Statistica Sinica Preprint No: SS-2023-0356	
Title	Inference on Large-scale Generalized Functional Linear Model
Manuscript ID	SS-2023-0356
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0356
Complete List of Authors	Kaijie Xue and Riquan Zhang
Corresponding Authors	Riquan Zhang
E-mails	zhangriquan@163.com
Notice: Accepted author version.	

Statistica Sinica

Inference on Large-scale Generalized Functional Linear Model

Kaijie Xue and Riquan Zhang*

Shanghai University of International Business and Economics

**corresponding author: zhangriquan@163.com*

Abstract: In this work, we extend the classical generalized functional linear model to a large-scale generalized functional linear model to handle a variety of complex situations where the response (possibly discrete) can be nonlinearly linked to an ultra-high number of functional predictors. Unlike most existing requirements on functional data, we don't need to impose any conditions regarding eigenvalue-decay or square-integrability on those functional predictors, resulting in a more flexible but challenging model framework. Based on a penalized model estimator, we develop a general inferential method to assess the significance of an arbitrary group of regression curves. Concretely, a pseudo score function is adopted to construct the associated confidence region for the regression curves of interest. Notably, the proposed test is justified uniformly convergent to nominal level, without any demand on estimation consistency of the regression curves. Finally, numerical studies are carried out to show the empirical performance of the proposed test.

Key words and phrases: high dimensions, eigenvalue-decay-free, square-integrable-free, estimation-consistency-relaxed, multiplier bootstrap.

1. Introduction

A series of work (Ramsay and Dalzell, 1991; Yuan and Cai, 2010; Malfait and Ramsay, 2003; Fan and Zhang, 2000; Cardot et al., 1999) have been devoted to the study of classical functional linear model (FLM) containing a single functional predictor, focusing on either theoretical basis (Hall and Horowitz, 2007; Cai and Yuan, 2012; Ramsay and Silverman, 2005) or inferential methods (Cardot et al., 2003; Shang and Cheng, 2015a; Lei, 2014; Hilgert et al., 2013; Zhang and Chen, 2007). As an important extension of FLM, the generalized functional linear model (GFLM) has been frequently employed to model the more complicated (possibly nonlinear) association between a response Y and a functional predictor $X(t) \in L^2(T)$, where the random process $X(\cdot)$ is defined and square-integrable on a compact subset $T \subseteq \mathbb{R}$. This model has been intensively studied by many articles (e.g., Müller and Stadtmüller, 2005; Shang and Cheng, 2015a; Escabias et al., 2004). Concretely, given a sample of n i.i.d pairs $\{Y_i, X_i(\cdot)\}$, the conditional density of the classical GFLM under the commonly-used canonical link belong to an exponential family, which takes the simple form:

$$f(y_i|X_i, \beta, \phi) = \exp\left\{\frac{y_i\tilde{\delta}_i - b(\tilde{\delta}_i)}{a(\phi)} + c(y_i, \phi)\right\}, \qquad i = 1, \ldots, n \tag{1.1}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are known functions, and ϕ is the dispersion parameter. The linear predictor $\tilde{\delta}_i = \alpha_0 + \int_T X_i(t)\beta(t)dt$, where α_0 is the intercept. The random processes X_i are set as mean zero, i.e., $EX_i(t) = 0$ for all $t \in T$, and the regression functions $\beta(t)$ are assumed square-integrable satisfying $\beta \in L^2(T)$. The errors denoted by $\epsilon_i = Y_i - E(Y_i|X_i) = Y_i - b'(\tilde{\delta}_i)$ are independent of each other, having mean zero and conditional variances var$(\epsilon_i|X_i) = a(\phi)b''(\tilde{\delta}_i)$. Moreover, the classical GFLM has been generalised to settings allowing for a finite number of functional predictors (e.g., Li and Zhu, 2020; Xiao et al., 2021).

Nevertheless, when a GFLM is considered under large-scale setting, the number of predictors p_n can be potentially much greater than the sample size n, in spite of the parsimonious assumption that the sparsity level $q_n = card\{j : \beta_j \neq 0\}$ is of a fraction logarithmic order of n. As an illustrative example, the illness status of a certain psychiatric disorder may rely on a small portion of a great many potential brain areas, where the electroencephalography data measured over time can be obtained for each area. Besides the curse of low dimensionality, another drawback of the classical GFLM is that it demands its linear subcase (FLM) to have homogeneous errors ϵ_i, whereas it is common to encounter heterogeneous errors in practice. To address these shortcomings, the conditional density function of a large-scale generalized functional linear model with a possibly heterogeneous linear subcase ($LGFLM_{hete}$) is formulated as:

$$f_i(y_i|X_i,\beta,\phi_i) = \exp\left\{\frac{y_i\delta_i - b(\delta_i)}{a_i(\phi_i)} + c_i(y_i,\phi_i)\right\}, \quad i = 1,\dots,n \tag{1.2}$$

with the linear predictor $\delta_i = \alpha_0+\sum_{j=1}^{p_n}\int_T X_{ij}(t)\beta_j(t)dt$, where the data dimension p_n is permitted to grow exponentially in n. For technical convenience, we set the leading q_n regression curves $\{\beta_j : j = 1\dots,q_n\}$ nonzero. Notably, under the linear subcase of model (1.2) with $b(t) = 2^{-1}t^2$, it is not hard to verify that the errors $\epsilon_i = Y_i - E(Y_i|X_i) = Y_i - b'(\delta_i)$ can be heterogeneous with mean zero and possibly different variances and distributions, due to the reason that $a_i(\phi_i)$ may vary in i. Like other literature on GLM, we assume $\max_{i\leq n} a_i(\phi_i) < \infty$ throughout the paper, without loss of generality. Notice that the canonical link function of (1.2) is given by $g(\cdot) = (b')^{-1}(\cdot)$, which is completely determined by the function $b(\cdot)$. Apparently, the $LGFLM_{hete}$ largely extends our previous work Xue and Yao (2021) by permitting the framework of a large-scale generalized linear model with a possibly heterogeneous linear subcase. Indeed, extending the standard high-dimensional linear model to accommodate heterogeneity or nonlinearity is not a straightforward task, to say nothing of functional data. In terms of the basis representations of those random processes X_{ij}, we adopt a common pre-fixed orthornormal and complete basis $\{b_k : k \geq 1\}$ on T, instead of seeking data-driven bases (e.g., FPCA) which are computationally intensive especially in the context of $p_n \gg n$.

In this article, we concentrate on constructing a confidence region for any given subset of regression curves $\{\beta_j : j \leq p_n\}$, giving rise to a valid inferential method on the general hypothesis regarding that subset. The difficulty of developing the method comes from five major challenges/advantages. The first challenge is due to the more flexible framework of a generalized linear model, coupled with the rather complicated correlation structure of a large-scale number of functional predictors, which permits exponentially growing p_n. In comparison, existing work on the classical GFLM (and its variants) is more stringent requiring a fixed data dimension p. Our previous work Xue and Yao (2021) did propose a testing procedure regarding ultra-high functional predictors, but only the linear framework is considered and it fails to cover a confidence region. The second challenge is the allowance of heterogeneous errors under the linear subcase of $LGFLM_{hete}$, whereas all current articles on GFLM demand $i.i.d.$ errors for their linear subcases. The third challenge is due to the "square-integrable-free" advantage of our $LGFLM_{hete}$ that permits any j-th functional predictor to satisfy $\int_T E(X_{ij}^2)dt = \infty$, whereas all existing literature on GFLM strictly demands $\int_T E(X_{ij}^2)dt < \infty$ for every j. The fourth challenge is due to the "eigenvalue-decay-free" advantage that sets $LGFLM_{hete}$ free from any restrictions on the decay of eigenvalues of functional data, whereas current works on GFLM either require explicit decay restrictions on eigenvalue-gap (e.g., $\omega_{jk} - \omega_{j,k+1} \gtrsim k^{-a-1}$ for some $a > 1$) or impose implicit decay constraints to ensure a bounded exponential expectation in Bahadur representation for the functional data (e.g.,Assumption A4 in Shang and Cheng, 2015b). For post-regularization inference on $LGFLM_{hete}$, it is common to construct the procedure based on a penalized estimator $\{\hat{\beta}_j : j \leq p_n\}$ such as in Xue and Yao (2021). The fifth challenge is due to the advantage of "estimation-consistency-relaxed" since no estimation consistency of estimated curves $\{\hat{\beta}_j : j \leq p_n\}$ is needed to conduct inference in the framework of the $LGFLM_{hete}$, which differs from most related works (e.g., Xue and Yao, 2021).

The organization of the remainder of the paper is as follows. Section 2 first details a regularized estimator $\{\hat{\beta}_j : j \leq p_n\} \cup \{\hat{\alpha}_0\}$ under a wide range of either nonconvex or convex penalties, where $\hat{\alpha}_0$ is the estimated intercept. Then, the estimation consistency of a scaled-version of $\{\hat{\beta}_j : j \leq p_n\}$ is established in Theorem 1. In section 3, the confidence region of a general hypothesis is provided in Theorem 2, leading to the proposed testing procedure. The power analysis of the test are then given in Theorems 3–4. Simulation results in section 4 and real data analysis in section 5 demonstrate the effectiveness of the inferential procedure. Appendix A summarizes the conditions imposed on $LGFLM_{hete}$.

Appendix B first presents the assumptions on penalty function, and then the algorithm to get the regularized estimator. For space economy, the lemmas with proofs and the proofs of theorems are relegated to an online Supplement.

2. Group penalized model estimation

For the $LGFLM_{hete}$ model in (1.2), given a pre-fixed orthonormal and complete basis $\{b_k : k \geq 1\}$, the basis representations of the regression curves β_j and predictors X_{ij} can be expressed as

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

with the zero-mean random variables $\theta_{ijk} = \int_T X_{ij}(t) b_k(t) dt$, whose variance is denoted by $\omega_{jk} = E(\theta_{ijk}^2) > 0$. In this context, $\{\omega_{jk} : k \geq 1\}$ are regarded as eigenvalues of the functional predictor X_{ij}. Thus the linear predictors δ_i in (1.2) has the equivalent form

$$\delta_i = \alpha_0 + \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \theta_{ijk} \eta_{jk}, \qquad i = 1, \ldots, n \tag{2.3}$$

which entails the equivalence between each β_j and an infinite-dimensional sequence of coefficients $\{\eta_{jk} : k \geq 1\}$. Due to the infeasibility to directly estimate an infinite sequence, we adopt the commonly-used technique of truncation to approach the model (e.g., Kong et al., 2016; Xue, 2023; Xue et al., 2023; Fan et al., 2015; Hall and Horowitz, 2007; Yao et al., 2005; Rice and Silverman, 1991). Specifically, using a common truncation size s_n that increases in n, the linear predictor δ_i in (2.3) can be approximated by the truncated version δ_i^* as

$$\delta_i^* = \delta_i - \sum_{j=1}^{p_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk}\eta_{jk} = \alpha_0 + \sum_{j=1}^{p_n}\sum_{k=1}^{s_n} \theta_{ijk}\eta_{jk}. \tag{2.4}$$

After truncation, our next step is to regularize those truncated predictors, in the same spirit as the group penalization procedure (Yuan and Lin, 2006), where each predictor X_j is viewed as a group with dimension s_n. Specifically for each j-th predictor, penalization is implemented on the quantity $n^{-5/9}\|\Theta_j\eta_j\|_2$, with matrix $\Theta_j = (\theta_{ijk})_{1\le i\le n; 1\le k\le s_n}$ and vector $\eta_j = (\eta_{j1}, \ldots, \eta_{js_n})'$. To this end, denoting $\eta = (\eta_1', \ldots, \eta_{p_n}')'$, we solve the following optimization problem with respect to (η, α_0) to obtain a penalized estimator,

$$\min_{\|\eta\|_1+|\alpha_0|\le B_n} \overbrace{\left\{ \underbrace{n^{-1}\sum_{i=1}^{n}\{b(\delta_i^*) - Y_i\delta_i^*\}}_{L_n(\eta,\alpha_0)} + \underbrace{\sum_{j=1}^{p_n} \rho_{\lambda_n}\left(n^{-5/9}\|\Theta_j\eta_j\|_2\right)}_{P_{\lambda_n}(\eta)} \right\}}^{Q_n(\eta,\alpha_0)} \tag{2.5}$$

in which the penalty function $\rho_\lambda(\cdot)$ that relies on a regularization parameter $\lambda > 0$ can be a large number of popular candidates including SCAD, MCP and

LASSO (Loh and Wainwright, 2015), provided that assumptions (B1)-(B5) are met in Appendix B. The upper bound B_n is allowed to be any positive real number so that the genuine value (η^*, α_0^*) is feasible (i.e., $\|\eta^*\|_1 + |\alpha_0^*| \leq B_n$). Based on any solution $(\hat{\eta}, \hat{\alpha}_0)$ of (2.5), the penalized estimator for β_j can be expressed as $\hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t)$. An algorithm adapted from Ravikumar et al. (2008) is provided in Appendix B to solve the optimization problem (2.5). In implementation, we tune the parameters s_n and λ_n through cross-validation. Before proceeding, we write a diagonal matrix as $\Lambda = diag\{\Lambda_1, \ldots, \Lambda_{p_n}\}$ with submatrices $\Lambda_j = diag\{\omega_{j1}, \ldots, \omega_{js_n}\}$. Notably, to conduct inference, it is adequate to use a penalized estimator $(\hat{\eta}, \hat{\alpha}_0)$ whose scaled-form $(\Lambda^{1/2}\hat{\eta}, \hat{\alpha}_0)$ satisfies estimation consistency, thus relaxing both estimation and selection consistencies on $(\hat{\eta}, \hat{\alpha}_0)$ and hence the curves $\hat{\beta}_j$. As a result, we present Theorem 1 below to establish the estimation consistency of the scaled-form $(\Lambda^{1/2}\hat{\eta}, \hat{\alpha}_0)$ under mild conditions (A1)-(A5) from Appendix A and (B1)-(B5) from Appendix B.

Theorem 1. *Under conditions (A1)–(A5) and (B1)–(B5), for any local minima $(\hat{\eta}, \hat{\alpha}_0)$ of $Q_n(\eta, \alpha_0)$ obtained from (2.5), we have with probability tending to 1:*

1) $\max\left\{\|\Lambda^{1/2}(\hat{\eta} - \eta)\|_2, |\hat{\alpha}_0 - \alpha_0|\right\} \leq c_1 \lambda_n q_n^{1/2} n^{-1/18}$*, for some constant* $c_1 > 0$.

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

It is noteworthy that Theorem 1 is essentially different from and more challenging to show than the corresponding results in Xue and Yao (2021), since a more general framework of a GLM including possibly ultrahigh number of functional predictors is under concern. Additionally, the regression functions $\hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t)$ are not necessarily consistent in estimation under Theorem 1, setting itself apart from all existing results on GFLM. For a concrete example, based on Theorem 1, if we impose extra restrictions that

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

then it can be deduced that $\sum_{j=1}^{p_n} \|\hat{\beta}_j - \beta_j\|_{L_2}^2 \to \infty$. As we shall see in the forthcoming section, Theorem 1 provides theoretical guarantee on using some inconsistent estimators $\hat{\beta}_j(t)$ in post-regularization inference. Such desired feature for inference or testing is described as “estimation-consistency-relaxed”.

3. Inferring a general hypothesis on $LGFLM_{hete}$

Based on an estimator from (2.5), it is of primary interest to conduct post-regularization inference on a general spectrum of hypotheses regarding $LGFLM_{hete}$. To motivate such general hypothesis, we first let the index set $\mathbb{P}_n = \{1, \ldots, p_n\}$ to denote all predictors. Second, we write an arbitrary nonzero subset as $\mathcal{H}_n \subseteq \mathbb{P}_n$ containing $|\mathcal{H}_n| = h_n > 0$ elements, whose complement is $\mathcal{H}_n^c = \mathbb{P}_n \setminus \mathcal{H}_n$.

To this end, we can formulate the general hypothesis as

$$H_0 : \|\beta_j\|_{L^2} = 0 \text{ for every } j \in \mathcal{H}_n \text{ v.s. } H_a : \|\beta_j\|_{L^2} \neq 0 \text{ for some } j \in \mathcal{H}_n, \quad (3.6)$$

whose generality stems from the arbitrary selection of the size $h_n \le p_n$.

To execute the inference on (3.6), the main solution is to construct a confidence region of $\{\beta_j : j \in \mathcal{H}_n\}$ based on a pseudo score function and a penalized estimator $(\hat{\eta}, \hat{\alpha}_0)$ from (2.5). Specifically, a pseudo score function represents any extended version of the conventional score function, that will result in a valid inference (see, for example, the decorrelated score function in Ning and Liu, 2017). To motivate the pseudo score function of $LGFLM_{hete}$, we first present some useful notations. Here, we let $\eta_{\mathcal{H}_n}$ to denote the vector of attaching $\{\eta_j : j \in \mathcal{H}_n\}$ vertically in a column, whose estimator $\hat{\eta}_{\mathcal{H}_n}$ is defined analogously. Also, we write $\beta_{\mathcal{H}_n} = \{\beta_j : j \in \mathcal{H}_n\}$ to represent the collection of regression curves. We further denote $F_{\{b_k : k \le s_n\}}(\beta_{\mathcal{H}_n}) = \eta_{\mathcal{H}_n}$ as the function of mapping $\beta_{\mathcal{H}_n}$ onto $\eta_{\mathcal{H}_n}$. We write the matrix $\Theta_{\mathcal{H}_n}$ by attaching $\{\Theta_j : j \in \mathcal{H}_n\}$ in a line, and let $\Theta = \Theta_{\mathbb{P}_n}$ for brevity. For each $j \le p_n$, we write the moment estimator of the diagonal matrix $\Lambda_j = diag\{\omega_{j1}, \ldots, \omega_{js_n}\}$ by $\hat{\Lambda}_j = diag\{\hat{\omega}_{j1}, \ldots, \hat{\omega}_{js_n}\}$, with each $\hat{\omega}_{jk} = n^{-1} \sum_{i=1}^{n} \theta_{ijk}^2$. We then formulate the two diagonal matrices $\Lambda_{\mathcal{H}_n} = diag\{\Lambda_j : j \in \mathcal{H}_n\}$ and $\Lambda = diag\{\Lambda_j : j \in \mathbb{P}_n\}$, whose moment estimates are given by $\hat{\Lambda}_{\mathcal{H}_n} = diag\{\hat{\Lambda}_j : j \in \mathcal{H}_n\}$ and $\hat{\Lambda} = diag\{\hat{\Lambda}_j : j \in \mathbb{P}_n\}$. To this end, we present a series of matrices (expressed in row vectors) as

$$\Theta = (G_1, \ldots, G_n)', \Theta_{\mathcal{H}_n} = (E_1, \ldots, E_n)', \Theta_{\mathcal{H}_n^c} = (F_1, \ldots, F_n)', \tilde{\Theta} = \Theta\Lambda^{-1/2} = (\tilde{G}_1, \ldots, \tilde{G}_n)',$$

$$\tilde{\Theta}_{\mathcal{H}_n} = \Theta_{\mathcal{H}_n}\Lambda_{\mathcal{H}_n}^{-1/2} = (\tilde{E}_1, \ldots, \tilde{E}_n)', \tilde{\Theta}_{\mathcal{H}_n^c} = \Theta_{\mathcal{H}_n^c}\Lambda_{\mathcal{H}_n^c}^{-1/2} = (\tilde{F}_1, \ldots, \tilde{F}_n)', \breve{\Theta} = \Theta\hat{\Lambda}^{-1/2} = (\breve{G}_1, \ldots, \breve{G}_n)'$$

$$\breve{\Theta}_{\mathcal{H}_n} = \Theta_{\mathcal{H}_n}\hat{\Lambda}_{\mathcal{H}_n}^{-1/2} = (\breve{E}_1, \ldots, \breve{E}_n)', \breve{\Theta}_{\mathcal{H}_n^c} = \Theta_{\mathcal{H}_n^c}\hat{\Lambda}_{\mathcal{H}_n}^{-1/2} = (\breve{F}_1, \ldots, \breve{F}_n)'.$$

Several scaled-forms of the vector η are abbreviated by $\tilde{\eta} = \Lambda^{1/2}\eta$ and $\breve{\eta} = \hat{\Lambda}^{1/2}\eta$, and similarly for $\tilde{\eta}_{\mathcal{H}_n}$ and $\breve{\eta}_{\mathcal{H}_n}$. Since the loss function $L_n(\eta, \alpha_0) = n^{-1}\sum_{i=1}^{n}\{b(\delta_i^*) - Y_i\delta_i^*\}$ with each $\delta_i^* = \alpha_0 + \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\theta_{ijk}\eta_{jk} = \alpha_0 + \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\omega_{jk}^{-1/2}\theta_{ijk}\tilde{\eta}_{jk}$, it causes no confusion to write $L_n(\tilde{\eta}, \alpha_0) = L_n(\eta, \alpha_0)$ for convenience. In our model, it is natural to adopt $L_n(\tilde{\eta}, \alpha_0)$ as a pseudo negative log-likelihood function, where the genuine version is usually inaccessible due to the unknown parameters ϕ_i such as in the case of a large-scale FLM with heterogenous errors. At this point, we denote an unknown matrix w by

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

with each $w_j = (w_{j1}, \ldots, w_{j,(p_n-h_n)s_n})'$. Here, we use the value $\rho_n = \sup_{j \le h_ns_n}\rho_{nj}$ ($\rho_{nj} = card\{l : w_{jl} \neq 0\}$) to stand for the degree of sparsity of w. Note that condition (A5.4) demands ρ_n to be sparse in n. To this end, the pseudo score function for $\tilde{\eta}_{\mathcal{H}_n}$ can be formed by

$$S(\tilde{\eta}_{\mathcal{H}_n};\eta_{\mathcal{H}_n^c},\alpha_0) = n^{-1}\sum_{i=1}^{n}\big(w'F_i - \tilde{E}_i\big)\{Y_i - b'(\delta_i^*)\}, \qquad (3.7)$$

where $\delta_i^* = \alpha_0 + \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\theta_{ijk}\eta_{jk}$. Nevertheless, the utility of $S(\tilde{\eta}_{\mathcal{H}_n};\eta_{\mathcal{H}_n^c},\alpha_0)$ is limited since we don't know the values of $\{\Lambda_{\mathcal{H}_n},\eta_{\mathcal{H}_n^c},\alpha_0,w\}$, which need to be estimated. To bridge this gap, we approximate $\Lambda_{\mathcal{H}_n}$ by $\hat{\Lambda}_{\mathcal{H}_n}$, and use the estimator $(\hat{\eta}_{\mathcal{H}_n^c},\hat{\alpha}_0)$ in Theorem 1 to estimate $(\eta_{\mathcal{H}_n^c},\alpha_0)$. Due to the possibly high-dimensionality of w, it is reasonable to use a penalized method for estimation. Precisely, for each $j \le h_n s_n$, one solves the following penalization problem

$$\hat{w}_j = \underset{w_j}{\text{argmin}}\big[(2n)^{-1}\sum_{i=1}^{n}(\breve{E}_{ij} - F_i'w_j)^2 + \lambda_n^*\|\hat{\Lambda}_{\mathcal{H}_n^c}^{1/2}w_j\|_1\big], \qquad (3.8)$$

leading to the estimate $\hat{w} = (\hat{w}_1,\ldots,\hat{w}_{h_n s_n})$, where the penalty parameter $\lambda_n^* > 0$ is tuned via cross-validation. Plugging $\hat{w}$ and $\hat{\Lambda}_{\mathcal{H}_n}$ into (3.7), we reach an estimated version of $S(\tilde{\eta}_{\mathcal{H}_n};\eta_{\mathcal{H}_n^c},\alpha_0)$ as

$$\hat{S}(\tilde{\eta}_{\mathcal{H}_n};\eta_{\mathcal{H}_n^c},\alpha_0) = n^{-1}\sum_{i=1}^{n}(\hat{w}'F_i - \breve{E}_i)\{Y_i - b'(\alpha_0 + \breve{E}_i'\tilde{\eta}_{\mathcal{H}_n} + F_i'\eta_{\mathcal{H}_n^c})\}. \qquad (3.9)$$

Further substituting $(\eta_{\mathcal{H}_n^c},\alpha_0)$ by $(\hat{\eta}_{\mathcal{H}_n^c},\hat{\alpha}_0)$ gives rise to the estimated version of the pseudo score function for $\tilde{\eta}_{\mathcal{H}_n}$, denoted by

$$\hat{S}(\tilde{\eta}_{\mathcal{H}_n}; \hat{\eta}_{\mathcal{H}_n^c}, \hat{\alpha}_0) = n^{-1}\sum_{i=1}^{n}(\hat{w}'F_i - \breve{E}_i)\{Y_i - b'(\hat{\alpha}_0 + \breve{E}_i'\tilde{\eta}_{\mathcal{H}_n} + F_i'\hat{\eta}_{\mathcal{H}_n^c})\}, \tag{3.10}$$

which plays a crucial role in establishing the inferential procedure. For brevity of notation, we introduce a new function $\hat{T}(\beta_{\mathcal{H}_n})$ as

$$\hat{T}(\beta_{\mathcal{H}_n}) = n^{1/2}\hat{S}(\hat{\Lambda}_{\mathcal{H}_n}^{1/2}F_{\{b_k:k\leq s_n\}}(\beta_{\mathcal{H}_n}); \hat{\eta}_{\mathcal{H}_n^c}, \hat{\alpha}_0), \tag{3.11}$$

where $\beta_{\mathcal{H}_n} = \{\beta_j : j \in \mathcal{H}_n\}$ and $F_{\{b_k:k\leq s_n\}}(\beta_{\mathcal{H}_n}) = \eta_{\mathcal{H}_n}$. Before introducing the cutoff value in inference, we propose a related term $\hat{T}_e$, which takes the form

$$\hat{T}_e = n^{-1/2}\sum_{i=1}^{n} e_i(\hat{w}'F_i - \breve{E}_i)\{Y_i - b'(\hat{\alpha}_0 + G_i'\hat{\eta})\},$$

where $e = \{e_1, \ldots, e_n\}$ represents a collection of i.i.d. $N(0,1)$, that are independent of the data. Then, the cutoff value is defined as the $(1-\alpha)$th quantile of $\|\hat{T}_e\|_\infty$, denoted by

$$c_B(\alpha) = \inf\{t \in \mathbb{R} : P_e(\|\hat{T}_e\|_\infty \leq t) \geq 1-\alpha\}, \qquad \alpha \in (0,1) \tag{3.12}$$

where $P_e(\cdot)$ represents the conditional probability that only treats e as random. The computation of $c_B(\alpha)$ can be fastly achieved through multiplier bootstrap on e. The following Theorem 2 lays down a theoretical basis for conducting general inference under fairly flexible conditions (A1)-(A5) from Appendix A and (B1)-(B5) from Appendix B.

Theorem 2. *Under conditions (A1)–(A5) and (B1)–(B5), the Kolmogorov distance between the distributions of* $\|\hat{T}(\beta_{\mathcal{H}_n})\|_\infty$ *and* $\|\hat{T}_e\|_\infty$ *satisfies*

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

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

The quantities $\hat{T}(\beta_{\mathcal{H}_n})$ and $c_B(\alpha)$ can be referred to (3.11) and (3.12). Based on Theorem 2, a $100(1-\alpha)\%$ confidence region of $\beta_{\mathcal{H}_n}$ takes the form

$$CR_{1-\alpha} = \{\beta_{\mathcal{H}_n} : \|\hat{T}(\beta_{\mathcal{H}_n})\|_\infty \leq c_B(\alpha)\}. \tag{3.13}$$

Thus, the corresponding test is to reject the H_0 specified by (3.6) at nominal level $\alpha \in (0,1)$ provided that

$$\|\hat{T}(0)\|_\infty > c_B(\alpha), \tag{3.14}$$

with cutoff value $c_B(\alpha)$ and test statistic $\|\hat{T}(0)\|_\infty$. Notably, the inferential procedure induced by Theorem 2 is uniformly consistent for every $0 < \alpha < 1$. Also notice that Theorem 2 is essentially distinguished from its counterpart in our previous work Xue and Yao (2021), due to not only the more challenging framework of a large-scale generalized linear model, but also the fact that a confidence region is established instead of just an asymptotic test. As a result, such confidence region can be used to analyze power in the following Theorems 3–4. Since (A3.1) in Appendix A is the only assumption on the eigenvalues ω_{jk}, it indicates that the inferential procedure enjoys the distinguishing advantage of imposing no constraint on the decay of those eigenvalues, in contrast to most current articles on generalized functional linear models that require ω_{jk} to satisfy either $\lambda_{\min}(\Lambda) \gtrsim s_n^{-a}$ or $\omega_{jk} \gtrsim k^{-a}$, for a constant $a > 1$. Such desired advantage is described as "eigenvalue-decay-free". It is also seen from (A3.1) that the proposed inferential procedure enjoys another distinguishing advantage of permitting $\sup_{j \le p_n} \sum_{k=1}^{\infty} \omega_{jk} = \infty$, in contrast to all existing literature that demand a square-integrable condition: $\sup_{j \le p_n} \sum_{k=1}^{\infty} \omega_{jk} < \infty$. This advantage is regarded as "square-integrable-free".

To evaluate power at the given genuine $\beta_{\mathcal{H}_n}$, first note that the true power function takes the form $Power(\beta_{\mathcal{H}_n}) = P\{\|\hat{T}(0)\|_\infty > c_B(\alpha)|\beta_{\mathcal{H}_n}\}$. Nevertheless, the true power cannot be assessed due to the unknown distribution of $\|\hat{T}(0)\|_\infty$. To bridge this gap, the key idea is to exploit a proper estimate of $Power(\beta_{\mathcal{H}_n})$. In a similar spirit to Theorem 2, another procedure of multiplier bootstrap is adopted to imitate the distribution of $\|\hat{T}(0)\|_\infty$, thus approximating the genuine power. More precisely, the estimated version of $Power(\beta_{\mathcal{H}_n})$ is expressed by

$$\begin{aligned} Power^*(\beta_{\mathcal{H}_n}) =& P_{e^*}[\|\hat{T}_{e^*} + n^{-1/2}\sum_{i=1}^{n}(\hat{w}'F_i - \breve{E}_i)\{b'(\hat{\alpha}_0 + E_i' \cdot F_{\{b_k:k\le s_n\}}(\beta_{\mathcal{H}_n}) + F_i'\hat{\eta}_{\mathcal{H}_n^c}) \\ & - b'(\hat{\alpha}_0 + F_i'\hat{\eta}_{\mathcal{H}_n^c})\}\|_\infty > c_B(\alpha)], \end{aligned} \tag{3.15}$$

with $e^* = \{e_1^*, \ldots, e_n^*\}$ as an independent copy of e. The computation of (3.15) can be achieved via a multiplier bootstrap procedure on e^*. The asymptotic equivalence between $Power(\beta_{\mathcal{H}_n})$ and $Power^*(\beta_{\mathcal{H}_n})$ is then established in the Theorem below.

Theorem 3. *Under conditions (A1)–(A5) and (B1)–(B5), given the true version $\beta_{\mathcal{H}_n}$, we have:* $\lim_{n\to\infty}\left|Power(\beta_{\mathcal{H}_n}) - Power^*(\beta_{\mathcal{H}_n})\right| = 0$.

Regarding the analysis of power, the consistency of $Power^*(\beta_{\mathcal{H}_n})$ is ensured for a quite general set of alternatives in the following Theorem.

Theorem 4. *Assume the conditions (A1)–(A5) and (B1)–(B5) and that the true version $\beta_{\mathcal{H}_n}$ belongs to the alternative set*

$$\begin{aligned} &\mathcal{F}_n = \{\beta_{\mathcal{H}_n} : P\big(\|n^{-1}\sum_{i=1}^{n}(\tilde{E}_i - w'F_i)\{b'(\hat{\alpha}_0 + E_i'\eta_{\mathcal{H}_n} + F_i'\hat{\eta}_{\mathcal{H}_n^c}) - b'(\hat{\alpha}_0 + F_i'\hat{\eta}_{\mathcal{H}_n^c})\}\|_\infty \\ &\ge Kn^{-1/2}(\rho_n^{1/2} + \log n)\{\log(np_ns_n)\}^{1/2}\exp\{Kq_n\log^{1/2}(nq_ns_n)\}\big) \to 1\}, \end{aligned}$$

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

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

4. Simulation Studies

Since the finite sample performance of the linear model is thoroughly evaluated in Xue and Yao (2021) and Xue and Yao (2024), we consider another two crucial cases of a $LGFLM_{hete}$, which are the logistic and poisson models as follows.

4.1 Logistic model

For logistic model, the response data $\{Y_i, i = 1, \ldots, n\}$ are independently simulated from the distribution as follows

$$Y_i|X_i \sim Bernoulli(\frac{\exp(\delta_i)}{1+\exp(\delta_i)}),$$

with the linear predictor $\delta_i = \alpha_0+\sum_{j=1}^{p_n} \int_T X_{ij}(t)\beta_j(t)dt = \alpha_0+\sum_{j=1}^{q_n} \int_T X_{ij}(t)\beta_j(t)dt = \alpha_0+\sum_{j=1}^{q_n} \sum_k \eta_{jk}\theta_{ijk}$ under the ultra-high-dimensional setting $(n, p_n) = (100, 200)$. We set the intercept $\alpha_0 = 0$ without loss of generality. The basis representation of the nonzero regression functions are defined by $\beta_j(t) = \sum_{k=1}^{50} \eta_{jk}\phi_k(t)$, with coefficients $\eta_{jk} = h_j[(1.2-0.2k)1_{\{k\leq 5\}} + 0.2(k-4)^{-3}1_{\{6\leq k\leq 50\}}]$ for $j \leq q_n$, where the parameters $\{h_j : j \leq q_n\}$ control the intensity level of each predictor. The fourier basis $\{\phi_k(\cdot) : k \geq 1\}$ defined on $T = [0, 1]$ is orthonormal and can be expressed as $\phi_1 = 1$, $\phi_{2l} = 2^{1/2}\cos\{l\pi(2t-1)\}$, $l = 1, \ldots, 25$ and $\phi_{2l-1} = 2^{1/2}\sin\{(l-1)\pi(2t-1)\}$, $l = 2, \ldots, 25$. In order to obtain the predictors $\{X_{ij}(\cdot) : j \leq p_n\}$, we start by introducing a collection of independent random processes $\{Z_{ij}(\cdot) : j \leq p_n\}$ as

$$Z_{ij}(t) = \sum_{k=1}^{50} \tilde{\xi}_{ijk}\phi_k(t), \quad t \in [0,1],$$

with the scores $\{\tilde{\xi}_{ijk}\}$ independently simulated from $N(0, \tau_k^2)$. The sequence $\tau_1, \ldots, \tau_{50}$ are independently generated from $uniform(1/4, 1/2)$, and are set fixed through all simulations. To this end, the functional predictors X_{ij} can be formulated via autoregressive correlation:

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

where $\theta_{ijk} = \sum_{j'=1}^{p_n} \rho^{|j-j'|}\tilde{\xi}_{ij'k}$, and the value of $\rho \in (0,1)$ governs the overall correlation among these predictors. Without loss of generality, we consider the setting of $\rho = 0.3$ in simulation. In addition, when compared to existing studies on functional data, it follows from the definition of τ_k that our setting is more challenging since it not only imposes no decay restriction on eigenvalues of the processes X_{ij}, but also violates the square-integrable condition due to the fact that $\int_T E(X_{ij}^2)dt \asymp \sum_{k=1}^{d} \tau_k \to \infty$ for large d. The observed measurements of $\{X_{ij}(\cdot), j = 1, \ldots, p_n\}$ are taken discretely at $m = 100$ time points $\{t_{ijl} = \frac{l-1}{m-1} : l = 1, \ldots, m\} \in T$. An orthonormal cubic spline basis is then adopted to obtain the fitted curves, with the tuning parameters s_n, λ_n, λ_n^* chosen from 5-fold cross-validation based on the algorithm in Appendix B using SCAD penalty. To conduct test, we adopt nominal level $\alpha = 5\%$ and bootstrap resample size $N = 10000$. In the upper part of Table 1, a parsimonious model with $q_n = 3$ is considered and we report the rejection proportions for testing several H_0 under various settings of $\{h_j : j \leq q_n\}$, on the basis of 1000 Monte Carlo repetitions. Furthermore, a denser case of $q_n = 6$ is summarized in the lower part of Table 1.

The upper part of Table 1 shows that the empirical size corresponding to the null $\mathcal{H}_n = \{4, \ldots, 6\}$ approximates the prespecified level $\alpha = 5\%$ very well, which is consistent with the theory of the proposed test. Besides, as we include more nonzero regressors into $\mathcal{H}_n$, the rejection proportion (empirical power) ascends fast, showing the pattern of a power curve. As expected, the empirical power also rises significantly when the signal strength level is raised from $.3 \times \mathbf{1}_{1\times 3}$ to $.5 \times \mathbf{1}_{1\times 3}$. Analogous patterns are shown in the lower part of Table 1, which illustrate the validity of the test under relative denser signals.

Table 1: Results of the logistic model for different settings of the regression curves $\{\beta_j : j \leq q_n\}$ specified by $\{h_k : k \leq q_n\}$, under various values of q_n and hypotheses $\mathcal{H}_n$ based on 1000 Monte Carlos, in the context of large-scale setting: $(n, p_n) = (100, 200)$. Shown are the empirical rejection proportions.

Value of q_n	$(h_1, \ldots, h_{q_n})$	$H_0 : \mathcal{H}_n$	Rejection proportion
$q_n = 3$	$.3 \times \mathbf{1}_{1\times3}$	$\{4, \ldots, 6\}$	**.052**
	$.3 \times \mathbf{1}_{1\times3}$	$\{3, \ldots, 6\}$	.082
	$.3 \times \mathbf{1}_{1\times3}$	$\{2, \ldots, 6\}$	.116
	$.3 \times \mathbf{1}_{1\times3}$	$\{1, \ldots, 6\}$	.258
	$.5 \times \mathbf{1}_{1\times3}$	$\{4, \ldots, 6\}$	**.053**
	$.5 \times \mathbf{1}_{1\times3}$	$\{3, \ldots, 6\}$	.103
	$.5 \times \mathbf{1}_{1\times3}$	$\{2, \ldots, 6\}$	.163
	$.5 \times \mathbf{1}_{1\times3}$	$\{1, \ldots, 6\}$	.470
$q_n = 6$	$.3 \times \mathbf{1}_{1\times6}$	$\{7, \ldots, 9\}$	**.054**
	$.3 \times \mathbf{1}_{1\times6}$	$\{5, \ldots, 9\}$	.101
	$.3 \times \mathbf{1}_{1\times6}$	$\{3, \ldots, 9\}$	.193
	$.3 \times \mathbf{1}_{1\times6}$	$\{1, \ldots, 9\}$	.425
	$.5 \times \mathbf{1}_{1\times6}$	$\{7, \ldots, 9\}$	**.051**
	$.5 \times \mathbf{1}_{1\times6}$	$\{5, \ldots, 9\}$	.180
	$.5 \times \mathbf{1}_{1\times6}$	$\{3, \ldots, 9\}$	.452
	$.5 \times \mathbf{1}_{1\times6}$	$\{1, \ldots, 9\}$	.784

4.2 Poisson model

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

$$Y_i|X_i \sim Poisson(\exp(\delta_i)),$$

where the linear predictors $\delta_i = \alpha_0 + \sum_{j=1}^{p_n} \int_T X_{ij}(t)\beta_j(t)dt = \alpha_0 + \sum_{j=1}^{q_n} \int_T X_{ij}(t)\beta_j(t)dt = \alpha_0 + \sum_{j=1}^{q_n} \sum_k \eta_{jk}\theta_{ijk}$ are identically generated as those in the logistic model. We adopt the same settings as those in Table 1, and report the corresponding results in Table 2 below. Since the pattern in Table 2 is similar to that of Table 1, it further justifies the validity of the proposed test for poisson model. The computation costs around 1 minute for each test in one Monte Carlo for both models.

Table 2: Results of the poisson model for different settings of the regression curves $\{\beta_j : j \leq q_n\}$ specified by $\{h_k : k \leq q_n\}$, under various values of q_n and hypotheses $\mathcal{H}_n$ based on 1000 Monte Carlos, in the context of large-scale setting: $(n, p_n) = (100, 200)$. Shown are the empirical rejection proportions.

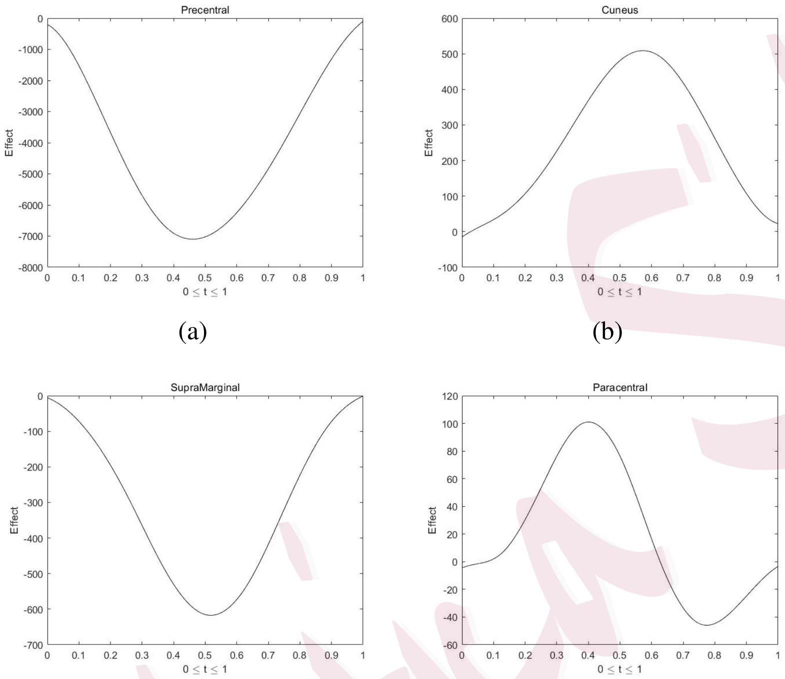
Value of q_n	$(h_1, \ldots, h_{q_n})$	$H_0 : \mathcal{H}_n$	Rejection proportion
$q_n = 3$	$.3 \times \mathbf{1}_{1\times 3}$	$\{4, \ldots, 6\}$	**.045**
	$.3 \times \mathbf{1}_{1\times 3}$	$\{3, \ldots, 6\}$	.084
	$.3 \times \mathbf{1}_{1\times 3}$	$\{2, \ldots, 6\}$	.204
	$.3 \times \mathbf{1}_{1\times 3}$	$\{1, \ldots, 6\}$	.488
	$.5 \times \mathbf{1}_{1\times 3}$	$\{4, \ldots, 6\}$	**.050**
	$.5 \times \mathbf{1}_{1\times 3}$	$\{3, \ldots, 6\}$	.270
	$.5 \times \mathbf{1}_{1\times 3}$	$\{2, \ldots, 6\}$	.635
	$.5 \times \mathbf{1}_{1\times 3}$	$\{1, \ldots, 6\}$	.905
$q_n = 6$	$.3 \times \mathbf{1}_{1\times 6}$	$\{7, \ldots, 9\}$	**.048**
	$.3 \times \mathbf{1}_{1\times 6}$	$\{5, \ldots, 9\}$	.188
	$.3 \times \mathbf{1}_{1\times 6}$	$\{3, \ldots, 9\}$	.476
	$.3 \times \mathbf{1}_{1\times 6}$	$\{1, \ldots, 9\}$	.572
	$.5 \times \mathbf{1}_{1\times 6}$	$\{7, \ldots, 9\}$	**.046**
	$.5 \times \mathbf{1}_{1\times 6}$	$\{5, \ldots, 9\}$	.305
	$.5 \times \mathbf{1}_{1\times 6}$	$\{3, \ldots, 9\}$	.660
	$.5 \times \mathbf{1}_{1\times 6}$	$\{1, \ldots, 9\}$	.760

5. Real Data

In this section, we adopt one real data set regarding attention deficit hyperactivity disorder (ADHD) to show the desired performance of the proposed method, where ADHD is known as a common neurodevelopmental disorder of childhood. More precisely, we use the preprocessed resting state data from the ADHD-200

Sample Initiative Project under Anatomical Automatic Labeling atalas (Tzourio-Mazoyer et al., 2002), where $n = 137$ individuals are included in the study after quality control. In dataset, each individual is linked to $p_n = 116$ brain regions of interest, where the mean grayscale is recorded over 172 evenly spaced time points for each region. The binary response of interest $Y_i \in \{0, 1\}$ is the diagnosis status, where $Y_i = 1$ means the sick state. Our target is to find the significant functional predictors among the 116 regions for predicting the ADHD status.

Given the discrete response, we assume the model as logistic, representing an important subcase of (1.2). At the significance level $\alpha = .05$, the proposed testing method is first conducted to access the simple hypotheses $H_0 : \beta_j = 0$ respectively. Based on these simple tests, it is found that 4 regression curves $(\beta_j : j = 2, 45, 64, 70)$ corresponding to brain regions of Precentral, Cuneus, SupraMarginal, and Paracentral are identified as significant, which makes sense due to the literature such as (Long et al., 2022; Hart et al., 2013; Hale et al., 2014; Liu et al., 2017; Griffiths et al., 2021). To confirm further, letting $\mathcal{H}_n = \{2, 45, 64, 70\}$, we apply the testing procedure to the two composite hypotheses $H_0^1 : \beta_j = 0$ for all $j \in \mathcal{H}_n$, and $H_0^2 : \beta_j = 0$ for all $j \in \mathcal{H}_n^c$, which rejects H_0^1 and accepts H_0^2. This further verifies the importance of regions in $\mathcal{H}_n$ for predicting disease status. For comprehension, the estimated regression curves for regions in $\mathcal{H}_n$ are shown in Figure 1, in spite of the scenario of estimation-consistency-relaxed.



(a) (b)

(c) (d)

Figure 1: The estimated regression functions are depicted for significant brain regions of Precentral, Cuneus, SupraMarginal, and Paracentral respectively.

Appendix

A. Assumptions on the $LGFLM_{hete}$

To ensure the theoretical results, we impose some mild conditions (A1)–(A5) on the model as follows. First of all, assumption (A1) is on the property of the function $b(t)$, which specify the canonical link function $g(\cdot) = (b')^{-1}(\cdot)$.

(A1). Given any $K > 0$, for every $t, t_1, t_2 \in [-K, K]$, we have

$$4^{-1}\exp(-K) \leq b''(t) \leq \exp(K), \quad |b''(t_1) - b''(t_2)| \leq \exp(3K) \cdot |t_1 - t_2|.$$

Condition (A1) is satisfied by a large class of generalized linear models including the linear regression, the logistic regression, and the poisson regression. Condition (A2) imposes mild distributional assumptions on several random quantities, which consists of parts (A2.1)–(A2.4).

(A2). (A2.1). The random terms $\omega_{jk}^{-1/2}\theta_{ijk}$, $w_t{}'F_i$ are centered sub-Gaussian with variance proxy σ^2 for some constant $\sigma > 0$, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, p_n$, $k = 1, \ldots, \infty$, $t = 1, \ldots, h_n s_n$.

(A2.2). There is a sufficiently large universal constant $C > 0$ such that:

$$P\{\max_{l \leq h_n s_n} \|n^{-1}\sum_{i=1}^{n}\left(\begin{smallmatrix}\tilde{F}_i \\ 1\end{smallmatrix}\right)(w_l{}'F_i - \tilde{E}_{il})b''(\delta_i^*)\|_\infty \leq C \max_{i \leq n} b''(\delta_i^*) \cdot$$

$$\max_{l \leq h_n s_n} \|n^{-1}\sum_{i=1}^{n}\left(\begin{smallmatrix}\tilde{F}_i \\ 1\end{smallmatrix}\right)(w_l{}'F_i - \tilde{E}_{il})\|_\infty\} \to 1.$$

(A2.3). The conditional distribution of the error ϵ_i can be either

$$\epsilon_i | X_i \sim \text{sub-Gaussian}(\sigma^{*2}\{1 + \text{var}(\epsilon_i | X_i)\})$$

$$\text{or} \quad \epsilon_i | X_i \sim \text{sub-Exponential}(\sigma^{*2}\{1 + \text{var}(\epsilon_i | X_i)\}),$$

where X_i stands for the data $\{X_{ij} : j \leq p_n\}$, and $\sigma^{*2}\{1 + \text{var}(\epsilon_i | X_i)\}$ is the variance proxy, with $\sigma^* > 0$ as a universal constant.

(A2.4). The errors $\epsilon_1, \ldots, \epsilon_n$ are mutually independent (may not have identical distribution), and meet the moment condition:

$$\min_{l \le h_n s_n} n^{-1} \sum_{i=1}^{n} E\{(w_l' F_i - \tilde{E}_{il})^2 \epsilon_i^2\} \ge c_1, \text{ for some universal constant } c_1 > 0.$$

Notice that (A2.1) demands the sub-Gaussianity of the predictors, which is commonly assumed in high-dimensional data analysis. (A2.2) further imposes a mild restriction on the distribution of the data $\{X_{ij} : i \le n, j \le p_n\}$, which holds apparently for the linear model by taking $C = 1$. (A2.3) is on the distribution type of error terms, which includes many commonly used GLMs. For instance, (A2.3) holds true for poisson regression with $\epsilon_i | X_i \sim$ sub-Exponential$(\sigma^{*2}\{1 +$ var$(\epsilon_i|X_i)\})$, and is also valid for logistic regression with $\epsilon_i | X_i \sim$ sub-Gaussian(σ^{*2}). (A2.4) correspond to a moment assumption on the error terms, which is also quite general permitting a portion of vanishing errors. In particular, (A2.3) and (A2.4) together indicate that the error terms are allowed to be non $i.i.d.$ or heterogeneous under the linear model of (1.2). Condition (A3) is used to regulate the smoothness and the correlation structure of the $LGFLM_{hete}$.

(A3). (A3.1). $\sup_{j \le p_n} \sup_{k \ge 1} \omega_{jk} < \infty$.

(A3.2). $\sup_{j \le q_n} \sum_{k=1}^{\infty} \eta_{jk}^2 k^{2\delta} < \infty$, for some constant $\delta > 0$.

(A3.3). $c_1^{-1} \le \lambda_{\min}(E(\tilde{G}_i \tilde{G}_i')) \le \lambda_{\max}(E(\tilde{G}_i \tilde{G}_i')) \le c_1$, for a constant $c_1 > 0$.

(A3.1) stands for the sole smoothness constraint on the functional predictors X_{ij}, which is extremely loose by permitting non-square-integrable processes in the sense of $\sup_{j \le p_n} \int_T E(X_{ij}^2)dt = \sup_{j \le p_n} \sum_{k=1}^{\infty} \omega_{jk} = \infty$ (i.e., square-integrable-free), and also requires no decay restrictions on ω_{jk} (i.e., eigenvalue-decay-free). In comparison, all existing literature on GFLM demand the square-integrable condition: $\sup_{j \le p_n} \sum_{k=1}^{\infty} \omega_{jk} < \infty$, as well as some eigenvalue-decay conditions such as $\omega_{jk} - \omega_{j,k+1} \gtrsim k^{-a-1}$ for some $a > 1$. (A3.2) regulates the smoothness level of the non-vanishing regression functions $\{\beta_j : j \le q_n\}$ via a positive constant δ. We require bounded eigenvalues of the correlation matrix in (A3.3), which is commonly assumed in large-scale data analysis. The relative magnitudes of a series of important parameters are provided in (A4).

(A4). (A4.1). For any constant $c > 0$, we have:

$$\frac{\{\log(np_n s_n)\}^{12} \exp\{c q_n \log^{1/2}(n q_n s_n)\}}{n} \to 0.$$

(A4.2). For any constant $c > 0$, we have:

$$\frac{s_n^2 \{\log(np_n s_n)\}^{6} \exp\{c q_n \log^{1/2}(n q_n s_n)\}}{n} \to 0,$$

$$\frac{s_n B_n^2 \{\log(np_n s_n)\}^{2} \exp\{c q_n \log^{1/2}(n q_n s_n)\}}{n} \to 0.$$

(A4.3). For any constant $c > 0$, we have:

$$\frac{s_n^{2\delta - 1}}{n\{\log(np_n s_n)\} \exp\{c q_n \log^{1/2}(n q_n s_n)\}} \to \infty.$$

(A4.1) incorporates the large-scale model framework by allowing for exponentially increasing p_n in n, and it also suggests a parsimonious model structure due to $q_n \ll n$. Regarding the truncation number s_n, (A4.2) demands s_n to grow relatively slowly in n to mitigate the negative effects caused by the infinite-dimensionality of these processes X_{ij}, whereas (A4.3) requires s_n to grow relatively fast in n to retain adequate information for valid inference. (A4.2) together with (A4.3) entails $\delta > 3/2$, where the smoothing parameter δ can be referred to (A3.2). Based on (A3) and (A4), it can be deduced that $\|\eta\|_1 \lesssim q_n$, which suggests us to choose $B_n \asymp q_n$, where B_n is defined in (2.5). (A5) specifies the magnitudes of the penalization parameters λ_n and λ_n^*, as well as the sparsity level $\rho_n = \sup_{j\le h_n s_n} \|w_j\|_0$ of $w = \{E(F_i F_i')\}^{-1}E(F_i\tilde{E}_i') = (w_1, \ldots, w_{h_n s_n})$.

(A5). (A5.1). For any constant $c > 0$, we have:

$$n^{-1/9}s_n\{\log(np_n s_n)\}^2\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{-2}),$$

$$n^{-1/9}\rho_n\{\log(np_n s_n)\}^2\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{-2}),$$

$$n^{-1/9}\{\log(np_n s_n)\}^3\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{-2}),$$

$$n^{7/18}\{\log(np_n s_n)\}\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{-2}),$$

$$n^{-1/9}B_n^2 = o(\lambda_n^{-2}).$$

(A5.2). For any constant $c > 0$, we have:

$$n^{-8/9}s_n\{\log(np_n s_n)\}^2\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{2}),$$

$$n^{1/9}s_n^{-2\delta+2}\exp\{cq_n\log^{1/2}(nq_n s_n)\} = o(\lambda_n^{2}).$$

(A5.3). $K_1\{\log(np_n s_n)/n\}^{1/2} \le \lambda_n^* \le K_2\{\log(np_n s_n)/n\}^{1/2}$, for some sufficiently large universal constants $K_2 > K_1 > 0$.

(A5.4). For any constant $c > 0$, we have:

$$\frac{\rho_n^2\{\log(np_ns_n)\}^4\exp\{cq_n\log^{1/2}(nq_ns_n)\}}{n} \to 0.$$

The order of the parameter λ_n is given in (A5.1) and (A5.2), while the magnitude of λ_n^* is quantified by (A5.3). The order of ρ_n is provided in (A5.4), suggesting a sparse matrix w. As an illustrative example, assumptions (A1)–(A5) hold true for the large-scale scenario $\big\{\log p_n \asymp n^{1/20};\ \ q_n \asymp B_n \asymp \log^{1/4}(n);\ \ s_n \asymp \rho_n \asymp n^{1/9};\ \ \lambda_n \asymp n^{-1/3};\ \ \lambda_n^* \asymp n^{-19/40};\ \ \delta = 6\big\}$.

B. Penalties and algorithm

For the sake of simplicity, we write each $\hat{f}_j = \Theta_j\hat{\eta}_j$ for $j = 1, \ldots, p_n$, and denote the vector $\hat{\delta}^* = (\hat{\delta}_1^*, \ldots, \hat{\delta}_n^*)' = \hat{\alpha}_0 1_n + \sum_{j=1}^{p_n}\hat{f}_j$. We use $Y = (Y_1, \ldots, Y_n)'$ to represent the response vector. The coordinate descent algorithm to solve (2.5) is adapted from Fan et al. (2015) and Ravikumar et al. (2008). Such algorithm can be implemented to a wide spectrum of penalties ρ_λ provided that the mild conditions (B1)–(B5) on ρ_λ listed below are satisfied. Similar conditions can be also found in Loh and Wainwright (2015). To be specific, we assume:

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

(B2) $\rho_\lambda(t)$ is nondecreasing on the interval $[0, \infty)$.

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

(B4) $\rho_\lambda(t)$ has derivatives at every $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t\to 0^+}\rho_\lambda'(t) = \lambda L$ for a universal constant $L > 0$.

(B5) There is a constant $\mu > 0$ so that the function $\rho_{\lambda,\mu}(t) = \rho_\lambda(t) + 2^{-1}\mu t^2$ is convex.

Notably, a wide range of commonly seen regularizers (e.g., SCAD, LASSO, MCP) fulfill (B1)–(B5). Lastly, the algorithm to fit (2.5) is presented below

```
(i)      Initialize $\hat{\alpha}_0 = 0$ and $\hat{f}_j = 0$, for every $j = 1, \ldots, p_n$.
(ii)     Compute $\hat{\delta}^* = (\hat{\delta}_1^*, \ldots, \hat{\delta}_n^*)' = \hat{\alpha}_0 1_n + \sum_{j=1}^{p_n} \hat{f}_j$.
(iii)    Compute $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)'$ with each $\hat{\pi}_i = b'(\hat{\delta}_i^*)$.
(iv)     Compute $v = \|\hat{\pi}\|_\infty$.
(v)      Compute $\tilde{Y} = \hat{\delta}^* + \frac{Y - \hat{\pi}}{v}$.
(vi)     For $j = 1, \ldots, p_n$
             set $z_j = n^{-1}(\tilde{Y} - \hat{\delta}^*) + n^{-1}\hat{f}_j$,
             set $\tilde{f}_j = \max\{n\|z_j\|_2 - n^{4/9}v^{-1}\rho'_{\lambda_n}(n^{-5/9}\|\hat{f}_j\|_2), 0\} z_j/(\|z_j\|_2 + 1_{\{\|z_j\|_2 = 0\}})$,
             set $\hat{\delta}^* = \hat{\delta}^* + (\tilde{f}_j - \hat{f}_j)$,
             set $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)'$ with each $\hat{\pi}_i = b'(\hat{\delta}_i^*)$,
             set $v = \|\hat{\pi}\|_\infty$,
             set $\tilde{Y} = \hat{\delta}^* + \frac{Y - \hat{\pi}}{v}$,
             set $\hat{f}_j = \tilde{f}_j$,
             set $\tilde{\alpha}_0 = n^{-1}1_n'(\tilde{Y} - \hat{\delta}^*) + \hat{\alpha}_0$,
             set $\hat{\delta}^* = \hat{\delta}^* + (\tilde{\alpha}_0 - \hat{\alpha}_0)1_n$,
             set $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)'$ with each $\hat{\pi}_i = b'(\hat{\delta}_i^*)$,
             set $v = \|\hat{\pi}\|_\infty$,
             set $\tilde{Y} = \hat{\delta}^* + \frac{Y - \hat{\pi}}{v}$,
             set $\hat{\alpha}_0 = \tilde{\alpha}_0$,
         end.
(vii)    Repeat (ii)–(vi) until convergence to obtain the final estimators $\hat{\alpha}_0$ and $\{\hat{f}_j : j = 1, \ldots, p_n\}$.
(viii)   Using the final estimators from (vii), obtain the final estimators $\hat{\alpha}_0$, and $\hat{\eta}_j = ({\Theta_j}'\Theta_j)^{-1}{\Theta_j}'\hat{f}_j$ for $j = 1, \ldots, p_n$.
```

Supplementary Materials

The auxiliary lemmas with their proofs, and the proofs of the main theorems are delegated to an online Supplementary Material for space economy.

Acknowledgements

Kaijie Xue is the first author. Riquan Zhang is the corresponding author: zhangriquan@163.com. This research was supported by the National Natural Science Foundation of China (12371272, 12371268) and the Basic Research Project of Shanghai Science and Technology Commission (22JC1400800).

References

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

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.

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 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.

Griffiths, K. R., T. A. Braund, M. R. Kohn, S. Clarke, L. M. Williams, and M. S. Korgaonkar (2021). Structural brain network topology underpinning adhd and response to methylphenidate treatment. *Translational psychiatry 11*(1), 150.

Hale, T. S., A. M. Kane, K. L. Tung, O. Kaminsky, J. J. McGough, G. Hanada, and S. K. Loo (2014). Abnormal parietal brain function in adhd: replication and extension of previous eeg beta asymmetry findings. *Frontiers in Psychiatry 5*, 87.

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

Hart, H., J. Radua, T. Nakao, D. Mataix-Cols, and K. Rubia (2013). Meta-analysis of functional magnetic resonance imaging studies of inhibition and attention in attention-deficit/hyperactivity disorder: exploring task-specific, stimulant medication, and age effects. *JAMA psychiatry 70*(2), 185–198.

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. *Biometrika 103*, 147–159.

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

Li, T. and Z. Zhu (2020). Inference for generalized partial functional linear regression. *Statistica Sinica 30*(3), 1379–1397.

Liu, T., Y. Chen, C. Li, Y. Li, and J. Wang (2017). Altered brain structural networks in attention deficit/hyperactivity disorder children revealed by cortical thickness. *Oncotarget 8*(27), 44785.

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.

Long, Y., N. Pan, S. Ji, K. Qin, Y. Chen, X. Zhang, M. He, X. Suo, Y. Yu, S. Wang, et al. (2022). Distinct brain structural abnormalities in attention-deficit/hyperactivity disorder and substance use disorders: A comparative meta-analysis. *Translational Psychiatry 12*(1), 368.

Malfait, N. and J. O. Ramsay (2003). The historical functional linear model. *Canadian Journal of Statistics 31*(2), 115–128.

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

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* (2nd ed.). 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.

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.

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

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

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, and B. M. . M. Joliot (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.

Xiao, W., Y. Wang, and H. Liu (2021). Generalized partially functional linear model. *Scientific reports 11*(1), 23428.

Xue, K. (2023). Distribution/correlation-free test for two-sample means in high-dimensional functional data with eigenvalue decay relaxed. *Science China Mathematics 66*, 2337–2346.

Xue, K., J. Yang, and F. Yao (2023). Optimal linear discriminant analysis for high-dimensional functional data. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2022.2164288.

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

Xue, K. and F. Yao (2024). Inference on large-scale partially functional linear model with heterogeneous errors. *Statistica Sinica 34*, 679–697.

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