resonance imaging brain image data.

Key words and phrases: Eigenvalue-decay-relaxation, high dimensions, multiplier bootstrap.

1. Introduction

The classical functional linear model (FLM) is often used to model the linear association between a continuous response Y and a functional predictor that is often supposed 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 FLM takes the simple form,

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

where both Y_i and X_i are generally assumed to have zero means; that is, $EY_i = 0$ and $EX_i(t) = 0$ for $t \in T$, the errors ϵ_i that are independent of X_i are required to be i.i.d. with mean zero and finite variance $E(\epsilon_i^2) = \sigma^2 \in (0, \infty)$, and the unknown regression parameter function $\beta(t)$ is square-integrable such that $\beta \in L^2(T)$. This model is investigated thoroughly by numerous works in functional data analysis (e.g., Ramsay and Dalzell (1991); Fan and Zhang (2000); Yuan and Cai

*Corresponding author.

(2010)), from the perspectives of either statistical inference (Cardot et al. (2003); Hilgert, Mas and Verzelen (2013); Lei (2014); Shang and Cheng (2015)) or a theoretical construct (Hall and Horowitz (2007); Cai and Yuan (2012)). In addition, the classical FLM has been extended to a variety of settings, including the additive regression (Müller and Yao (2008); Zhu, Yao and Zhang (2014); Fan, James and Radchenko (2015)), partially FLM (Kong et al. (2016)), and large-scale FLM (Xue and Yao (2021)), among others.

In modern applications, the response Y can potentially be linked to a large number of scalar covariates and a large number of functional predictors, simultaneously. For example, Kong et al. (2016) propose a penalized estimation and variable selection procedure for a partially FLM comprising a finite number of functional predictors and high-dimensional scalar covariates, and Xue and Yao (2021) consider hypothesis testing on a large-scale FLM involving only high-dimensional functional predictors. Nevertheless, in the context of large-scale data and a partially FLM, the potential numbers of functional predictors p_n and scalar covariates d_n can far exceed the sample size n, despite the sparsity assumption that the sizes of the significant functional predictors q_n and significant scalar covariates r_n grow at a fraction polynomial rate of n. For instance, in a neuroimage analysis, a certain disease marker is potentially associated with a large number of brain regions of interest (ROIs), scanned over time, in addition to high-dimensional scalar covariates, including age, sex, and so forth. Furthermore, in practice, the errors ϵ_i may not be as homogeneous as they are in the classical FLM. Motivated by these concerns, denoting $Z_i = (Z_{i1}, \ldots, Z_{id_n})^\top$ as a vector of scalar covariates and $\gamma = (\gamma_1, \ldots, \gamma_{d_n})^\top$ as a vector of coefficients, we can formulate a large-scale partially FLM with heterogeneous errors ($LPFLM_{hete}$) as

$$\begin{aligned} Y_i &= \sum_{j=1}^{p_n} \int_T X_{ij}(t)\beta_j(t)dt + Z_i^\top \gamma + \epsilon_i \\ &= \sum_{j=1}^{p_n} \int_T X_{ij}(t)\beta_j(t)dt + \sum_{l=1}^{d_n} Z_{il}\gamma_l + \epsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{1.2}$$

where both p_n and d_n are allowed to grow exponentially with the sample size n, and, without loss of generality, assume the first q_n regression functions $\{\beta_j : j = 1 \ldots, q_n\}$ and the first r_n regression coefficients $\{\gamma_l : l = 1 \ldots, r_n\}$ are significant, whereas the rest are zero. The errors ϵ_i are heterogeneous, with mean zero and possibly different variances and distributions. Note that the $LPFLM_{hete}$ differs from previous works, such as those on the partially FLM (Kong et al. (2016)) and the large-scale FLM (Xue and Yao (2021)), by allowing p_n and d_n to grow exponentially, simultaneously, in the context of heterogeneous errors.

The basis representation of each predictor X_{ij} can be achieved using either a pre-fixed basis (i.e., B-splines, wavelets) or a data-driven basis (i.e.,

eigenfunctions). Despite the efficiencies of data-driven bases, they need to be estimated from p_n separate functional principal component analysis procedures, which is computationally intensive whenever $p_n \gg n$. As a tradeoff, researches use a common pre-fixed basis $\{b_k : k \geq 1\}$ that is complete and orthonormal in $L^2(T)$ to represent all random processes X_{ij}'s as suggested in Xue and Yao (2021). Thus, we do not consider other basis-seeking procedures, such as the functional partial least squares (Reiss and Ogden (2007)).

Our main contribution is to develop a confidence region for an arbitrary subset of regression functions and scalar coefficients $\{\beta_j : j \leq p_n\} \cup \{\gamma_l : l \leq d_n\}$, leading to a rigorous inferential procedure for a general hypothesis on that subset. Here, we face three main challenges. The first arises from the complex inter-correlation between the ultrahigh-dimensional functional and scalar parts, where both p_n and d_n can grow exponentially in n. Although the partially FLR (and its variants) has been well studied, such as in Kong et al. (2016), these methods usually require the number of functional predictors to be fixed at p in order to control its inter-correlation with the scalar part, and thus no inferential procedure is available. (Xue and Yao (2021)) consider a general testing procedure based on ultrahigh-dimensional p_n functional predictors. However, the model fails to include an ultrahigh-dimensional scalar part, and neither a confidence region nor a power assessment is provided. The second challenge stems from the heterogeneous error assumption under the $LPFLM_{hete}$, which means all existing works, such as Xue and Yao (2021), require *i.i.d.* errors in the presence of high-dimensional functional predictors. In fact, it is nontrivial to extend the ordinary high-dimensional linear model to its heterogeneous counterpart. Similarly to Xue and Yao (2021), an important ingredient for carrying out inference on the $LPFLM_{hete}$ is a penalized estimator $\{\hat{\beta}_j : j \leq p_n\} \cup \{\hat{\gamma}_l : l \leq d_n\}$. The third challenge is that we do not require estimation consistency for the estimated regression curves $\{\hat{\beta}_j : j \leq p_n\}$, in contrast to Xue and Yao (2021), making our method more applicable in practical situations, but more difficult to derive.

The rest of the article is organized as follows. In Section 2, we first introduce a penalized estimator $\{\hat{\beta}_j : j \leq p_n\} \cup \{\hat{\gamma}_l : l \leq d_n\}$ under a broad class of convex or nonconvex penalties. We then establish the estimation consistency for both $\{\hat{\gamma}_l : l \leq d_n\}$ and a scaled-version of $\{\hat{\beta}_j : j \leq p_n\}$ in Theorem 1. In Section 3, we first establish a confidence region for a general hypothesis in Theorem 2, which results in the proposed test. Then, we propose an estimated power function for the test, and analyze it in Theorems 3 and 4. In Sections 4 and 5, we present a simulation study and a real-data analysis to demonstrate the desired performance of the proposed inferential method. We collect the conditions on the $LPFLM_{hete}$ in Appendix A. The algorithm to obtain the penalized estimator and the conditions on the penalties are summarized in Appendix B. We relegate the auxiliary lemmas, with their proofs, and the proofs of the main theorems to the online Supplementary Material.

2. Model Estimation Using Group-Regularized Least Squares

Consider the $LPFLM_{hete}$ defined in model (1.2). Given a complete and orthonormal basis $\{b_k : k \geq 1\}$, the infinite-dimensional basis representations of each random process X_{ij} and the regression function β_j have the following form:

$$X_{ij} = \sum_{k=1}^{\infty} \theta_{ijk} b_k, \quad \beta_j = \sum_{k=1}^{\infty} \eta_{jk} b_k,$$

where the projected coefficients $\theta_{ijk} = \int_T X_{ij}(t) b_k(t) dt$ are random variables with mean zero and variance $E(\theta_{ijk}^2) = \omega_{jk} > 0$. We regard the eigenvalues of the jth functional predictor as the eigenvalues of the covariance structure, $\int_T cov(\nu_{ij}(t))dt = diag\{\omega_{j1}, \omega_{j2}, \dots\}$, where $\nu_{ij}(t)$ is infinite-dimensional and denoted by $\nu_{ij}(t) = (\theta_{ij1} b_1(t), \theta_{ij2} b_2(t), \dots)'$, for notational convenience. Accordingly, model (1.2) can be rewritten as

$$Y_i = \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \theta_{ijk} \eta_{jk} + Z_i^\top \gamma + \epsilon_i, \quad i = 1, \dots, n, \tag{2.1}$$

indicating that estimating the regression functions β_j and parameters γ_l is equivalent to estimating the unknown coefficients η_{jk} and γ_l, respectively. However, it is prohibitive to minimize the square loss with respect to η_{jk} and γ_l directly, owing to the infinite dimensionality of the sequence η_{jk}. A common way of addressing this problem is to approximate the model by truncating up to the first s_n leading bases that can grow in n, where s_n governs the complexity of β_j to balance the bias–variance tradeoff, analogously to the way it dose in a classical nonparametric regression. As a result, model 2.1 can be reformulated as

$$Y_i = \sum_{j=1}^{p_n} \sum_{k=1}^{s_n} \theta_{ijk} \eta_{jk} + Z_i^\top \gamma + \left(\epsilon_i + \sum_{j=1}^{p_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right). \tag{2.2}$$

Similar approaches can be found in Rice and Silverman (1991), Yao, Müller and Wang (2005), Hall and Horowitz (2007), Fan, James and Radchenko (2015), Kong et al. (2016), and Xue and Yao (2021), among others. Although we wish to apply a distinct truncation size to each β_j, it is computationally improper to select various truncations in the presence of a large number of functional predictors. As suggested by Kong et al. (2016), adopting a common s_n is sufficient for theoretical development and implementation.

In addition to truncation, it is important to regularize the regression functions and the scalar coefficients on a comparable scale. To regularize each functional predictor X_j, similar to the group regularization (Yuan and Lin (2006)), we impose a suitable group penalty on the scaled term $n^{-5/9} \|\Theta_j \eta_j\|_2$, where $\Theta_j = (\theta_{ijk})_{1 \leq i \leq n; 1 \leq k \leq s_n}$, $\eta_j = (\eta_{j1}, \dots, \eta_{js_n})^\top$, and $\|\cdot\|_2$ is the ℓ_2-norm. To

regularize each scalar covariate Z_l, we impose the same penalty on the scaled term $n^{-5/9}(\sum_{i=1}^n Z_{il}^2)^{1/2}|\gamma_l|$, using a technique similar to that of (Fan and Li (2001)). Hence, our goal is to minimize the regularized square loss function with respect to (η, γ), as follows, where $\eta = (\eta_1', \ldots, \eta_{p_n}')'$ and $\|\cdot\|_1$ is the ℓ_1-norm:

$$\min_{\|\eta\|_1+\|\gamma\|_1\le B_n} \overbrace{\Bigg\{\underbrace{(2n)^{-1}\sum_{i=1}^n\Bigg(Y_i-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\theta_{ijk}\eta_{jk}-Z_i^\top\gamma\Bigg)^2}_{L_n(\eta,\gamma)}+}^{Q_n(\eta,\gamma)}$$
$$\underbrace{\sum_{j=1}^{p_n}\rho_{\lambda_n}\big(n^{-5/9}\|\Theta_j\eta_j\|_2\big)+\sum_{l=1}^{d_n}\rho_{\lambda_n}\Bigg(n^{-5/9}\Bigg(\sum_{i=1}^n Z_{il}^2\Bigg)^{1/2}|\gamma_l|\Bigg)}_{P_{\lambda_n}(\eta,\gamma)}\Bigg\}, \tag{2.3}$$

where the univariate $\rho_\lambda(\cdot)$ that depends on a tuning parameter $\lambda > 0$ covers a broad class of commonly used convex or nonconvex penalties fulfilling the conditions (B1)–(B5) in Appendix B, such as the lasso, SCAD, and MCP (Loh and Wainwright (2015)). Here, the parameter $B_n > 0$ can take any value, provided that the true version (η^*, γ^*) is feasible in the sense that $\|\eta^*\|_1+\|\gamma^*\|_1 \le B_n$. Note that many existing works on high-dimensional linear models implicitly impose an upper bound on the l_∞- or l_1-norms of the linear coefficient; see Loh and Wainwright (2015), for example, with R playing the role of B_n. After obtaining any estimator $(\hat\eta, \hat\gamma)$ from (2.3), the corresponding estimator for each β_j takes the form $\hat\beta_j(t) = \sum_{k=1}^{s_n}\hat\eta_{jk}b_k(t)$. We use a coordinate descent algorithm modified from Ravikumar et al. (2008) for the implementation; the procedures are summarized in Appendix B, where the tuning parameters λ_n and s_n are evaluated using K-fold cross-validation, where, for example, $K = 5$. In our analysis, for general inferential purposes, it suffices to obtain a regularized estimator $(\hat\eta, \hat\gamma)$, the scaled version of which, $(\Lambda^{1/2}\hat\eta, \hat\gamma)$, as defined in Theorem 1, is consistent in both the ℓ_1- and ℓ_2-norms, while relaxing the consistency of the unscaled-version $(\hat\eta, \hat\gamma)$ and the oracle property. In contrast, other post-regularization inferential methods such as that of (Xue and Yao (2021)), are more restrictive, requiring estimation consistency for both $\hat\eta$ and the regression curves $\hat\beta_j$. Denoting the block matrix $\Lambda = diag\{\Lambda_1, \ldots, \Lambda_{p_n}\}$, with each $\Lambda_j = diag\{\omega_{j1}, \ldots, \omega_{js_n}\}$, Theorem 1 can be stated as follows, where the conditions (A1)–(A4) and (B1)–(B5) are relegated to Appendix A and B, respectively.

Theorem 1. *Under conditions* (A1)–(A4) *and* (B1)–(B5)*, for any local minima* $(\hat\eta, \hat\gamma)$ *of* $Q_n(\eta, \gamma)$ *obtained from* (2.3)*, we have the following with probability tending to one:*

1*)* $\max\{\|\Lambda^{1/2}(\hat\eta-\eta)\|_2, \|\hat\gamma-\gamma\|_2\} \le c_1\lambda_n(q_n+r_n)^{1/2}n^{-1/18}$*, for some constant* $c_1 > 0$.

2) $\max\{\|\Lambda^{1/2}(\hat{\eta}-\eta)\|_1, \|\hat{\gamma}-\gamma\|_1\} \leq c_2\lambda_n s_n^{1/2}(q_n+r_n)n^{-1/18}$, *for some constant* $c_2 > 0$.

Note that the aforementioned error bounds depend on the truncation s_n, through λ_n or itself, which is expected, because each regression function is represented by an s_n-dimensional vector η_j. Note too that Theorem 1 differs from and is more complicated to derive than its counterpart in Xue and Yao (2021), because we are examining both high-dimensional functional and scalar parts. In addition, Theorem 1 does not demand estimation consistency of the regression curves $\hat{\beta}_j(t)$, in contrast to all existing works on functional linear regression. As an illustrative example, if we further assume

$$|\hat{\eta}_{jk} - \eta_{jk}| \asymp \lambda_n \omega_{jk}^{-1/2} s_n^{-1/2} n^{-1/18} \quad \text{for} \quad j \leq q_n, k \leq s_n;$$
$$\lambda_n^2 n^{-1/9} s_n^{-1} \sum_{j=1}^{q_n}\sum_{k=1}^{s_n} \omega_{jk}^{-1} \to \infty$$

under Theorem 1, then we have $\sum_{j=1}^{p_n} \|\hat{\beta}_j - \beta_j\|_{L_2}^2 \to \infty$. Despite the possibly inconsistent regression curves $\hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t)$, Theorem 1 is sufficient for inferring the general hypothesis discussed in the next section. This significantly expands the range of application of our inferential method.

3. Inference on General Hypothesis in $LPFLM_{hete}$

Our goal is to infer a broad class of hypotheses of full generality in $LPFLM_{hete}$. Specifically, we write $\mathcal{P}_n = \{1, \ldots, p_n\}$ to represent all functional predictors, and similarly, $\mathcal{D}_n = \{1, \ldots, d_n\}$ to represent all scalar covariates. For any nonempty $\mathcal{H}_n \subseteq \mathcal{P}_n$ with cardinality $|\mathcal{H}_n| = h_n \leq p_n$, we write its complement as $\mathcal{H}_n^c = \mathcal{P}_n \setminus \mathcal{H}_n$. Likewise, for any nonempty $\mathcal{K}_n \subseteq \mathcal{D}_n$ with cardinality $|\mathcal{K}_n| = k_n \leq d_n$, we write its complement as $\mathcal{K}_n^c = \mathcal{D}_n \setminus \mathcal{K}_n$. Denoting the set $\mathcal{H}_n \times \mathcal{K}_n = \{(j,l) : j \in \mathcal{H}_n, l \in \mathcal{K}_n\}$, the general hypothesis takes the form

$$\begin{aligned} H_0 &: (\|\beta_j\|_{L^2}, |\gamma_l|) = (0,0) \text{ for all } (j,l) \in \mathcal{H}_n \times \mathcal{K}_n \\ \text{vs. } H_a &: (\|\beta_j\|_{L^2}, |\gamma_l|) \neq (0,0) \text{ for some } (j,l) \in \mathcal{H}_n \times \mathcal{K}_n, \end{aligned} \tag{3.1}$$

where the cardinalities h_n and k_n can be as large as the dimensions p_n and d_n, respectively, permitting a hypothesis of any size on $\{\beta_j : j = 1, \ldots, p_n\}$ and $\{\gamma_l : l = 1, \ldots, d_n\}$. Here, (3.1) contains two important degenerate hypotheses:

$$H_0 : \|\beta_j\|_{L^2} = 0 \text{ for all } j \in \mathcal{H}_n \quad \text{vs.} \quad H_a : \|\beta_j\|_{L^2} \neq 0 \text{ for some } j \in \mathcal{H}_n, \tag{3.2}$$

$$H_0 : \gamma_l = 0 \text{ for all } l \in \mathcal{K}_n \quad \text{vs.} \quad H_a : \gamma_l \neq 0 \text{ for some } l \in \mathcal{K}_n. \tag{3.3}$$

We focus on the inference of (3.1), without loss of generality.

To infer the general hypothesis (3.1), our idea is to develop a confidence region for $\{\beta_j : j \in \mathcal{H}_n\} \cup \{\gamma_l : l \in \mathcal{K}_n\}$ by combining the estimator $(\hat{\eta}, \hat{\gamma})$ from Theorem 1 with a pseudo score function. We refer to a pseudo score function as any generalization of the traditional score function that leads to the theoretical validity of the inferential method. For instance, the decorrelated score function used in Ning and Liu (2017) is a pseudo score function. Before proposing the pseudo score function for the $LPFLM_{hete}$, we discuss some notation. For functional predictors, we write the vector $\eta_{\mathcal{H}_n}$ as stacking $\{\eta_j : j \in \mathcal{H}_n\}$ in a column, and similarly for $\hat{\eta}_{\mathcal{H}_n}$. We abbreviate $\beta_{\mathcal{H}_n} = \{\beta_j : j \in \mathcal{H}_n\}$ as the sequence of regression functions. Given the fixed basis $\{b_k : k \geq 1\}$ and the truncation size s_n, we define the function $F_{\{b_k : k \leq s_n\}}(\beta_{\mathcal{H}_n}) = \eta_{\mathcal{H}_n}$ as mapping the regression curves $\beta_{\mathcal{H}_n}$ to the projection vector $\eta_{\mathcal{H}_n}$. For scalar covariates, we denote $\gamma_{\mathcal{K}_n}$ as restricting the vector γ to $\mathcal{K}_n$, and similarly for $\hat{\gamma}_{\mathcal{K}_n}$. For functional predictors, we denote its design matrix Θ as stacking $\{\Theta_j : j \leq p_n\}$ in a row, and similarly, the matrix $\Theta_{\mathcal{H}_n}$ as stacking $\{\Theta_j : j \in \mathcal{H}_n\}$ in a row. We denote the design matrix of the scalar covariates as $Z = (Z_1, \ldots, Z_n)'$, and similarly, the matrix $Z_{\mathcal{K}_n}$ as restricting the columns of Z to $\mathcal{K}_n$. Estimating each ω_{jk} by $\hat{\omega}_{jk} = n^{-1}\sum_{i=1}^{n} \theta_{ijk}^2$, we denote several diagonal matrices as $\Lambda = diag\{\Lambda_j : j \in \mathcal{P}_n\}$, $\Lambda_{\mathcal{H}_n} = diag\{\Lambda_j : j \in \mathcal{H}_n\}$, $\hat{\Lambda} = diag\{\hat{\Lambda}_j : j \in \mathcal{P}_n\}$, and $\hat{\Lambda}_{\mathcal{H}_n} = diag\{\hat{\Lambda}_j : j \in \mathcal{H}_n\}$, with submatrices $\Lambda_j = diag\{\omega_{j1}, \ldots, \omega_{js_n}\}$ and $\hat{\Lambda}_j = diag\{\hat{\omega}_{j1}, \ldots, \hat{\omega}_{js_n}\}$. We express design matrices in the form of row vectors as $(\Theta, Z) = (G_1, \ldots, G_n)'$, $(\tilde{\Theta}, Z) = (\Theta\Lambda^{-1/2}, Z) = (\tilde{G}_1, \ldots, \tilde{G}_n)'$, $(\breve{\Theta}, Z) = (\Theta\hat{\Lambda}^{-1/2}, Z) = (\breve{G}_1, \ldots, \breve{G}_n)'$, $(\Theta_{\mathcal{H}_n}, Z_{\mathcal{K}_n}) = (E_1, \ldots, E_n)'$, $(\tilde{\Theta}_{\mathcal{H}_n}, Z_{\mathcal{K}_n}) = (\Theta_{\mathcal{H}_n}\Lambda_{\mathcal{H}_n}^{-1/2}, Z_{\mathcal{K}_n}) = (\tilde{E}_1, \ldots, \tilde{E}_n)'$, $(\breve{\Theta}_{\mathcal{H}_n}, Z_{\mathcal{K}_n}) = (\Theta_{\mathcal{H}_n}\hat{\Lambda}_{\mathcal{H}_n}^{-1/2}, Z_{\mathcal{K}_n}) = (\breve{E}_1, \ldots, \breve{E}_n)'$, $(\Theta_{\mathcal{H}_n^c}, Z_{\mathcal{K}_n^c}) = (F_1, \ldots, F_n)'$, $(\tilde{\Theta}_{\mathcal{H}_n^c}, Z_{\mathcal{K}_n^c}) = (\Theta_{\mathcal{H}_n^c}\Lambda_{\mathcal{H}_n^c}^{-1/2}, Z_{\mathcal{K}_n^c}) = (\tilde{F}_1, \ldots, \tilde{F}_n)'$, and $(\breve{\Theta}_{\mathcal{H}_n^c}, Z_{\mathcal{K}_n^c}) = (\Theta_{\mathcal{H}_n^c}\hat{\Lambda}_{\mathcal{H}_n}^{-1/2}, Z_{\mathcal{K}_n^c}) = (\breve{F}_1, \ldots, \breve{F}_n)'$, where the invertibility of $\hat{\Lambda}$ is ensured by Lemma 2 in the Supplementary Material. Moreover, we denote $\tilde{\eta} = \Lambda^{1/2}\eta$, $\breve{\eta} = \hat{\Lambda}^{1/2}\eta$, $\tilde{\eta}_{\mathcal{H}_n} = \Lambda_{\mathcal{H}_n}^{1/2}\eta_{\mathcal{H}_n}$, and $\breve{\eta}_{\mathcal{H}_n} = \hat{\Lambda}_{\mathcal{H}_n}^{1/2}\eta_{\mathcal{H}_n}$. Recall from (2.3) the square loss function $L_n(\eta, \gamma) = (2n)^{-1}\|Y - \Theta\eta - Z\gamma\|_2^2 = (2n)^{-1}\|Y - \tilde{\Theta}\tilde{\eta} - Z\gamma\|_2^2$, with $Y = (Y_1, \ldots, Y_n)'$. Sometimes, we write $L_n(\tilde{\eta}, \gamma) = L_n(\eta, \gamma)$, with some abuse of notation. Although the true negative log-likelihood function is unknown under the heterogeneous error condition (A1.2) of (A1), it is rational to treat $L_n(\tilde{\eta}, \gamma)$ as a pseudo negative log-likelihood function. We define the matrix w as

$$w = \{E(F_i F_i')\}^{-1} E(F_i \tilde{E}_i') = (w_1, \ldots, w_{h_n s_n + k_n}) \in \mathbb{R}^{(p_n - h_n)s_n + (d_n - k_n) \times (h_n s_n + k_n)},$$

with column vectors $w_j = (w_{j1}, \ldots, w_{j,(p_n - h_n)s_n + (d_n - k_n)})'$. We denote the parameter $\rho_n = \sup_{j \leq h_n s_n + k_n} \rho_{nj}$, with each $\rho_{nj} = \|w_j\|_0$ as the cardinality of w_j, to measure the sparsity level of w. It is rational to impose a sparsity assumption on w, such as in condition (A4.4). For $LPFLM_{hete}$, we define the pseudo score function with respect to the primary parameters $(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ as follows:

$$S(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c}) = \nabla_{(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})} L_n(\tilde{\eta}, \gamma) - w' \nabla_{(\eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c})} L_n(\eta, \gamma)$$
$$= n^{-1} \sum_{i=1}^{n} \big(w' F_i - \tilde{E}_i\big) \left(Y_i - G_i' \begin{pmatrix} \eta \\ \gamma \end{pmatrix} \right). \tag{3.4}$$

However, the pseudo score function $S(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c})$ cannot be used directly because of the unknown w, $\Lambda_{\mathcal{H}_n}^{-1}$, $\eta_{\mathcal{H}_n^c}$, and $\gamma_{\mathcal{K}_n^c}$. We estimate $\Lambda_{\mathcal{H}_n}$ by its moment estimate $\hat{\Lambda}_{\mathcal{H}_n}$, the invertibility of which is ensured by Lemma 2 in the Supplementary Material. We estimate $(\eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c})$ by $(\hat{\eta}_{\mathcal{H}_n^c}, \hat{\gamma}_{\mathcal{K}_n^c})$, from Theorem 1. The moment estimate for $w = \{E(F_i F_i')\}^{-1} E(F_i \tilde{E}_i')$ is prohibitive, because $n^{-1} \sum_{i=1}^{n} F_i F_i'$ may be singular, owing to the high dimensionality. To resolve this problem, we estimate w by column-wisely solving the following lasso problems. Specifically, for each $j \leq h_n s_n + k_n$, we solve

$$\hat{w}_j = \underset{w_j}{\operatorname{argmin}} \left[(2n)^{-1} \sum_{i=1}^{n} (\breve{E}_{ij} - F_i' w_j)^2 + \lambda_n^* \| diag\{\hat{\Lambda}_{\mathcal{H}_n^c}^{1/2}, I_{(d_n - k_n)}\} w_j \|_1 \right], \tag{3.5}$$

where $\breve{E}_{ij}$ is the jth coordinate of $\breve{E}_i$, and the common tuning parameter λ_n^* is chosen using K-fold cross-validation, for example, $K = 5$. The above lasso procedures result in the estimator $\hat{w}$. Replacing w and $\tilde{E}_i$ in $S(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c})$ with $\hat{w}$ and $\breve{E}_i$, respectively, we obtain a new function $\hat{S}(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c})$, as

$$\hat{S}(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \eta_{\mathcal{H}_n^c}, \gamma_{\mathcal{K}_n^c}) = n^{-1} \sum_{i=1}^{n} \big(\hat{w}' F_i - \breve{E}_i\big) \left(Y_i - \breve{E}_i' \begin{pmatrix} \tilde{\eta}_{\mathcal{H}_n} \\ \gamma_{\mathcal{K}_n} \end{pmatrix} - F_i' \begin{pmatrix} \eta_{\mathcal{H}_n^c} \\ \gamma_{\mathcal{K}_n^c} \end{pmatrix} \right). \tag{3.6}$$

Plugging $\hat{\eta}_{\mathcal{H}_n^c}$ and $\hat{\gamma}_{\mathcal{K}_n^c}$ into the above function yields the estimated pseudo score function with respect to $(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$, as

$$\hat{S}(\tilde{\eta}_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}; \hat{\eta}_{\mathcal{H}_n^c}, \hat{\gamma}_{\mathcal{K}_n^c}) = n^{-1} \sum_{i=1}^{n} \big(\hat{w}' F_i - \breve{E}_i\big) \left(Y_i - \breve{E}_i' \begin{pmatrix} \tilde{\eta}_{\mathcal{H}_n} \\ \gamma_{\mathcal{K}_n} \end{pmatrix} - F_i' \begin{pmatrix} \hat{\eta}_{\mathcal{H}_n^c} \\ \hat{\gamma}_{\mathcal{K}_n^c} \end{pmatrix} \right), \tag{3.7}$$

which can be used to construct the test statistic for the general hypothesis (3.1). For notational convenience, we define a random function $\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ as

$$\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) = n^{1/2} \hat{S}(\hat{\Lambda}_{\mathcal{H}_n}^{1/2} F_{\{b_k : k \leq s_n\}}(\beta_{\mathcal{H}_n}), \gamma_{\mathcal{K}_n}; \hat{\eta}_{\mathcal{H}_n^c}, \hat{\gamma}_{\mathcal{K}_n^c}). \tag{3.8}$$

Recall that $F_{\{b_k : k \leq s_n\}}(\beta_{\mathcal{H}_n}) = \eta_{\mathcal{H}_n}$, where $\beta_{\mathcal{H}_n} = \{\beta_j : j \in \mathcal{H}_n\}$ denotes a sequence of regression functions. To obtain the critical value for an inference, we first define a random quantity $\hat{T}_e$ as

$$\hat{T}_e = n^{-1/2} \sum_{i=1}^{n} e_i \big(\hat{w}' F_i - \breve{E}_i\big) \left(Y_i - G_i' \begin{pmatrix} \hat{\eta} \\ \hat{\gamma} \end{pmatrix} \right),$$

where $e = \{e_1, \ldots, e_n\}$ is a set of i.i.d. standard normals, independent of the data. Denoting $\|\cdot\|_\infty$ as the infinite norm, we further define

$$c_B(\alpha) = \inf\{t \in \mathbb{R} : P_e(\|\hat{T}_e\|_\infty \le t) \ge 1 - \alpha\}, \quad \alpha \in (0,1) \tag{3.9}$$

as the $100(1-\alpha)$th percentile of $\|\hat{T}_e\|_\infty$, where $P_e(\cdot)$ means the probability with respect to $\{e_1, \ldots, e_n\}$ only. This serves as the critical value in the inference, and can be calculated rapidly using a multiplier bootstrap based on the set $e = \{e_1, \ldots, e_n\}$. Theorem 2 establishes the theoretical foundation for our proposed inferential method under some mild conditions. The conditions (A1)–(A4) and (B1)–(B5) are deferred to Appendices A and B.

Theorem 2. *Under conditions* (A1)–(A4) *and* (B1)–(B5), *the Kolmogorov distance between the distributions of* $\|\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})\|_\infty$ *and* $\|\hat{T}_e\|_\infty$ *satisfies*

$$\lim_{n\to\infty} \sup_{t\ge 0} \left| P(\|\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})\|_\infty \le t) - P_e(\|\hat{T}_e\|_\infty \le t) \right| = 0,$$

and, consequently, $\lim_{n\to\infty} \sup_{\alpha\in(0,1)} \left| P\{\|\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})\|_\infty \le c_B(\alpha)\} - (1-\alpha) \right| = 0.$

Recall that $\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ and $c_B(\alpha)$ are defined in (3.8) and (3.9), respectively. Theorem 2 defines the $100(1-\alpha)$th confidence region for $(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ as

$$CR_{1-\alpha} = \{(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) : \|\hat{T}(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})\|_\infty \le c_B(\alpha)\}. \tag{3.10}$$

Hence, our proposed testing procedure for (3.1) is such that we reject the null $H_0 : (\|\beta_j\|_{L^2}, |\gamma_l|) = (0,0)$ at the significance level $\alpha \in (0,1)$ if and only if

$$\|\hat{T}(0,0)\|_\infty > c_B(\alpha), \tag{3.11}$$

where $\|\hat{T}(0,0)\|_\infty$ is the test statistic, and $c_B(\alpha)$ serves as the critical value. Under Theorem 2, the proposed test based on the multiplier bootstrap method is valid uniformly over all $\alpha \in (0,1)$. Note that Theorem 2 is quite different from that of Xue and Yao (2021), not only because of the aforementioned heterogeneous errors, but also because we derive a general confidence region, rather than just the asymptotic null distribution, which facilitates the power analysis in Theorems 3–4. Furthermore, it follows from Theorems 1–2 that the proposed inferential method imposes no restriction on the decay rate of the eigenvalues ω_{jk}, as in (A2.1), whereas all existing FLMs demand $\omega_{jk} \gtrsim k^{-a}$ or $\lambda_{\min}(\Lambda) \gtrsim s_n^{-a}$, for some $a > 1$.

For the power assessment, given the true $(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$, the true power function is $Power(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) = P\{\|\hat{T}(0,0)\|_\infty > c_B(\alpha) | \beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}\}$. However, this is inaccessible, because the distribution of $\|\hat{T}(0,0)\|_\infty$ is intractable. The solution is to find an approximation of $Power(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ that is convenient to compute. Motivated by Theorem 2, we use another independent bootstrap procedure to

approximate the distribution of $\|\hat{T}(0,0)\|_\infty$, and hence the true power function. Specifically, the estimated power function $Power^*(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ can be formulated as

$$\begin{aligned} &Power^*(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) \\ &= P_{e^*}\left\{\|\hat{T}_{e^*} + n^{-1/2}\sum_{i=1}^{n}\big(\hat{w}'F_i - \breve{E}_i\big)E_i'\big(\tfrac{F_{\{b_k:k\le s_n\}}(\beta_{\mathcal{H}_n})}{\gamma_{\mathcal{K}_n}}\big)\|_\infty > c_B(\alpha)\right\}, \end{aligned} \tag{3.12}$$

where $e^* = \{e_1^*, \ldots, e_n^*\}$ is a set of *i.i.d.* standard normals, independent of the e used to compute $c_B(\alpha)$ and the data. Here, (3.12) can be computed using a multiplier bootstrap on e^*. In the next theorem, we establish the asymptotic equivalence between $Power(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ and $Power^*(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$.

Theorem 3. *Under conditions* (A1)–(A4) *and* (B1)–(B5)*, given the true version* $(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$*, we have:* $\lim_{n\to\infty}\big|Power(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) - Power^*(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})\big| = 0$.

For the power analysis, the following theorem establishes the consistency of the asymptotic power (3.12) under a fairly general alternative set $\mathcal{F}_n$.

Theorem 4. *Assume the conditions* (A1)–(A4) *and* (B1)–(B5) *hold, and that the true version* $(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n})$ *belongs to the alternative set*

$$\begin{aligned} \mathcal{F}_n = \Bigg\{(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) : \|\{E(\tilde{E}_i\tilde{E}_i') - w'E(F_i\tilde{E}_i')\}\begin{pmatrix}\Lambda_{\mathcal{H}_n}^{1/2}\eta_{\mathcal{H}_n}\\ \gamma_{\mathcal{K}_n}\end{pmatrix}\|_\infty \ge \\ K\rho_n(q_n + r_n)\left[\frac{\log\{n(p_n s_n + d_n)\}}{n}\right]^{1/2}\Bigg\}, \end{aligned}$$

where $K > 0$ *is a sufficiently large universal constant. Then, we have*

$$\lim_{n\to\infty} Power^*(\beta_{\mathcal{H}_n}, \gamma_{\mathcal{K}_n}) = 1.$$

4. Simulation Studies

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

$$\begin{aligned} Y_i &= \sum_{j=1}^{p_n}\int_0^1 \beta_j(t)X_{ij}(t)dt + \sum_{l=1}^{d_n} Z_{il}\gamma_l + \epsilon_i = \sum_{j=1}^{q_n}\int_0^1 \beta_j(t)X_{ij}(t)dt + \sum_{l=1}^{r_n} Z_{il}\gamma_l + \epsilon_i \\ &= \sum_{j=1}^{q_n}\sum_{k}\eta_{jk}\theta_{ijk} + \sum_{l=1}^{r_n} Z_{il}\gamma_l + \epsilon_i, \end{aligned}$$

with the high-dimensional setting $(n, p_n, d_n) = (100, 200, 200)$. To simulate the errors ϵ_i, we first generate $\{\delta_{ih} : i \le n, h \le 2\}$ as a set of mutually independent $U(1,2)$ random variables, and keep them fixed throughout the simulations. Then,

we consider three challenging error settings (I)–(III):

$$\text{(I)}\ \ \epsilon_i = (\delta_{i1}^2 + \delta_{i2}^2)^{1/2}\epsilon_{i1}, \quad \text{(II)}\ \ \epsilon_i = (\delta_{i1}^2 + \delta_{i2}^2)^{1/2}\epsilon_{i2}, \quad \text{(III)}\ \ \epsilon_i = \delta_{i1}\epsilon_{i1} + \delta_{i2}\epsilon_{i2},$$

where $\epsilon_{i1} \overset{i.i.d.}{\sim} (5/3)^{-1/2}t(5)$ and $\epsilon_{i2} \overset{i.i.d.}{\sim} 8^{-1/2}\{\chi^2(4) - 4\}$ have zero means and unit variances. The errors $\epsilon_1, \ldots, \epsilon_n$ are heterogeneous in all three settings. In setting (I), the errors are heavy tailed. The errors are skewed in setting (II). In setting (III), the errors are both heavy tailed and skewed. For the functional part, we define the nonzero regression curves as $\beta_j(t) = \sum_{k=1}^{50} \eta_{jk}\phi_k(t)$, for $j \leq q_n = 3$, where $\eta_{jk} = c_j(1.2 - 0.2k)$ for $k \leq 5$ and $\eta_{jk} = 0.2c_j(k-4)^{-3}$ for $6 \leq k \leq 50$, with the parameters $\{c_j : j \leq q_n\}$ chosen for different settings, and $\{\phi_k(\cdot) : k \geq 1\}$ is a complete orthonormal Fourier basis on $[0,1]$, with $\phi_1 = 1$, $\phi_{2\ell} = 2^{1/2}\cos\{\ell\pi(2t-1)\}$, for $\ell = 1, \ldots, 25$, and $\phi_{2\ell-1} = 2^{1/2}\sin\{(\ell-1)\pi(2t-1)\}$, for $\ell = 2, \ldots, 25$. To construct X_{ij}, for $j \leq p_n$, we first generate p_n *i.i.d.* functional predictors $\{W_{ij}(\cdot) : j \leq p_n\}$, defined on $[0,1]$ as

$$W_{ij}(t) = \sum_{k=1}^{50} \tilde{\xi}_{ijk}\phi_k(t),$$

where $\{\tilde{\xi}_{ijk}\}$ are independently distributed as $N(0, \delta_k^{-2})$. The sequence $\delta_1, \ldots, \delta_{50}$ is a random permutation of $1, \ldots, 50$, and is kept fixed throughout the simulations. The p_n functional predictors are defined using the following autoregressive relationship:

$$X_{ij}(t) = \sum_{j'=1}^{p_n} \rho_1^{|j-j'|}W_{ij'}(t) = \sum_{k=1}^{50}\sum_{j'=1}^{p_n} \rho_1^{|j-j'|}\tilde{\theta}_{ij'k}\phi_k(t) = \sum_{k=1}^{50} \theta_{ijk}\phi_k(t),$$

with each $\theta_{ijk} = \sum_{j'=1}^{p_n} \rho_1^{|j-j'|}\tilde{\theta}_{ij'k}$, where the parameter $\rho_1 \in (0,1)$ controls the correlation between the functional predictors. In the simulation, we present the case of $\rho_1 = 0.3$. In contrast to prior studies, we adopt a more challenging setting that requires no decaying pattern on the eigenvalues of X_{ij}, owing to the random permutation operation. For the observed measurements, we take discrete realizations of $\{X_{ij}(\cdot), j = 1, \ldots, p_n\}$ at $m = 100$ equally spaced times $\{t_{ijl}, l = 1, \ldots, 100\} \in \mathcal{T} = [0,1]$, and use an orthonormal cubic spline basis to fit the data. For the scalar part, we define the nonzero regression coefficients as $\gamma_l = c_{l+q_n}(-1)^{l-1}$, for $l \leq r_n$, where the parameters $\{c_{l+q_n} : l \leq r_n\}$ are chosen for different settings. To generate Z_{il} for $l \leq d_n$, we first simulate $V_{il} \overset{iid}{\sim} N(0,1)$. The d_n scalar covariates are then defined using autoregressive relationship,

$$Z_{il} = \sum_{l'=1}^{d_n} \rho_2^{|l-l'|}V_{il'} + \sum_{j=1}^{p_n} \rho_3^{|p_n-j+l|}\theta_{ij1},$$

where the parameter $\rho_2 \in (0,1)$ controls the correlation between the scalar covariates, and the parameter $\rho_3 \in (0,1)$ controls the correlation between the functional and the scalar parts. In the simulation, we present the case of $\rho_2 = \rho_3 = 0.3$. For the model fitting, the tuning parameters s_n and λ_n are selected using five-fold cross-validation and the algorithm given in Appendix B with the SCAD penalty. Given the selected s_n, the tuning parameter λ_n^* is further chosen using five-fold cross-validation based on the lasso regularization in (3.5). In terms of inference, we set the nominal level as $\alpha = 5\%$, and the resampling size as $N = 10{,}000$. In Table 1, we consider $(q_n, r_n) = (3, 6)$, and summarize the empirical size and power under the null and several alternative hypotheses in settings specified by $\{c_j : j \leq q_n + r_n\}$, based on the rejection proportion over 1,000 Monte Carlo simulations. We also consider the case of $(q_n, r_n) = (6, 12)$ in Table 2. The computation takes about four minutes for each case in one Monte Carlo run.

From Table 1, under error setting I, the rejection proportions under the null hypothesis $\mathcal{H}_n \times \mathcal{K}_n = \{4, \ldots, 6\} \times \{7, \ldots, 9\}$ are, as expected, close to the nominal level $\alpha = 5\%$. Moreover, the rejection proportions of the first four null hypotheses increase rapidly as we add significant functional predictors and scalar covariates into the null hypothesis, which is expected for a power function curve. As the intensity level increases from $0.4 \times \mathbf{1}_{1\times 9}$ to $0.6 \times \mathbf{1}_{1\times 9}$, the empirical power also increases rapidly, as expected. Similar patterns are observed for error settings II and III, indicating the proposed test is valid for various complex heterogeneous errors. The results in Table 2 share the same spirit as those in Table 1, further demonstrating the validity of the proposed test.

5. Real-Data Example

In this section, we apply our method to data on attention deficit hyperactivity disorder (ADHD), taken from the Attention Deficit Hyperactivity Disorder-200 Sample Initiative Project. ADHD is the most commonly diagnosed behavioral disorder in childhood, and can continue through adolescence and adulthood. Symptoms include lack of attention, hyperactivity, and impulsive behavior. The data set we use is the filtered preprocessed resting state data from the New York University Child Study Center, namely, the Anatomical Automatic Labeling (Tzourio-Mazoyer et al. (2002)) atlas, which contains $p_n = 116$ regions of interest, fractionated into 390 functional space using nearest-neighbor interpolation. After cleaning the raw data that failed the quality control, we include $n = 137$ individuals in the analysis.

The response of interest is the ADHD index, a continuous behavior score that reflects the severity of the disease. In the Anatomical Automatic Labeling atlas data, the mean gray scale in each region is calculated for 172 equally spaced time points. There are $d_n = 8$ scalar covariates of primary interest, including

Table 1. Simulation results for different settings of the regression curves $\{\beta_j : j \leq q_n\}$ and scalar coefficients $\{\gamma_l : l \leq r_n\}$ specified by $\{c_k : k \leq q_n + r_n\}$, under various error settings and hypotheses, measured over 1,000 Monte Carlo runs, where $(n, p_n, d_n, q_n, r_n) = (100, 200, 200, 3, 6)$. Shown are the empirical rejection proportions.

Error setting	$(c_1, \ldots, c_9)$	$H_0 : \mathcal{H}_n \times \mathcal{K}_n$	Rejection proportion
I	$0.4 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.046**
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.166
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.416
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.554
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.050**
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.386
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.812
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.918
II	$0.4 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.052**
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.192
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.410
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.544
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.054**
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.376
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.796
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.894
III	$0.4 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.042**
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.158
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.400
	$0.4 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.550
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{4, \ldots, 6\} \times \{7, \ldots, 9\}$	**0.048**
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{3, \ldots, 6\} \times \{5, \ldots, 9\}$	0.378
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{2, \ldots, 6\} \times \{3, \ldots, 9\}$	0.810
	$0.6 \times \mathbf{1}_{1\times 9}$	$\{1, \ldots, 6\} \times \{1, \ldots, 9\}$	0.924

gender (female/male), age, handedness (continuous between zero and one, where zero denotes totally left-handed, and one denotes totally right-handed), diagnosis status (categorical, with three levels: ADHD-combined, ADHD-inattentative, and Control as a baseline), medication status (yes/no), verbal IQ, performance IQ, and Full4 IQ. Our goal is to identify the important factors for the ADHD index from among these eight scalar covariates and 116 functional predictors. The model is given in (1.2). At the nominal level $\alpha = .05$, we first apply the proposed method to test the simple hypotheses $H_0 : \beta_j = 0$ and $H_0 : \gamma_l = 0$. These tests indicate that eight regression curves $(\beta_j : j = 1, 11, 41, 53, 68, 69, 70, 90)$ associated with the precentral, frontal, amygdala, occipital, supramarginal, precuneus and paracentral are significant, and that the two significant scalar coefficients $(\gamma_l : l = 4, 8)$ corresponding to diagnosis status and medication status are significant. These findings are reasonable, based on articles such as

Table 2. Simulation results for different settings of the regression curves $\{\beta_j : j \le q_n\}$ and scalar coefficients $\{\gamma_l : l \le r_n\}$ specified by $\{c_k : k \le q_n + r_n\}$, under various error settings and hypotheses, measured over 1000 Monte Carlo runs, where $(n, p_n, d_n, q_n, r_n) = (100, 200, 200, 6, 12)$. Shown are the empirical rejection proportions.

Error setting	$(c_1, \ldots, c_{18})$	$H_0 : \mathcal{H}_n \times \mathcal{K}_n$	Rejection proportion
I	$0.4 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.044**
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.388
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.782
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.860
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.041**
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.703
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.977
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.993
II	$0.4 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.042**
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.408
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.804
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.848
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.044**
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.714
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.972
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.984
III	$0.4 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.048**
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.398
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.774
	$0.4 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.856
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{7, \ldots, 9\} \times \{13, \ldots, 15\}$	**0.050**
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{5, \ldots, 9\} \times \{9, \ldots, 15\}$	0.708
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{3, \ldots, 9\} \times \{5, \ldots, 15\}$	0.978
	$0.6 \times \mathbf{1}_{1\times 18}$	$\{1, \ldots, 9\} \times \{1, \ldots, 15\}$	0.992

(Sasayama et al. (2010); Sidlauskaite et al. (2015)). Next, we consider the two general hypotheses $H_{01} : (\beta_j, \gamma_l) = (0, 0)$, for all $(j, l) \in \mathcal{H}_n \times \mathcal{K}_n$, and $H_{02} : (\beta_j, \gamma_l) = (0, 0)$, for all $(j, l) \in \mathcal{H}_n^c \times \mathcal{K}_n^c$, where $\mathcal{H}_n = \{1, 11, 41, 53, 68, 69, 70, 90\}$ and $\mathcal{K}_n = \{4, 8\}$. Because the proposed test rejects H_{01} and accepts H_{02}, it further supports our findings.

Appendix

A. Conditions on the $LPFLM_{hete}$

For the asymptotic properties, we require the following mild conditions (A1)–(A4). Condition (A1) is on the distribution type of the random variables from the model, which is comprised of parts (A1.1) and (A1.2).

Condition A1.

(A1.1) The random variables ϵ_i, $\omega_{jk}^{-1/2}\theta_{ijk}$, Z_{il}, $w_t{}'F_i$ are sub-Gaussian with mean zero and variance proxy σ^2 for some universal constant $\sigma > 0$, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, p_n$, $k = 1, \ldots, \infty$, $l = 1, \ldots, d_n$, $t = 1, \ldots, h_n s_n + k_n$.

(A1.2) The centered errors $\epsilon_1, \ldots, \epsilon_n$ are mutually independent (not necessarily identically distributed), and satisfy the moment condition:

$$n^{-1}\sum_{i=1}^{n} E(\epsilon_i^2) \geq c_1, \quad \text{for some universal constant} \quad c_1 > 0.$$

Note that (A1.1) only requires the sub-Gaussianity of random variables. Different from the underlying assumption of *i.i.d.* data $\{Y_i, (X_{ij} : j \leq p_n), Z_i, \epsilon_i\}_{i=1}^n$ in existing literature, the moment-type assumption (A1.2) takes one step further to allow for non i.i.d. responses, or heterogeneous errors. Condition (A2) specifies the smoothness and the covariance structure of the model.

Condition A2.

(A2.1) $\sup_{j \leq p_n} \sum_{k=1}^{\infty} \omega_{jk} < \infty$.

(A2.2) $\sup_{j \leq q_n} \sum_{k=1}^{\infty} \eta_{jk}^2 k^{2\delta} < \infty$, for some constant $\delta > 0$.

(A2.3) $\|\gamma\|_\infty < \infty$.

(A2.4) $c_1^{-1} \leq \lambda_{\min}(E(\tilde{G}_i \tilde{G}_i')) \leq \lambda_{\max}(E(\tilde{G}_i \tilde{G}_i')) \leq c_1$, for a constant $c_1 > 0$.

(A2.1) is the only smoothness condition on X_{ij}, which require the square integrability $\sup_{j \leq p_n} \int_T E(X_{ij}^2) dt < \infty$, without any decaying restriction on ω_{jk}. In contrast, all existing work demands either $\omega_{jk} - \omega_{j,k+1} \gtrsim k^{-a-1}$ or $\lambda_{\min}(\Lambda) \gtrsim s_n^{-a}$ for some $a > 1$. (A2.2) assumes the nonzero regression curves $\{\beta_j : j \leq q_n\}$ to belong to a Sobolev ball whose smoothness relies on a regularity constant $\delta > 0$. (A2.3) regulates the smoothness of scalar coefficients. (A2.4) specifies the covariance structure of the model, which is a standard condition in high-dimensional regression analysis. The relationship among the data dimensions, the truncation size and several critical parameters are given in condition (A3).

Condition A3.

(A3.1) $\log^9\{n(p_n s_n + d_n)\}/n \to 0$.

(A3.2) $\{s_n(q_n + r_n)B_n\}^{18/7}/n \to 0$.

(A3.3) $s_n^{2\delta}/[n q_n^2 \log^2\{n(p_n s_n + d_n)\}] \to \infty$.

Notice that (A3.1) permits both data dimensions p_n and d_n to grow exponentially in sample size, while (A3.2) reflects the sparsity of the model since both q_n and r_n are small in n. On one hand, (A3.2) requires the truncation size s_n to be relatively small in n to control the overall variation induced from the infinite-dimensional functional predictors. On the other hand, (A3.3) demands s_n to be relatively large in n to capture sufficient information for inference. Putting (A3.2) and (A3.3) together yields that $\delta > 9/7$, with δ defined in (A2). It follows from (A2) and (A3) that $\|\eta\|_1 + \|\gamma\|_1 \lesssim (q_n + r_n)$, suggesting the choice of $B_n \asymp (q_n + r_n)$ in practice based on the definition of B_n in (2.3). Condition (A4) quantifies the orders of tuning parameters λ_n and λ_n^*, and the sparsity parameter $\rho_n = \sup_{j \le h_n s_n + k_n} \|w_j\|_0$ of $w = \{E(F_i F_i')\}^{-1} E(F_i \tilde{E}_i') = (w_1, \ldots, w_{h_n s_n + k_n})$.

Condition A4.

(A4.1) $n^{-1/9} s_n (q_n + r_n)^2 \log^4\{n(p_n s_n + d_n)\} = o(\lambda_n^{-2})$.

(A4.2) $\max\{(s_n^{2\delta-1}/(n^{1/9} q_n^2))^{-1}, s_n B_n^2 / n^{7/9}\} = o(\lambda_n^2)$.

(A4.3) $K_1[\log\{n(p_n s_n + d_n)\}/n]^{1/2} \le \lambda_n^* \le K_2[\log\{n(p_n s_n + d_n)\}/n]^{1/2}$, for some sufficiently large universal constants $K_2 > K_1 > 0$.

(A4.4) $\rho_n^2 \log^5\{n(p_n s_n + d_n)\}/n \to 0, \rho_n^2 (q_n + r_n)^2 \log\{n(p_n s_n + d_n)\}/n \to 0$.

(A4.1) requires the tuning parameter λ_n in (2.3) to be small to retain important information about the significant predictors, while (A4.2) demands relatively larger λ_n to remove most irrelevant predictors via the regularization procedure. The order of the tuning parameter λ_n^* in (3.5) is specified in (A4.3). (A4.4) demands ρ_n to be small in n, corresponding to the sparseness assumption on w.

For a concrete example, under the high-dimensional setting $\{\log p_n \asymp \log d_n \asymp n^{1/20}; s_n \asymp q_n \asymp r_n \asymp \rho_n \asymp B_n \asymp n^{1/9}; \lambda_n \asymp n^{-13/60}; \lambda_n^* \asymp n^{-19/40}; \delta = 6\}$, conditions (A1)–(A13) hold simultaneously.

B. Regularizers and Algorithm

For technical convenience, the data are presumed centered such that $n^{-1}\sum_{i=1}^n Y_i = 0$, $n^{-1}\sum_{i=1}^n \theta_{ijk} = 0$ and $n^{-1}\sum_{i=1}^n Z_{il} = 0$, for all $j \le p_n$, $k \le s_n$, $l \le d_n$. We let $\hat{f}_j = \Theta_j \hat{\eta}_j$ and $U_j = \Theta_j (\Theta_j{}' \Theta_j)^{-1} \Theta_j{}'$ for each $j \le p_n$. For each $j = p_n + 1, \ldots, p_n + d_n$, we let $\hat{f}_j = V_{j-p_n} \hat{\gamma}_{j-p_n}$ and $U_j = V_{j-p_n} (V_{j-p_n}{}' V_{j-p_n})^{-1} V_{j-p_n}{}'$ with the vectors $V_l = (Z_{1l}, \ldots, Z_{nl})'$ for each $l \le d_n$. The optimization problem in (2.3) can be solved by a coordinate descent algorithm modified from Ravikumar et al. (2008) and Fan, James and Radchenko (2015). This algorithm is valid for a general class of regularizers ρ_λ as long as the technical conditions (B1)–(B5) below as in Loh and Wainwright (2015) are met. Specifically, for any $\lambda > 0$,

(B1) $\rho_\lambda(0) = 0$ and $\rho_\lambda(t) = \rho_\lambda(-t)$ for any $t \in \mathbb{R}$.

(B2) $\rho_\lambda(t)$ is nondecreasing in $t \in [0, \infty)$.

(B3) $g_\lambda(t) = t^{-1}\rho_\lambda(t)$ is nonincreasing in $t \in (0, \infty)$.

(B4) $\rho_\lambda(t)$ is differentiable at all $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t\to 0^+} \rho'_\lambda(t) = \lambda L$ for some constant $L > 0$.

(B5) For some constant $\mu > 0$, the function $\rho_{\lambda,\mu}(t) = \rho_\lambda(t) + 2^{-1}\mu t^2$ is convex in t.

Most of the convex or nonconvex penalties, e.g., LASSO, SCAD and MCP, satisfy the above conditions. Finally, the fitting algorithm for (2.3) is as follows.

Algorithm 1

(i) Initialize $\hat{f}_j = 0$, for every $j = 1, \ldots, p_n + d_n$.

(ii) Compute the residual $R_j = Y - \sum_{k\neq j} \hat{f}_k$, while keeping other $\{\hat{f}_k : k \neq j\}$ fixed.

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

(iv) Set $\hat{f}_j = \max\{1 - n^{4/9}\rho'_{\lambda_n}(n^{-5/9}||\hat{f}_j||_2)/||\hat{P}_j||_2, 0\}\hat{P}_j$.

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

(vi) Do (ii)–(v) for $j = 1, \ldots, p_n + d_n$ respectively and iterate until convergence to obtain the final estimators $\hat{f}_j$, for $j = 1, \ldots, p_n + d_n$.

(vii) Using the final estimators $\hat{f}_j$ from (vi), calculate the final estimators $\hat{\eta}_j = (\Theta_j{}'\Theta_j)^{-1}\Theta_j{}'\hat{f}_j$ for $j = 1, \ldots, p_n$, and $\hat{\gamma}_l = (V_l{}'V_l)^{-1}V_l{}'\hat{f}_{l+p_n}$ for $l = 1, \ldots, d_n$.

Supplementary Material

The auxiliary lemmas used to derive the theorems, as well as the proofs of those lemmas and theorems are contained in an online Supplementary Material.

Acknowldgments

This research was supported by the National Key R&D Program of China (No. 2020YFE0204200), National Natural Science Foundation of China (No. 12292981, 11931001, 12371268, 11901313), LMAM, the Fundamental Research Funds for the Central Universities and LMEQF, the KLMDASR, the LEBPS and the LPMC.

References

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

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

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

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

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

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

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

Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–159.

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

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

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.

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

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

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

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

Sasayama, D., Hayashida, A., Yamasue, H., Harada, Y., Kaneko, T., Kasai, K. et al. (2010). Neuroanatomical correlates of attention-deficit-hyperactivity disorder accounting for comorbid oppositional defiant disorder and conduct disorder. *Psychiatry and Clinical Neurosciences* **64**, 394–402.

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

Sidlauskaite, J., Caeyenberghs, K., Sonuga-Barke, E., Roeyers, H. and Wiersemaa, J. R. (2015). Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: Preserved global-disturbed local network organization. *NeuroImage: Clinical* **9**, 506–512.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N. et al. (2002). Automated anatomical labeling of activations in spm using a macroscopic 545 anatomical parcellation of the mni mri single-subject brain. *NeuroImage* **15**, 273–289.

Xue, K. and Yao, F. (2021). Hypothesis testing in large-scale functional linear regression. *Statistica Sinica* **31**, 1101–1123.

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

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics* **38**, 3412–3444.

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

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

Kaijie Xue

School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China.

E-mail: kaijie@utstat.toronto.edu

Fang Yao

School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing 100871, China.

E-mail: fyao@utstat.toronto.edu

(Received March 2022; accepted August 2022)