*The Annals of Statistics*
2020, Vol. 48, No. 5, 2622–2645
https://doi.org/10.1214/19-AOS1900


# TEST OF SIGNIFICANCE FOR HIGH-DIMENSIONAL LONGITUDINAL DATA

BY ETHAN X. FANG[1,*], YANG NING[2] AND RUNZE LI[1,**]

[1]*Department of Statistics, Pennsylvania State University,* [*]*xxf13@psu.edu;* [**]*rzli@psu.edu*
[2]*Department of Statistics and Data Science, Cornell University, yn265@cornell.edu*

This paper concerns statistical inference for longitudinal data with ultrahigh dimensional covariates. We first study the problem of constructing confidence intervals and hypothesis tests for a low-dimensional parameter of interest. The major challenge is how to construct a powerful test statistic in the presence of high-dimensional nuisance parameters and sophisticated within-subject correlation of longitudinal data. To deal with the challenge, we propose a new quadratic decorrelated inference function approach which simultaneously removes the impact of nuisance parameters and incorporates the correlation to enhance the efficiency of the estimation procedure. When the parameter of interest is of fixed dimension, we prove that the proposed estimator is asymptotically normal and attains the semiparametric information bound, based on which we can construct an optimal Wald test statistic. We further extend this result and establish the limiting distribution of the estimator under the setting with the dimension of the parameter of interest growing with the sample size at a polynomial rate. Finally, we study how to control the false discovery rate (FDR) when a vector of high-dimensional regression parameters is of interest. We prove that applying the Storey (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** (2002) 479–498) procedure to the proposed test statistics for each regression parameter controls FDR asymptotically in longitudinal data. We conduct simulation studies to assess the finite sample performance of the proposed procedures. Our simulation results imply that the newly proposed procedure can control both Type I error for testing a low dimensional parameter of interest and the FDR in the multiple testing problem. We also apply the proposed procedure to a real data example.

**1. Introduction.** Longitudinal data are ubiquitous in many scientific studies in biology, social science, economy and medicine. The major challenge in traditional longitudinal data analysis is how to construct more accurate estimates for regression coefficients by incorporating the within-subject correlation. Liang and Zeger (1986) proposed a generalized estimating equation (GEE) method to improve efficiency by using working correlation structure. Qu, Lindsay and Li (2000) proposed a quadratic inference function (QIF) approach to further improve the GEE method. Theoretical results for GEE and QIF have been well established by these authors for longitudinal data with fixed dimensional covariates.

In many scientific studies such as genomic studies and neuroscience researches, the dimension of covariates $d$ can far exceed the sample size $n$. Due to space limitation, we present two concrete motivating examples in the Supplementary Material, where $d$ is comparable to $n$ and $d$ is much larger than $n$. Motivated by these applications, it is of great interest to develop statistical inference procedures for longitudinal data with ultra-high dimensional covariates. Variable selection and model selection for longitudinal data have been studied by Wang and Qu (2009), Xue, Qu and Zhou (2010) and Ma, Song and Wang (2013) under the

Received May 2018; revised May 2019.
*MSC2020 subject classifications.* Primary 62F03; secondary 62F05.
*Key words and phrases.* False discovery rate, generalized estimating equation, quadratic inference function.

finite dimensional setting. Wang, Zhou and Qu (2012) proposed penalized GEE methods under the setting of $d = O(n)$. However, theories developed in the aforementioned works are not applicable for ultra-high dimensional setting with $\log d = o(n)$.

Some statistical inference procedures have been developed for independent and identically distributed (i.i.d.) observations with $\log d = o(n)$. van de Geer et al. (2014), Javanmard and Montanari (2013) and Zhang and Zhang (2014) developed a debiased estimator for i.i.d. data under linear and generalized linear models, and constructed confidence intervals for low-dimensional parameters. Ning and Liu (2017) proposed a hypothesis testing procedure based on a decorrelated score function method for i.i.d. data, and Fang, Ning and Liu (2017) further extended the method to the partial likelihood for survival data. These existing methods and theories are not applicable for longitudinal data under the high-dimensional setting, due to the following two challenges. First, the construction of the optimal QIF (or GEE) depends on the existence of the inverse of the sample covariance matrix of a set of high-dimensional estimating equations (Qu, Lindsay and Li (2000)). When the number of features is greater than the sample size, the matrix is not invertible, and, therefore, the quadratic inference function is not well defined. Second, the existing estimation result (Wang, Zhou and Qu (2012)) does not hold under the regime $\log d = o(n)$ so that their penalized estimator cannot be used as the initial estimator for asymptotic inference. Due to these difficulties, the existing debiased and decorrelation methods are not applicable to the quadratic inference function for ultra-high dimensional longitudinal data.

In this paper we propose a new inference procedure for longitudinal data under the regime $\log d = o(n)$ by decorrelating the QIF. We first consider how to construct confidence intervals and hypothesis tests for a low-dimensional parameter of interest. Specifically, we start by constructing multiple decorrelated quasi-score functions following the generalized estimating equations (GEE) instead of the likelihood or partial-likelihood function developed in the literature. Each decorrelated quasi-score function aims to capture a particular correlation pattern of the repeated measurements, specified by a basis of correlation matrices. Unlike Wang, Zhou and Qu (2012), who estimated the nuisance parameters by penalized generalized estimating equations with unstructured correlation matrix, we estimate the nuisance parameter under the working independence assumption. This is crucial to guarantee the fast rate of convergence of a preliminary estimator under the regime $\log d = o(n)$. Then, we propose to optimally combine the multiple decorrelated quasi-score functions to improve the efficiency of the inference procedures using the generalized method of moment. The resulting loss function is a quadratic form of the decorrelated quasi-score functions, and, therefore, we call it quadratic decorrelated inference function (QDIF). Since the dimensionality of the estimating equations is reduced by using the decorrelated quasi-score functions, its sample covariance matrix is invertible with high probability. Thus, the proposed QDIF is always well defined, whereas the QIF may not exist in high dimensions. In theory, the asymptotic properties of the estimator corresponding to QDIF are studied in the following two regimes. First, when the parameter of interest is of fixed dimension, we show that the proposed estimator is asymptotically normal and attains the semiparametric information bound, based on which we can construct an optimal Wald test statistic. Second, when the dimension of the parameter of interest grows with the sample size at a polynomial rate, we give the characterization of the limiting distribution of the proposed estimator and the associated test statistic.

To further broaden the applicability of the proposed method, we study the following multiple testing problem:

$$H_{0j} : \beta_j^* = 0 \quad \text{versus} \quad H_{1j} : \beta_j^* \neq 0,$$

for $j = 1, \ldots, d$, where $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_d)^T$ is regression coefficient vector. The null hypothesis $H_{0j}$ is rejected if our test statistic for $\beta_j^*$ is greater than a cutoff. To guarantee most

of the rejected null hypothesis being real discoveries, we aim to control the false discovery rate (FDR) within a given significance level by choosing a suitable cutoff for our test statistics. Due to the correlation among repeated measurements, the test statistics for different null hypothesis $H_{0j}$ become correlated which makes the FDR control challenging. While the Benjamini–Hochberg method can control FDR if the test statistics are independent or positively dependent (Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)), unfortunately, these dependence structures do not hold for our test statistics. To control FDR, we apply the procedure in Storey (2002) which is known to be more powerful than the Benjamini–Hochberg method. Our main result shows that the proposed method can control FDR asymptotically under the dependent test statistics. The intuition is that, by decorrelating the quasi-score function, the correlation among different test statistics becomes weak so that the correlation only contributes to the higher order terms in the FDR approximation which can be well controlled. The proof of this result relies on a moderate deviation lemma of Liu (2013), who applies the Benjamini–Hochberg procedure to control FDR under Gaussian graphical models. While the FDR control under linear models is recently studied by Barber and Candès (2015), Grazier G'Sell et al. (2016), the corresponding sequential procedure and the knockoff method cannot be directly extended to the longitudinal data, due to the dependence structure. To the best of our knowledge, how to control FDR in the analysis of longitudinal data under the generalized linear model remains an open problem. Finally, we note that the proposed method is a general recipe for correlated data which can be easily modified to handle family data and clustered data. To facilitate the presentation, we consider the longitudinal data throughout the paper.

*Paper organization.* The rest of this paper is organized as follows. In Section 2 we propose the QDIF method and the resulting estimator. We further derive the asymptotic distribution of the estimator and construct the test statistic and confidence interval. In Section 3 we consider the FDR control problem. In Section 4 we investigate the empirical performance of the proposed methods using simulation examples and real data example. The proof and technical details are deferred to Appendix. Proofs of technical lemmas are given in the Supplementary Material of this paper (Fang, Ning and Li (2020)).

*Notation.* We adopt the following notation throughout this paper. For a vector $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$ and $1 \le q \le \infty$, we define $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, and let $\|\mathbf{v}\|_0 = |\operatorname{supp}(\mathbf{v})|$, where $\operatorname{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set $A$. Denote $\|\mathbf{v}\|_\infty = \max_{1\le i\le d} |v_i|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. For a matrix $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$, let $\|\mathbf{M}\|_{\max} = \max_{1\le j,k\le d} |M_{jk}|$, $\|\mathbf{M}\|_1 = \sum_{jk} |M_{jk}|$ and $\|\mathbf{M}\|_\infty = \max_j \sum_k |M_{jk}|$. If the matrix $\mathbf{M}$ is symmetric, we let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be the minimal and maximal eigenvalues of $\mathbf{M}$, respectively. We denote by $\mathbf{I}_d$ the $d \times d$ identity matrix. For $S \subseteq \{1, \ldots, d\}$, let $\mathbf{v}_S = \{v_j : j \in S\}$, and let $\bar{S}$ be the complement of $S$. The gradient and subgradient of a function $f(\mathbf{x})$ are denoted by $\nabla f(\mathbf{x})$ and $\partial f(\mathbf{x})$, respectively. Let $\nabla_S f(\mathbf{x})$ denote the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}_S$. For two positive sequences, $a_n$ and $b_n$, we write $a_n \asymp b_n$ if $C \le a_n/b_n \le C'$ for some constants $C, C' > 0$. Similarly, we use $a \lesssim b$ to denote $a \le Cb$ for some constant $C > 0$. For a sequence of random variables $X_n$, we write $X_n \rightsquigarrow X$, for some random variable $X$ if $X_n$ converges weakly to $X$, and we write $X_n \to_p a$, for some constant $a$ if $X_n$ converges in probability to $a$. Given $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of $a$ and $b$, respectively. For notational simplicity, we use $C$, $C'$ to denote generic constants, and their values may change from line to line.

**2. Inference in high-dimensional longitudinal data.** Let $Y_{ij}$ denote the response variable for the $j$th observation of the $i$th subject, where $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Let $\boldsymbol{X}_{ij} \in \mathbb{R}^d$ denote the corresponding $d$-dimensional covariates. Our proposed procedures are still directly applicable for the setting in which $m_i$s are different from subject to subject, but

the correlation structure such as the AR and compound symmetry retains the same. We refer to the Supplementary Material for further discussion. In most applications $m$ is relatively small comparing with $n$ and $d$, and we assume throughout the paper that $m$ is fixed.

Denote by $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im})^T \in \mathbb{R}^{m\times d}$ and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^T \in \mathbb{R}^m$. We further assume that $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$ are independent, while the within-subject observations are correlated.

2.1. *Quadratic inference function in low-dimensional setting.* Under the framework of generalized linear models, we assume that the regression function follows $\mathbb{E}(Y_{ij} \mid \boldsymbol{X}_{ij}) = \mu_{ij}(\eta_{ij}^*)$, where $\mu_{ij}(\cdot)$ is a known function and $\eta_{ij}^* = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^*$ being the regression coefficient vector. Liang and Zeger (1986) proposed the GEE method to incorporate the within subject correlation to improve the estimation efficiency of $\boldsymbol{\beta}^*$. A brief description of this method is given in the Section S.2 of the supplementary material (Fang, Ning and Li (2020)). The GEE yields consistent estimators for any working correlation structure, while the resulting estimator can be far less efficient when the working correlation structure is misspecified. To overcome this drawback, Qu, Lindsay and Li (2000) proposed an alternative approach, called QIF, which avoids direct estimation of the correlation structure and provides optimal estimator even if the working correlation structure is misspecified. Denote by $\eta_i = (\eta_{i1}, \ldots, \eta_{im})^T = (\boldsymbol{X}_{i1}^T\boldsymbol{\beta}, \ldots, \boldsymbol{X}_{im}^T\boldsymbol{\beta})^T \in \mathbb{R}^m$, $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \{\mu_{i1}(\eta_{i1}), \ldots, \mu_{im}(\eta_{im})\}^T \in \mathbb{R}^m$, and $\mathbf{V}_i = \mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ is the true covariance matrix of $\boldsymbol{Y}_i$. We can decompose $\mathbf{V}_i$ as $\mathbf{V}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{R}\mathbf{A}_i^{1/2}(\boldsymbol{\beta})$. Here, $\mathbf{R}$ is the corresponding correlation matrix, and $\mathbf{A}_i(\boldsymbol{\beta})$ is a diagonal matrix in which the $(j, j)$th entry is the variance of $Y_{ij}$, given the covariates, and can be written as $[\mathbf{A}_i(\boldsymbol{\beta})]_{jj} = \phi V_{ij}(\mu_{ij})$, where $\phi$ is the dispersion parameter and $V_{ij}(\cdot)$ is a given variance function. We further assume that $V_{ij}(\mu_{ij}(\eta)) = \mu_{ij}'(\eta)$ which corresponds to the canonical link function under generalized linear models (while we do not impose the distributional assumptions as in GLMs). As seen later, the quasi-score function (2.1) is proportional to the dispersion parameter $\phi$, and thus the root of the quasi-score function does not depend on $\phi$. For simplicity, we assume $\phi = 1$ in the rest of the paper.

In QIF it is assumed that $\mathbf{R}^{-1}$ can be approximated by the linear space generated by some known basis matrices $\mathbf{T}_1, \ldots, \mathbf{T}_K$, that is, $\sum_{k=1}^K a_k\mathbf{T}_k$, where $a_1, \ldots, a_K$ are unknown parameters. Given these basis matrices, the quasi-score function of $\boldsymbol{\beta}$ is defined as

$$\mathbf{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta}) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \boldsymbol{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{T}_1\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \\ \vdots \\ \sum_{i=1}^n \boldsymbol{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta})\mathbf{T}_K\mathbf{A}_i^{-1/2}(\boldsymbol{\beta})\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \end{pmatrix}. \tag{2.1}$$

The QIF proposed by Qu, Lindsay and Li (2000) is

$$Q_n(\boldsymbol{\beta}) = \mathbf{g}_n^T(\boldsymbol{\beta})\mathbf{C}_n^{-1}\mathbf{g}_n(\boldsymbol{\beta}) \quad \text{where } \mathbf{C}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^T(\boldsymbol{\beta}), \tag{2.2}$$

which combines the quasi-score function $\mathbf{g}_n(\boldsymbol{\beta})$ using the generalized method of moment. Naturally, we estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}). \tag{2.3}$$

Qu, Lindsay and Li (2000) showed that $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and efficient under the classical fixed dimensional regime. The "large $n$, diverging $d$" asymptotics is studied under the generalized additive partial linear models by Wang et al. (2014) when $d = o(n^{1/5})$. The variable selection consistency of the penalized QIF estimator is established under the same conditions.

While the estimation and variable selection properties of the penalized GEE and QIF methods have been investigated under the regime $d = O(n^\alpha)$ for some $\alpha \leq 1$, how to perform optimal estimation and inference by incorporating the unknown correlation structure remains a challenging problem under the ultra-high dimensional regime, that is, $\log d = o(n^\alpha)$ for some $\alpha > 0$. In particular, to optimally combine the quasi-score function in QIF, one has to calculate $\mathbf{C}_n^{-1}$ in (2.2). However, $\mathbf{C}_n^{-1}$ does not exist when $d > n$. This is the main difficulty for extending the QIF method to high-dimensional data.

2.2. *Optimal inference under high-dimensional setting.* In this section we consider how to make inference on a low-dimensional component of the parameter $\boldsymbol{\beta}$ in longitudinal data. We focus on the high-dimensional regime, that is, $\log d = o(n^\alpha)$ for some $\alpha > 0$ which is a more challenging setting in comparison with existing works. For ease of presentation, we partition the vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T)^T$, where $\boldsymbol{\theta}$ is a $d_0$-dimensional parameter of interest with $d_0 \ll n$ and $\boldsymbol{\gamma}$ is a high-dimensional nuisance parameter with dimension $d - d_0$. Our goal is to construct the confidence region and test the hypotheses $H_0 : \boldsymbol{\theta}^* = 0$ vs. $H_1 : \boldsymbol{\theta}^* \neq 0$. Similarly, we denote the corresponding partition of $\boldsymbol{X}_i$ by $\boldsymbol{X}_i = (\boldsymbol{Z}_i, \boldsymbol{U}_i)$, where $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im})^T \in \mathbb{R}^{m \times d_0}$ and $\boldsymbol{U}_i = (\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{im})^T \in \mathbb{R}^{m \times (d-d_0)}$. In this section we assume that there exists an initial estimator $\widehat{\boldsymbol{\beta}}$ which converges to the true $\boldsymbol{\beta}^*$ sufficiently fast. Section 2.3 presents a procedure to construct such an initial estimator $\widehat{\boldsymbol{\beta}}$.

Before we propose the new procedure, we note that inference in high-dimensional problems has been studied under the linear and generalized linear models with independent data (Javanmard and Montanari (2013), Ning and Liu (2017), van de Geer et al. (2014), Zhang and Zhang (2014)). Their methods require the existence of a (pseudo)-likelihood function and a penalized estimator such as Lasso. One may attempt to apply their methods to the associated quasi likelihood of longitudinal data. However, this simple approach is only feasible under the working independence assumption and in general leads to suboptimal results as the within-subject correlation is ignored (Liang and Zeger (1986)). To increase the efficiency, one may incorporate the within-subject correlation and apply their methods to the quadratic inference function $Q_n$ in (2.2). As explained above, the matrix $\mathbf{C}_n$ in (2.2) is not invertible in high dimensions, and the function $Q_n$ is not well defined. Thus, we cannot directly apply the existing methods for efficient inference in high-dimensional longitudinal data.

To address the challenges, we propose a novel quadratic decorrelated inference function (QDIF) approach. Our proposed method relies on the generalized estimating equations and is distinguished from the methods that directly correct the bias of the Lasso type estimators. Instead, we modify the decorrelation idea in Ning and Liu (2017) to construct estimating equations that are insensitive to the impact of high-dimensional nuisance parameters. As aforementioned, how to design the decorrelation step is challenging in the setting of high-dimensional longitudinal data, as a (pseudo)-likelihood function is not available. Unlike the decorrelated score function constructed from the likelihood in Ning and Liu (2017), we construct a decorrelated quasi-score function directly from the generalized estimating equations in (2.1). Borrowing the idea from the QIF method, we replace the inverse of correlation matrix $\mathbf{R}^{-1}$ in GEE by $\sum_{k=1}^K a_k \mathbf{T}_k$, for some unknown parameters $a_1, \ldots, a_K$ and some pre-specified basis matrices $\mathbf{T}_1, \ldots, \mathbf{T}_K$. For any $1 \leq i \leq n$ and $1 \leq k \leq K$, we define the decorrelated quasi-score function for subject $i$ with correlation basis $\mathbf{T}_k$ as

$$\widehat{S}_{ik}(\boldsymbol{\theta}) = (\boldsymbol{Z}_i - \boldsymbol{U}_i \widehat{\mathbf{W}}_k)^T \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}) \mathbf{T}_k \mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}) \{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}})\}, \tag{2.4}$$

where $\widehat{\mathbf{W}}_k \in \mathbb{R}^{(d-d_0) \times d_0}$, to be defined later, is an estimator of the decorrelation matrix for the $k$th basis $\mathbf{T}_k$ and $\widehat{\boldsymbol{\beta}} := (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}^T)^T$ is an initial estimator defined in Section 2.3. In comparison with the component of the standard quasi-score function $\mathbf{g}_n(\boldsymbol{\beta})$ in (2.1), $\widehat{S}_{ik}(\boldsymbol{\theta})$ decorrelates

the score functions for $\boldsymbol{Z}_i$ and $\boldsymbol{U}_i$ by projection via $\widehat{\mathbf{W}}_k$. Denote

$$\widehat{\mathbf{H}}_{k\boldsymbol{\gamma}\boldsymbol{\theta}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{U}_i^T \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\boldsymbol{Z}_i,$$

$$\widehat{\mathbf{H}}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{U}_i^T \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\boldsymbol{U}_i,$$

$$\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}} = \mathbb{E}\{\boldsymbol{U}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{Z}_i\},$$

$$\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}} = \mathbb{E}\{\boldsymbol{U}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\boldsymbol{U}_i\},$$

and $\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}$ is defined similarly. Then, we define the estimator $\widehat{\mathbf{W}}_k$ in (2.4) as

$$\begin{aligned}\widehat{\mathbf{W}}_k = \operatorname*{argmin}_{\mathbf{W}} \frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}\{(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W})^T \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})\mathbf{T}_k\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}})(\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W})\} \\ + \lambda' \sum_{k,l} |w_{kl}|,\end{aligned} \tag{2.5}$$

where $w_{kl}$ is the $(k,l)$th element of $\mathbf{W}$, and $\lambda'$ is a tuning parameter. This estimator $\widehat{\mathbf{W}}_k$ is introduced to estimate the true decorrelation matrix

$$\mathbf{W}_k^* = \mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{k\boldsymbol{\gamma}\boldsymbol{\theta}}. \tag{2.6}$$

Then, we define the decorrelated quasi-score function of $\boldsymbol{\theta}$ by combining $\widehat{S}_{ik}(\boldsymbol{\theta})$'s from the different basis matrices,

$$\overline{\mathbf{S}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{S}}_i(\boldsymbol{\theta}) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}\widehat{S}_{i1}(\boldsymbol{\theta}) \\ \vdots \\ \sum_{i=1}^{n}\widehat{S}_{iK}(\boldsymbol{\theta}) \end{pmatrix}. \tag{2.7}$$

Note that the decorrelated quasi-score function $\overline{\mathbf{S}}_n(\boldsymbol{\theta})$ is of dimension $d_0K$ instead of dimension $dK$ as $\mathbf{g}_n(\boldsymbol{\beta})$ in (2.1). As $d_0 \cdot K \ll n$ in our setting, this decorrelated quasi-score function can be used to define the optimal quadratic inference function $\widetilde{Q}_n(\boldsymbol{\theta})$. In particular, given our initial estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\theta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$, we define our QDIF estimator as

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta_n} \widetilde{Q}_n(\boldsymbol{\theta}) \quad \text{where } \widetilde{Q}_n(\boldsymbol{\theta}) = n\overline{\mathbf{S}}_n(\boldsymbol{\theta})^T\widehat{\mathbf{C}}^{-1}\overline{\mathbf{S}}_n(\boldsymbol{\theta}). \tag{2.8}$$

Here, $\Theta_n = \{\boldsymbol{\theta} \in \mathbb{R}^{d_0} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \le cd_0^{-1/2}\}$ is a neighborhood around the initial estimator $\widehat{\boldsymbol{\theta}}$ for some small constant $c > 0$, and

$$\widehat{\mathbf{C}} = n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{S}}_i(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{S}}_i^T(\widehat{\boldsymbol{\theta}}) \in \mathbb{R}^{d_0K\times d_0K}. \tag{2.9}$$

Since $\widetilde{Q}_n(\boldsymbol{\theta})$ is generally a nonconvex function of $\boldsymbol{\theta}$, there may exist multiple local solutions, especially when $d_0$ is large. To alleviate these issues, we propose the above localized estimator by minimizing $\widetilde{Q}_n(\boldsymbol{\theta})$ in a small neighborhood around the initial estimator $\widehat{\boldsymbol{\theta}}$. In the theoretical analysis, we show that $\widetilde{Q}_n(\boldsymbol{\theta})$ is strongly convex for $\boldsymbol{\theta} \in \Theta_n$ with probability tending to one. Thus, any off-the-shelf convex optimization algorithm is applicable to solving the problem (2.8).

2.3. *An initial estimator based on working independence structure.* As seen in the previous section, the decorrelated quasi-score function (2.4) requires the knowledge of an initial estimator of $\boldsymbol{\beta}$. In this subsection, we shall construct such an initial estimator in the ultra-high dimensional regime, that is, $\log d = o(n^{\alpha})$ for some $\alpha > 0$.

Since the initial estimator is only used to estimate the nuisance parameter in (2.4), we allow the estimator to be less efficient in terms of incorporating the within-subject dependence. The following penalized maximum quasi-loglikelihood estimator under the working independence structure serves this purpose,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \mathcal{P}_{\lambda}(\boldsymbol{\beta}), \tag{2.10}$$

where

$$\mathcal{L}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{Y_{ij}}^{\mu_{ij}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta})} \frac{Y_{ij}-u}{V_{ij}(u)}\,du.$$

Here, $\mathcal{L}_n(\boldsymbol{\beta})$ is known as the negative quasi loglikelihood under the working independence assumption, and $\mathcal{P}_{\lambda}(\cdot)$ is a penalization term encouraging sparsity of $\widehat{\boldsymbol{\beta}}$ with some tuning parameter $\lambda > 0$. The penalization term can be either convex, such as Lasso (Tibshirani (1996)), or nonconvex (e.g., SCAD (Fan and Li (2001))). Before we pursue the statistical properties of $\widehat{\boldsymbol{\beta}}$ further, let us introduce some definitions.

DEFINITION 1 (Subexponential variable and subexponential norm). A random variable $X$ is called subexponential if there exists some positive constant $K_1$ such that $\mathbb{P}(|X| > t) \le \exp(1 - t/K_1)$ for all $t \ge 0$. The subexponential norm of $X$ is defined as $\|X\|_{\psi_1} = \sup_{p\ge 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.

Furthermore, denoting by $s = \|\boldsymbol{\beta}^*\|_0$ the sparsity of $\boldsymbol{\beta}^*$, we impose the following assumptions:

ASSUMPTION 2.1. Assume that the error $\epsilon_{ij} = Y_{ij} - \mu_{ij}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*)$ is subexponential, that is, $\|\epsilon_{ij}\|_{\psi_1} \le C$ for some constant $C > 0$. The covariates are uniformly bounded, $\max_{i\in[n]} \|\boldsymbol{X}_i\|_{\infty} = \mathcal{O}(1)$.

Note that the subexponential and bounded covariate assumptions are technical assumptions in concentration inequalities and hold in most applications. Similar assumptions are widely used in the literature, for example, van de Geer et al. (2014) and Ning and Liu (2017), for analyzing high-dimensional generalized linear models.

ASSUMPTION 2.2. For any set $\mathcal{S} \subset \{1, \dots, d\}$, where $|\mathcal{S}| \asymp s$ and any vector $\mathbf{v}$ belonging to the cone $\mathcal{C}(\xi, \mathcal{S}) = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}_{\bar{\mathcal{S}}}\|_1 \le \xi\|\mathbf{v}_{\mathcal{S}}\|_1\}$, it holds that

$$\mathrm{RE}(\xi, \mathcal{S}; \nabla^2\mathcal{L}_n(\boldsymbol{\beta}^*)) = \inf_{\mathbf{0}\neq\mathbf{v}\in\mathcal{C}(\xi,\mathcal{S})} \frac{\mathbf{v}^T\nabla^2\mathcal{L}_n(\boldsymbol{\beta}^*)\mathbf{v}}{\|\mathbf{v}_{\mathcal{S}}\|_2^2} \ge \lambda_{\min} > 0.$$

This assumption is known as the restricted eigenvalue condition (Bühlmann and van de Geer (2011)) and provides the necessary curvature of the loss function within a cone. Specifically, it bounds the minimal eigenvalue of the Hessian matrix $\nabla^2\mathcal{L}_n(\boldsymbol{\beta}^*)$ from below within the cone $\mathcal{C}(\xi, \mathcal{S})$. Under Assumptions 2.1 and 2.2 and the technical conditions in Section 2.4,

if $\lambda \asymp \sqrt{n^{-1}\log d}$ and $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, a simple modification of Theorem 5.2 in van de Geer and Müller (2012) implies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q = \mathcal{O}_{\mathbb{P}}\Big(s^{1/q}\sqrt{\frac{\log d}{n}}\Big) \quad \text{for } q = 1, 2, \tag{2.11}$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}[\boldsymbol{X}_{ij}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 = \mathcal{O}_{\mathbb{P}}\Big(\frac{s\log d}{n}\Big). \tag{2.12}$$

The rate in (2.11) shows that even if we ignore the correlation structure, the penalized maximum quasi-loglikelihood estimator still converges to $\boldsymbol{\beta}^*$ with the optimal rate in the high-dimensional regime. Assumption 2.2 can be further relaxed by using nonconvex penalties and more tailored statistical optimization algorithms as discussed in Fan et al. (2018), Loh and Wainwright (2013), Wang, Kim and Li (2013), Zhao, Liu and Zhang (2018). It can be easily verified that Assumption 2.2 holds under the conditions in the next subsection and the minimum eigenvalue condition on the population matrix $\mathbb{E}(\boldsymbol{X}_i^T\boldsymbol{X}_i)$. For ease of presentation, we assume that Assumptions 2.1 and 2.2 hold throughout our later discussion, and, therefore, the rate of convergence (2.11) and (2.12) is achieved.

2.4. *Theoretical properties.* In this subsection we establish the asymptotic distribution of $\widetilde{\boldsymbol{\theta}}$ in (2.8). In the analysis we assume $m$, $K$ are fixed, and $n$, $d$ increase to infinity with $\log d = o(n^\alpha)$ for some $\alpha > 0$. To make the proposed framework more flexible, we allow $d_0$ to diverge together with $n$ and $d$. We note that the theoretical results also hold for fixed $d_0$.

To facilitate our discussion, we impose some technical assumptions. Let

$$S_{ik}^*(\boldsymbol{\theta}) = (\boldsymbol{Z}_i - \boldsymbol{U}_i\mathbf{W}_k^*)^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}^*)\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}, \boldsymbol{\gamma}^*)\}, \tag{2.13}$$

$$\mathbf{S}_i^*(\boldsymbol{\theta}) = \big(S_{i1}^{*T}(\boldsymbol{\theta}), \ldots, S_{iK}^{*T}(\boldsymbol{\theta})\big)^T \in \mathbb{R}^{d_0K}, \tag{2.14}$$

be the "ideal" versions of $\widehat{S}_{ik}(\boldsymbol{\theta})$ and $\overline{\mathbf{S}}_n(\boldsymbol{\theta})$, respectively. Also, we let

$$\mathbf{g}_0(\boldsymbol{\theta}^*) = \mathbb{E}\{\nabla\mathbf{S}_i^*(\boldsymbol{\theta}^*)\} \tag{2.15}$$

denote the population version of the gradient of $\mathbf{S}^*(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$, and let

$$\mathbf{C}^* = \mathbb{E}\{\mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*)\} \tag{2.16}$$

denote the population version of the matrix $\widehat{\mathbf{C}}$.

ASSUMPTION 2.3. The decorrelation matrix $\mathbf{W}_k^*$ is column-wise sparse, that is, $\max_{\ell\in[d_0]}\|(\mathbf{W}_k^*)_{\cdot\ell}\|_0 \le s'$, for $1 \le k \le K$. $\max_{k\in[K]}\max_{i\in[n]}\|\boldsymbol{U}_i\mathbf{W}_k^*\|_\infty = \mathcal{O}(1)$ and $\max_{i\in[n]}\max_{j\in[m]}|\boldsymbol{X}_{ij}^T\boldsymbol{\beta}^*| = \mathcal{O}(1)$.

ASSUMPTION 2.4. The mean function $\mu_{ij}(\cdot)$ is third order continuously differentiable and satisfies

$$\max_{i\in[n],\,j\in[m]}\{\mu_{ij}'(\eta_{ij}^*), 1/\mu_{ij}'(\eta_{ij}^*), \mu_{ij}''(\eta_{ij}^*), \mu_{ij}'''(\eta_{ij}^*)\} = \mathcal{O}(1).$$

ASSUMPTION 2.5. The eigenvalues of $\mathbf{T}_k$ and $\mathbf{C}^*$ are bounded, that is, $C^{-1} \le \lambda_{\min}(\mathbf{T}_k) \le \lambda_{\max}(\mathbf{T}_k) \le C$ for any $1 \le k \le K$, $C^{-1} \le \lambda_{\min}(\mathbf{C}^*) \le \lambda_{\max}(\mathbf{C}^*) \le C$. In addition, the following eigenvalue conditions on the design matrix hold, $\lambda_{\min}(\mathbb{E}(\boldsymbol{X}_i^T\boldsymbol{X}_i)) \ge C^{-1}$ and $\lambda_{\max}(\mathbb{E}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)) \le C$ for some constant $C > 0$.

Assumption 2.3 specifies the sparsity of $\mathbf{W}_k^*$ and the bounded covariate effect which ensure the fast rate of convergence of the estimators $\widehat{\mathbf{W}}_k$'s. To understand the sparsity assumption on $\mathbf{W}_k^*$, let us consider $d_0 = 1$. Denote $\widetilde{\boldsymbol{Z}}_i = \mathbf{A}^{1/2}\boldsymbol{Z}_i$ and $\widetilde{\boldsymbol{U}}_i = \mathbf{A}^{1/2}\boldsymbol{U}_i$. If there exists $\mathbf{W}^* \in \mathbb{R}^{d-1}$ such that

$$\widetilde{\boldsymbol{Z}}_i = \widetilde{\boldsymbol{U}}_i\mathbf{W}^* + \boldsymbol{\delta} \quad \text{and} \quad \mathbb{E}(\boldsymbol{\delta}|\widetilde{\boldsymbol{U}}_i) = 0, \tag{2.17}$$

we can verify that $\mathbf{W}_k^* = \mathbf{W}^*$ for any $1 \le k \le K$. For instance, if $\mu_{ij}(\eta_{ij})$ is a quadratic function (corresponding to the Gaussian response) and $(Z_{ij}, \boldsymbol{U}_{ij}^T)^T \sim N(0, \boldsymbol{\Sigma})$ for $1 \le j \le m$ follows the Gaussian design, then (2.17) holds and the sparsity assumption on $\mathbf{W}_k^*$ (and $\mathbf{W}^*$ in (2.17)) is implied by the sparsity of $\boldsymbol{\Sigma}^{-1}$ which is a standard condition for high-dimensional inference in the generalized linear model (van de Geer et al. (2014)).

Assumption 2.4 provides some local smoothness conditions of $\boldsymbol{\mu}_i(\cdot)$'s around $\eta_i^*$'s, and it is easy to verify that this assumption is satisfied for many commonly used regression functions $\mu_{ij}(\cdot)$. In Assumption 2.5 we require the basis matrices $\mathbf{T}_k$ to be positive definite. In practice, we usually choose the following matrices as the bases: $\mathbf{T}_1$ an identity matrix $\mathbf{I}_m$, $\widetilde{\mathbf{T}}_2$ a matrix of diagonal elements set to be 0's and off-diagonal elements set to be 1's, $\widetilde{\mathbf{T}}_3$ a matrix with two main diagonals set to be 1's and 0's elsewhere and $\widetilde{\mathbf{T}}_4$ with 1's on the corners $(1, 1)$ and $(m, m)$ and 0 elsewhere. As shown by Qu, Lindsay and Li (2000), the commonly used equal correlation and AR(1) models can be written as the linear combination of the above four basis matrices. However, the matrices $\widetilde{\mathbf{T}}_2$, $\widetilde{\mathbf{T}}_3$ and $\widetilde{\mathbf{T}}_4$ are not positive definite. To meet Assumption 2.5, we can add an identity matrix to $\widetilde{\mathbf{T}}_j$ and define $\mathbf{T}_j = \widetilde{\mathbf{T}}_j + \lambda\mathbf{I}_m$ for some constant $\lambda > 0$. The eigenvalue condition on $\mathbf{C}^*$ in Assumption 2.5, as we shall see later, guarantees the existence of the asymptotic variance which is even essential in the low-dimensional setting. Finally, the minimum eigenvalue condition on $\mathbb{E}(\boldsymbol{X}_i^T\boldsymbol{X}_i)$ is used to verify the restricted eigenvalue condition for $\widehat{\mathbf{W}}_k$, and the maximum eigenvalue condition on $\mathbb{E}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)$ is used to control $\|\mathbf{H}_{k\boldsymbol{\theta}\boldsymbol{\theta}}\|_2$, especially when $d_0$ diverges.

Denote $\|\mathbf{A}\|_{L_1} := \max_j \sum_i |A_{ij}|$ to be the maximum absolute column sum of the matrix $\mathbf{A}$. The following lemma shows the rate of convergence of the estimation and prediction errors of $\widehat{\mathbf{W}}_k$:

LEMMA 2.6. *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5, if* $\lambda \asymp \lambda' \asymp \sqrt{n^{-1}\log d}$ *and* $s\sqrt{\frac{\log d}{n}} = o(1)$, *we have*

$$\sup_{1\le k\le K} \|\widehat{\mathbf{W}}_k - \mathbf{W}_k^*\|_{L_1} = \mathcal{O}_{\mathbb{P}}\bigg(\max(s, s')\sqrt{\frac{\log d}{n}}\bigg),$$

$$\sup_{1\le k\le K} \operatorname{tr}\bigg(\frac{1}{n}\sum_{i=1}^n [\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]^T\boldsymbol{\Psi}_i[\boldsymbol{U}_i(\widehat{\mathbf{W}}_k - \mathbf{W}_k^*)]\bigg) = \mathcal{O}_{\mathbb{P}}\bigg(\frac{\max(s, s')d_0\log d}{n}\bigg),$$

*where* $\boldsymbol{\Psi}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)\mathbf{T}_k\mathbf{A}_i^{1/2}(\boldsymbol{\beta}^*)$.

Based on Lemma 2.6, we first establish the rate of convergence of the decorrelated estimator $\widetilde{\boldsymbol{\theta}}$ in (2.8).

THEOREM 2.7. *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5, if* $\lambda \asymp \lambda' \asymp \sqrt{n^{-1}\log d}$, *and as* $n, d \to \infty$,

$$\max\bigg\{\frac{d_0}{n^{1/2}}, \frac{\{d_0(s\vee s')\log d(\log n)^2\}^{1/2}}{n^{1/2}}, \frac{(s\vee s')\log d\log n}{n^{1/2}}\bigg\} = o(1), \tag{2.18}$$

*then the rate of convergence of the estimator* $\widetilde{\boldsymbol{\theta}}$ *is*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0/n}). \tag{2.19}$$

When the dimension of $\boldsymbol{\theta}^*$ diverges, the convergence rate (2.19) is comparable to Theorem 3.6 of Wang (2011) in which the author showed that the convergence rate of the GEE estimator is $\mathcal{O}_{\mathbb{P}}(\sqrt{d/n})$ with diverging number of covariates $d = o(n^{1/2})$. This also agrees with our condition $d_0 = o(n^{1/2})$ in (2.18). When $d_0$ is fixed, (2.19) implies that the estimator has the standard root-$n$ rate under the sparsity condition $(s \vee s')\log d \log n = o(n^{1/2})$ which agrees with the weakest possible assumption in the literature for generalized linear models up to logarithmic factors in $d$ and $n$ (van de Geer et al. (2014)).

In order to conduct valid inference, we need to understand the asymptotic distribution of the estimator. The following main theorem of this section establishes the asymptotic normality of our estimator $\widetilde{\boldsymbol{\theta}}$.

THEOREM 2.8. *Under Assumptions* 2.1, 2.2, 2.3, 2.4 *and* 2.5, *if* $\lambda \asymp \lambda' \asymp \sqrt{n^{-1}\log d}$, *and as* $n, d \to \infty$,

$$\max\left\{\frac{d_0^{3/2}}{n^{1/2}}, \frac{d_0\{(s \vee s')\log d(\log n)^2\}^{1/2}}{n^{1/2}}, \frac{d_0^{1/2}(s \vee s')\log d \log n}{n^{1/2}}\right\} = o(1), \tag{2.20}$$

*then the estimator* $\widetilde{\boldsymbol{\theta}}$ *satisfies, as* $d_0 \to \infty$,

$$\frac{n(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}) - d_0}{\sqrt{2d_0}} \rightsquigarrow N(0, 1), \tag{2.21}$$

*where* $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \{\mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1}\mathbf{g}_0(\boldsymbol{\theta}^*)\}^{-1}$ *and* $\mathbf{g}_0(\boldsymbol{\theta})$ *is defined in* (2.15). *If* $d_0$ *is fixed, we have*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{2.22}$$

Theorem 2.8 characterizes the asymptotic distribution of the decorrelated estimator. In particular, we note that Theorem 2.8 holds whether or not the inverse of the within-subject correlation matrix $\mathbf{R}^{-1}$ is correctly specified via the basis matrices $\{\mathbf{T}_k\}$. Thus, similar to the classical QIF and GEE methods, our estimator is robust to the specification of the within-subject dependence structure.

When $d_0$ diverges, (2.21) can be interpreted as $n(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}) \rightsquigarrow \chi^2_{d_0}$, and one can further approximate $\chi^2_{d_0}$ by $N(d_0, 2d_0)$. To justify the normal approximation of the decorrelated estimator, the required condition $d_0 = o(n^{1/3})$ in (2.20) is stronger than that in Theorem 2.7, and again it is comparable to the condition in Theorem 3.8 of Wang (2011). When $d_0$ is fixed, (2.22) implies that the estimator is asymptotically normal under the same sparsity condition $(s \vee s')\log d \log n = o(n^{1/2})$ as in Theorem 2.7. Moreover, if $\mathbf{R}^{-1} = \sum_{k=1}^K a_k \mathbf{T}_k$ is correctly specified, as shown in Corollary 2.10, our estimator $\widetilde{\boldsymbol{\theta}}$ is semiparametrically efficient.

In order to use the above result to construct confidence regions and statistical tests, we need to estimate the asymptotic variance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ in (2.21). This can be accomplished by using the plug-in estimator

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \{\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1}\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})\}^{-1}. \tag{2.23}$$

LEMMA 2.9. *Under the same assumptions as in Theorem 2.8, we have as* $d_0 \to \infty$,

$$\frac{n(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}})^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}) - d_0}{\sqrt{2d_0}} \rightsquigarrow N(0, 1),$$

*where* $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ *and* $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ *are defined in* (2.21) *and* (2.23), *respectively. If* $d_0$ *is fixed*,

$$n(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}})^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}) \rightsquigarrow \chi^2_{d_0},$$

*where* $\chi^2_{d_0}$ *denotes the chi-square distribution with* $d_0$ *degrees of freedom.*

Consider the following hypothesis testing problem:

$$H_0 : \boldsymbol{\theta}^* = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta}^* \neq 0.$$

Based on the above result, we define the Wald-type test statistic as follows,

$$\widehat{T}_n = n\widetilde{\boldsymbol{\theta}}^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} \widetilde{\boldsymbol{\theta}}. \tag{2.24}$$

Lemma 2.9 implies that the distribution of the test statistic $\widehat{T}_n$ can be approximated by a chi-squared distribution with $d_0$ degrees of freedom under $H_0$. In addition, we can obtain an asymptotic $(1-\alpha) \times 100\%$ confidence region of $\boldsymbol{\theta}^*$:

$$\{\boldsymbol{\theta} : n(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \leq \chi^2_{d_0, 1-\alpha}\},$$

where $\chi^2_{d_0,1-\alpha}$ denotes the $(1-\alpha) \times 100\%$th percentile of a chi-square distribution with $d_0$ degrees of freedom.

To conclude this section, we compare the efficiency of the proposed estimator with the decorrelated estimator based on the quasi likelihood. The latter corresponds to the special case of the estimator (2.8) with $K = 1$ and $\mathbf{T}_1 = \mathbf{I}$. Consider the case $d_0 = 1$. We denote the estimator by $\widetilde{\theta}_I$. The following corollary shows that our estimator is more efficient than $\widetilde{\theta}_I$ and attains the semiparametric information bound. Thus, the proposed QDIF method provides the optimal inference for high-dimensional longitudinal data.

COROLLARY 2.10. *Assume that the assumptions in Theorem 2.8 hold. By taking* $\mathbf{T}_1 = \mathbf{I}$, *we obtain* $\text{Avar}(\widetilde{\theta}) \leq \text{Avar}(\widetilde{\theta}_I)$, *where* Avar *denotes the asymptotic variance of the estimator. Moreover, if the true correlation matrix* $\mathbf{R}$ *satisfies* $\mathbf{R}^{-1} = \sum_{k=1}^K a_k \mathbf{T}_k$ *for some constants* $a_1, \ldots, a_K$ *and* (2.17) *holds, then the proposed estimator* $\widetilde{\theta}$ *is semiparametrically efficient.*

**3. False discovery rate control.** In the previous section we develop valid inferential methods to test a low-dimensional parameter of interest in high-dimensional longitudinal data. However, in many practical applications, the parameter of interest may not be prespecified. Instead, we are interested in simultaneously testing all $d$ hypotheses with $\theta = \beta_j$, that is, $H_{0j} : \beta_j^* = 0$ versus $H_{1j} : \beta_j^* \neq 0$ for all $j = 1, \ldots, d$. The knowledge of which null hypotheses are rejected can help us identify important features in the longitudinal data. When conducting multiple hypothesis tests, it is a common practice to control the false discovery rate (FDR) to avoid spurious discoveries. Under the high-dimensional setting, due to the potential dependence among test statistics, how to control FDR is a challenging problem. In this direction, Liu (2013) and Barber and Candès (2015) applied the Benjamini–Hochberg procedure and the knockoff procedure to control the FDR under the Gaussian graphical model and linear model, respectively. Both of their methods crucially depend on the linearity structure and are not directly applicable to the generalized linear model, let alone generalized estimating equations for longitudinal data. Thus, the FDR control for high-dimensional longitudinal data is still largely unexplored while it is of substantial practical importance.

In this section we extend the procedure discussed in the previous section to control the FDR in multiple hypothesis testing for $H_{0j} : \beta_j^* = 0$ vs. $H_{1j} : \beta_j^* \neq 0$ where $j = 1, \ldots, d$. We first construct the test statistic $\Lambda_{nj} = n\widehat{\sigma}_j^{-2}\widetilde{\beta}_j^2$ for hypothesis $H_{0j}$, where $\widehat{\sigma}_j^2$ is defined as in (2.23). Then, we obtain the asymptotic p-value $P_j = 1 - \chi_1^2(\Lambda_{nj})$, where $\chi_1^2(\cdot)$ is the cumulative distribution function of a chi-squared random variable with degree of freedom 1. Given a decision rule that rejects $H_{0j}$ if and only if $P_j \leq u$ for some cutoff $u$, we define the false discovery proportion (FDP) and false discovery rate (FDR) as

$$\mathrm{FDP}(u) = \frac{\sum_{j \in \mathcal{S}_0} \mathbb{1}(P_j \leq u)}{\max\{\sum_{j \in [d]} \mathbb{1}(P_j \leq u), 1\}} \quad \text{and} \quad \mathrm{FDR}(u) = \mathbb{E}\big[\mathrm{FDP}(u)\big],$$

where $\mathcal{S}_0 = \{j : \beta_j^* = 0\}$ denotes the set of true null hypotheses. Given the desired FDR level $\alpha$, we aim to find the cutoff $\widehat{u}_\alpha$ such that $\mathrm{FDR}(\widehat{u}_\alpha) \leq \alpha$. However, in practice we cannot directly compute $\mathrm{FDP}(u)$ and $\mathrm{FDR}(u)$ as the set $\mathcal{S}_0$ is unknown. To approximate $\mathrm{FDP}(u)$, we utilize the following procedure proposed by Storey (2002) which is known to be more powerful than the Benjamini–Hochberg procedure. Let $t \in (0, 1)$ be a tuning parameter. For any $u \in (0, 1)$, we define

$$\widehat{\mathrm{FDP}}_t(u) := \frac{\pi(t) \cdot u \cdot d}{\max\{\sum_{j=1}^d \mathbb{1}\{P_j \leq u\}, 1\}} \tag{3.1}$$

as an approximation of $\mathrm{FDP}(u)$, where

$$\pi(t) := \min\left\{(1/d)\sum_{j=1}^d \mathbb{1}\{P_j \geq t\}/(1-t), 1\right\}.$$

Comparing (3.1) with $\mathrm{FDP}(u)$, the denominators are identical. For the numerator, by taking expectation we have that $\mathbb{E}(\sum_{j \in \mathcal{S}_0} \mathbb{1}(P_j \leq u)) \approx u \cdot |\mathcal{S}_0|$, as $P_j \sim \mathrm{Uniform}(0, 1)$ asymptotically for all $j \in \mathcal{S}_0$. It turns out the quantity $\pi(t)$ in (3.1) tends to slightly overestimate $|\mathcal{S}_0|/d$, the proportion of null hypotheses among all hypotheses. To see this, we have $\mathbb{E}[d \cdot \pi(t)] \approx |\mathcal{S}_0| + \sum_{j \notin \mathcal{S}_0} \mathbb{P}(P_j \geq t)/(1-t)$, where $\sum_{j \notin \mathcal{S}_0} \mathbb{P}(P_j \geq t)/(1-t)$ is usually small as $P_j$ are close to 0 if $j \notin \mathcal{S}_0$. This leads to a slightly conservative cutoff. However, we show in the proof that this overestimation is asymptotically negligible, that is, $|\mathcal{S}_0|/(\pi(t) \cdot d) \to 1$ in probability as $|\mathcal{S}_0|/d \to 1$ in the setting of a sparse high-dimensional model. Since $|\mathcal{S}_0|/d \leq 1$, we force $\pi(t) \leq 1$ by taking the minimum with 1 in the definition of $\pi(t)$.

Given $\widehat{\mathrm{FDP}}_t(u)$, we define $\widehat{u}_{\alpha,t}$ to be the largest value of $u$ such that $\widehat{\mathrm{FDR}}_t(u)$ is controlled at level $\alpha$:

$$\widehat{u}_{\alpha,t} := \sup\big\{0 \leq u \leq 1 : \widehat{\mathrm{FDP}}_t(u) \leq \alpha\big\}.$$

Then, we reject all the null hypotheses of which the corresponding p-values are below $\widehat{u}_{\alpha,t}$. It is easily seen that the well known Benjamini–Hochberg procedure corresponds to choosing $\pi(t) = 1$ in $\widehat{\mathrm{FDP}}_t(u)$. By introducing $\pi(t) \leq 1$, the cutoff of p-values $\widehat{u}_{\alpha,t}$ is no smaller than that in the Benjamini–Hochberg procedure and, therefore, leads to more discoveries. Thus, the method is more powerful than the Benjamini–Hochberg procedure. In the literature the theoretical justification of the Storey (2002) procedure requires that the p-values are independent and uniformly distributed under the null hypothesis. To prove the validity of the procedure, the main technical difficulty is that the p-values from the proposed test statistics are not independent, and their distribution holds only asymptotically. For $1 \leq j \leq d$, we define

$$A_{ij} = \{\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1}\mathbf{g}_{0j}(\beta_j^*)\}^{-1}\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1} \cdot \mathbf{S}_{ij}^*(\beta_j^*)/\sigma_j,$$

where $\mathbf{g}_{0j}^*$, $\mathbf{C}_j^*$ and $\mathbf{S}_{ij}^*(\beta_j^*)$ denote the corresponding $\mathbf{g}_0$, $\mathbf{C}^*$ and $\mathbf{S}_i^*(\beta_j^*)$ in the previous section for $\beta_j^*$ and $\sigma_j^2$ is defined in (2.21) with $d_0 = 1$. Denote $\mathcal{A} := \{(j,k) : j \neq k, |\Omega_{jk}| \geq (\log d)^{-5/2}\}$, where $\Omega_{jk}$ is the correlation between $A_{ij}$ and $A_{ik}$. For some constant $C > 0$, let $\mathcal{S}_1 = \{j : |\beta_j| \geq C\sqrt{\frac{\log d}{n}}\}$ denote the set of strong signals. As the main result in this section, the following theorem shows that our procedure controls the FDR and FDP asymptotically:

THEOREM 3.1. *Assume that Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 hold for all $\beta_j$ and $s \asymp s'$. In addition, assume $n^{-1/2}s(\log d)(\log n) = o(1/\log d)$, $s/d = o(1)$, $|\mathcal{S}_1| \to \infty$, $d = \mathcal{O}(n^C)$ for some constant $C > 0$, and*

$$\sum_{(j,k)\in\mathcal{A}} d^{-2(1+|\Omega_{jk}|)^{-1}} = o\big((\log d)^{-3/2}\big). \tag{3.2}$$

*Let $c > 0$ be a small constant. For any $t \in [c, 1)$, we have as $n, d \to \infty$,*

$$\mathrm{FDP}(\widehat{u}_{\alpha,t}) \leq \alpha \quad \textit{and} \quad \mathrm{FDR}(\widehat{u}_{\alpha,t}) \leq \alpha \quad \textit{in probability}.$$

In the following, we comment on the conditions in the theorem. First, we require $n^{-1/2}s(\log d)(\log n) = o(1/\log d)$ which is identical to the condition in Theorem 2.8 for fixed $d_0$ up to a logarithmic factor of $d$. This guarantees the validity of the p-values under the null, asymptotically. Since we consider the sparse model, most $\beta_j$'s are 0 implying $s/d = o(1)$. We also require the number of strong signals tends to infinity, $|\mathcal{S}_1| \to \infty$, which is needed to control FDP. We require $d = \mathcal{O}(n^C)$ to apply the moderate deviation lemma of Liu (2013). While we cannot allow $d$ to grow exponentially fast in $n$ as in Theorem 2.8, by choosing $C > 1$ $d$ can still be much larger than $n$. Finally, (3.2) imposes conditions on the correlation of the test statistics. Recall that $\Omega_{jk}$ is the correlation between $A_{ij}$ and $A_{ik}$. Denote $\mathcal{A}_j = \{1 \leq k \leq d : |\Omega_{jk}| \geq (\log d)^{-5/2}\}$. If $|\Omega_{jk}| \leq a$ for some constant $a < 1$, then (3.2) holds under the assumption $|\mathcal{A}_j| = o(d^{\frac{1-a}{1+a}-\delta})$ for some small $\delta > 0$. In particular, under (2.17) and Assumption 2.3, we can show that $|\mathcal{A}_j| \leq s$ and therefore (3.2) is true, provided $d$ is sufficiently large. Thus, the FDR and FDP control is still feasible under dependent test statistics, provided their correlation satisfies (3.2).

**4. Numerical studies.** In this section we evaluate the numerical performance of our proposed method by Monte Carlo simulation and a real data example. We further provide more simulation results in Supplementary Material Section S.4.

4.1. *Simulation studies.* We first assess the empirical performance of the proposed method using simulated data. In all settings we randomly select $s$ out of the $d$ components to be nonzero. We consider two settings where the nonzero components are all 1's (in what follows we refer to this setting as Dirac), or each of the nonzero components are generated from a uniform distribution over $[0, 2]$, independently. We consider linear model where each $Y_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}^* + \epsilon_{ij}$, and $\epsilon_{ij}$ follows a normal distribution with variance equals 1. Note that the noise $\epsilon_{ij}$'s are correlated with other $\epsilon_{ij'}$'s, as specified later. The cardinality $s$ of the active set is set as 5, 10 or 20; we let $d = 200, 500, 1000$, and $n = 50, 100$. In each simulation we generate the covariate $\boldsymbol{X}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$ for each $(i, j)$, where the $(k, \ell)$th element of $\boldsymbol{\Sigma}$ equals $\rho^{|k-\ell|}$, and $\rho = 0.25, 0.4, 0.6$ or $0.75$. We set $m = 3$ or 5, and we take the within-subject correlation to be either equal-correlation model or AR(1) model. In our method we include a broad class of matrices as the basis for the inverse of the correlation matrix. Specifically, we generate data from either the equal-correlation model or AR(1) model and include the basis matrices discussed in Section 2.4. For ease of presentation, we investigate Type I error, false discovery rate and power of our method.

TABLE 1

*Empirical Type I error rate (%) under equal correlation with correlation parameter being* 0.5. *The nominal level is set to be 5%*

| | | $\rho = 0.25$ | | $\rho = 0.4$ | | $\rho = 0.6$ | | $\rho = 0.75$ | |
|---|---|---|---|---|---|---|---|---|---|
| $(n, d)$ | $s$ | Dirac | U[0, 2] | Dirac | U[0, 2] | Dirac | U[0, 2] | Dirac | U[0, 2] |
| | | | | | $m = 3$ | | | | |
| (50, 200) | 5 | 5.9 | 5.5 | 5.5 | 5.3 | 5.8 | 4.7 | 4.4 | 4.6 |
| | 10 | 5.9 | 5.9 | 5.2 | 4.8 | 4.8 | 5.8 | 5.1 | 4.9 |
| | 20 | 6.1 | 5.7 | 5.9 | 4.8 | 6.2 | 6.0 | 5.4 | 5.5 |
| (100, 500) | 5 | 5.6 | 5.4 | 5.7 | 5.2 | 6.0 | 5.9 | 4.7 | 5.1 |
| | 10 | 5.8 | 5.2 | 5.5 | 5.3 | 5.8 | 5.3 | 3.9 | 5.4 |
| | 20 | 5.7 | 4.9 | 5.4 | 4.9 | 5.6 | 4.0 | 6.2 | 5.7 |
| (100, 1000) | 5 | 5.3 | 5.7 | 5.8 | 5.5 | 5.8 | 6.3 | 5.1 | 4.5 |
| | 10 | 5.9 | 5.4 | 5.3 | 5.5 | 5.3 | 5.1 | 4.0 | 4.3 |
| | 20 | 6.1 | 5.8 | 5.6 | 5.7 | 5.5 | 4.3 | 6.8 | 6.2 |
| | | | | | $m = 5$ | | | | |
| (50, 200) | 5 | 5.5 | 5.2 | 5.3 | 5.2 | 5.4 | 5.1 | 5.0 | 4.8 |
| | 10 | 5.3 | 5.2 | 5.3 | 5.1 | 4.9 | 5.5 | 5.3 | 5.0 |
| | 20 | 5.3 | 5.4 | 5.4 | 5.3 | 5.7 | 5.3 | 5.3 | 5.6 |
| (100, 500) | 5 | 5.3 | 5.3 | 5.5 | 5.3 | 5.4 | 5.6 | 4.9 | 5.2 |
| | 10 | 5.5 | 5.4 | 5.4 | 5.1 | 5.6 | 5.5 | 4.7 | 4.8 |
| | 20 | 5.5 | 5.3 | 5.3 | 5.0 | 4.8 | 4.4 | 5.8 | 5.9 |
| (100, 1000) | 5 | 5.2 | 5.5 | 5.5 | 5.4 | 5.5 | 5.8 | 5.3 | 5.4 |
| | 10 | 5.6 | 5.7 | 5.0 | 5.6 | 5.2 | 5.3 | 4.4 | 4.5 |
| | 20 | 5.9 | 5.6 | 5.3 | 5.5 | 5.4 | 4.8 | 5.7 | 6.0 |

We first consider the empirical Type I error. Specifically, we apply the proposed method to test the null hypothesis $H_0 : \beta_1^* = 0$, which we assume to be true in our setting. The tuning parameters $\lambda$ and $\lambda'$ are determined by the five-fold cross validation. The simulation is repeated 1000 times. We report the empirical Type I error at 5% significance level in Tables 1 and 2. It is clearly seen that the proposed test can control the empirical Type I errors at the desired nominal level. This implies the asymptotic distribution of our test statistic is reasonably accurate in finite sample.

Next, we consider the empirical false discovery rate by applying the methods described in Section 3. In particular, we simultaneously test all $d$ hypotheses that $H_{0j} : \beta_j^* = 0$ for all $j = 1, \ldots, d$. After getting $d$ different p-values, we apply the proposed method under the level $\alpha = 0.1$ or 0.2. Under the same data generating schemes for investigating empirical Type I error, we repeat the simulation 1000 times and report the averaged false discovery rate in Tables 3 and 4. We find that the empirical false discovery rates are well controlled under different settings. Furthermore, we plot the empirical false discovery rate against the nominal false discovery rate from zero to one in Figure 1 under several settings. Our approach well controls the false discovery rate for different desired levels. It is worth noting that, in the second subfigure, the empirical FDR deviates from the nominal one as the maximum possible false discovery rate is $(d - s)/d = 90\%$ when $(s, d) = (20, 200)$.

Finally, we investigate the empirical power of the proposed test and compare it with some other high-dimensional inference procedures—the debiased Lasso method (van de Geer et al. (2014), Zhang and Zhang (2014)) and the decorrelation method (Ning and Liu (2017)) by pretending all observations are independent. In particular, we test $H_0 : \beta_1 = 0$, under the

TABLE 2
*Empirical Type I error rate (%) under AR-correlation structure with correlation parameter being* 0.6. *The nominal level is set to be* 5%

| | | $\rho = 0.25$ | | $\rho = 0.4$ | | $\rho = 0.6$ | | $\rho = 0.75$ | |
|---|---|---|---|---|---|---|---|---|---|
| $(n, d)$ | $s$ | Dirac | U[0, 2] | Dirac | U[0, 2] | Dirac | U[0, 2] | Dirac | U[0, 2] |
| | | | | | $m = 3$ | | | | |
| (50, 200) | 5 | 5.4 | 5.3 | 5.4 | 5.5 | 5.2 | 5.0 | 5.3 | 5.1 |
| | 10 | 4.6 | 5.1 | 5.4 | 4.8 | 5.1 | 5.3 | 5.1 | 5.2 |
| | 20 | 5.8 | 5.7 | 5.2 | 5.5 | 5.6 | 5.2 | 4.7 | 5.4 |
| (100, 500) | 5 | 5.4 | 5.1 | 5.6 | 5.3 | 5.5 | 5.3 | 5.2 | 5.3 |
| | 10 | 5.5 | 5.6 | 5.7 | 5.4 | 5.8 | 5.6 | 5.3 | 5.5 |
| | 20 | 5.6 | 4.8 | 5.8 | 5.3 | 5.3 | 4.5 | 4.6 | 4.6 |
| (100, 1000) | 5 | 5.3 | 5.4 | 4.8 | 4.7 | 5.3 | 5.9 | 5.7 | 5.6 |
| | 10 | 5.4 | 5.6 | 5.3 | 5.4 | 5.5 | 5.3 | 4.6 | 4.4 |
| | 20 | 6.1 | 5.8 | 5.6 | 5.7 | 5.5 | 5.3 | 5.9 | 5.7 |
| | | | | | $m = 5$ | | | | |
| (50, 200) | 5 | 5.2 | 5.2 | 5.3 | 5.6 | 5.1 | 5.2 | 5.4 | 5.0 |
| | 10 | 5.1 | 5.3 | 5.2 | 5.1 | 5.3 | 5.2 | 5.2 | 5.3 |
| | 20 | 5.5 | 5.6 | 5.3 | 5.4 | 5.5 | 5.1 | 4.8 | 5.3 |
| (100, 500) | 5 | 5.3 | 5.3 | 5.4 | 5.5 | 5.4 | 5.2 | 5.1 | 5.2 |
| | 10 | 5.2 | 5.7 | 5.6 | 5.5 | 4.8 | 4.9 | 5.2 | 5.6 |
| | 20 | 5.5 | 5.8 | 5.5 | 5.4 | 5.6 | 5.2 | 4.9 | 4.7 |
| (100, 1000) | 5 | 5.5 | 5.7 | 5.4 | 5.6 | 5.5 | 5.6 | 5.4 | 5.7 |
| | 10 | 5.6 | 5.7 | 5.2 | 5.5 | 5.6 | 5.4 | 5.2 | 5.3 |
| | 20 | 5.8 | 6.1 | 5.8 | 5.7 | 6.0 | 5.9 | 4.5 | 4.3 |

Dirac setting, where the signal of $\beta_1$ gradually increase from 0 to 0.7, and we investigate the empirical rejection rate for different settings. The results are summarized in Figure 2. As expected, our QDIF approach achieves better empirical power, especially when the signal is weak and $s$ is relatively large. This is in line with our theoretical results.

4.2. *BMI dataset*. We further evaluate our method using a BMI genomic dataset from the Framingham Heart Study (FHS). This is a long-term, ongoing cardiovascular study beginning in 1948 under the direction of the National Heart, Lung and Blood institute (NHLBI) on residents of the City of Framingham, Massachusetts. The objective is to identify the important characteristics that contribute to cardiovascular disease. We refer to Jaquish (2007) for more details of the study. Recently, 913,854 SNPs from 24 chromosomes have been genotyped from the Offspring Cohort study. We investigate the issue of obesity as the body mass index (BMI), where $BMI = weight\ (\text{kg})/height\ (\text{m})^2$. Our dataset contains BMI of 977 samples, where each sample's BMI value is collected at 26 times. Since there are some missing values presented in the response of different samples, we first adopt $n = 234$ samples with time points $m_i$'s ranging from three to seven where their BMI values are recorded. The nonrare SNPs genotypes from the 23 chromosomes are also recorded. Taking the BMI values as response variables $Y$, we first screen the features by regressing the BMI values on each of the SNPs and only keep the SNPs with marginal P-value less than 0.05. This reduces the dimension to $d = 4294$. Then, for the $j$th SNP we treat this covariate as $\boldsymbol{Z}$ in Section 2 and the rest of the covariates as $\boldsymbol{U}$. We apply the proposed QDIF method to test whether the $j$th SNP is associated with BMI, where we use the same basis matrices as in the simulations. The

TABLE 3
*Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under equal-correlation structure with correlation parameter being 0.5*

| | | | $\rho = 0.25$ | | $\rho = 0.4$ | | $\rho = 0.6$ | | $\rho = 0.75$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n, d)$ | $s$ | $\alpha$: | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 |
| | | | | | | $m = 3$ | | | | |
| (50, 200) | 5 | | 9.3 | 19.1 | 9.6 | 19.6 | 8.9 | 18.9 | 10.6 | 20.8 |
| | 10 | | 8.8 | 19.3 | 9.2 | 18.9 | 9.3 | 20.7 | 10.4 | 21.0 |
| | 20 | | 8.7 | 18.7 | 8.8 | 18.8 | 9.4 | 19.3 | 9.4 | 20.9 |
| (100, 500) | 5 | | 9.6 | 19.2 | 10.3 | 20.2 | 11.0 | 20.9 | 10.9 | 21.3 |
| | 10 | | 9.7 | 20.1 | 9.5 | 21.1 | 8.7 | 20.8 | 8.8 | 20.8 |
| | 20 | | 9.4 | 18.9 | 9.2 | 18.9 | 10.5 | 20.3 | 11.1 | 21.3 |
| (100, 1000) | 5 | | 10.4 | 20.8 | 9.5 | 20.7 | 9.2 | 21.3 | 9.4 | 20.6 |
| | 10 | | 9.5 | 21.2 | 9.2 | 20.6 | 9.1 | 20.9 | 9.8 | 20.9 |
| | 20 | | 9.3 | 21.8 | 8.9 | 21.5 | 8.7 | 22.0 | 12.1 | 21.4 |
| | | | | | | $m = 5$ | | | | |
| (50, 200) | 5 | | 9.5 | 19.7 | 9.4 | 19.5 | 9.2 | 19.3 | 10.0 | 20.5 |
| | 10 | | 9.1 | 19.5 | 9.6 | 19.2 | 9.1 | 21.3 | 10.7 | 20.4 |
| | 20 | | 9.2 | 19.1 | 9.0 | 19.0 | 9.5 | 18.9 | 9.8 | 18.9 |
| (100, 500) | 5 | | 9.3 | 20.5 | 9.7 | 21.0 | 10.6 | 20.3 | 10.2 | 20.7 |
| | 10 | | 9.5 | 19.7 | 9.6 | 21.3 | 9.8 | 20.4 | 9.5 | 21.3 |
| | 20 | | 9.7 | 19.4 | 9.5 | 19.2 | 9.5 | 19.3 | 10.9 | 18.5 |
| (100, 1000) | 5 | | 9.8 | 20.4 | 10.8 | 20.4 | 8.9 | 19.3 | 9.6 | 19.5 |
| | 10 | | 9.2 | 20.7 | 9.3 | 20.4 | 9.3 | 20.5 | 9.5 | 20.7 |
| | 20 | | 9.1 | 21.0 | 10.5 | 19.2 | 9.2 | 21.7 | 11.3 | 20.6 |

obtained p-value is recorded as $P_j$ as in Section 3. We repeat this procedure for all the SNPs which yields a sequence of p-values $P_1, \ldots, P_d$. When we select important SNPs based on these p-values, we need to account for the fact that we have been looking at a large number of candidate SNPs (the so called multiple testing effect). Failure to account for the multiple testing effect causes irreproducibility of the results and may yield misleading scientific conclusions. Given the practical importance of this problem, we developed a rigorous result on the FDR control. Applying our result to the data analysis, we find that the 12,289th position of the 1st chromosome, 681st, 756th and 19,880th SNPs of the 10th chromosome, and 1189th and 12,075th SNPs of the 20th chromosome are the significant SNPs under the FDR at 10%. Interestingly, it is known that the 10th and 20th chromosomes are related to obesity (Dong et al. (2003)) which matches our results that the significant SNPs are mostly located at the 10th and 20th chromosomes.

## 5. Technical lemmas and proofs.

5.1. *Technical lemmas*. In this subsection we provide some technical lemmas used in our proofs in Sections 2 and 3. The proofs of these technical lemmas are given in the Supplementary Material, Fang, Ning and Li (2020).

The first lemma on the rate of convergence of random matrices in the spectral norm is derived from the matrix Bernstein's inequality and is fundamental for the rest of the proof.

TABLE 4

*Empirical false discovery rate (%) at level $\alpha = 0.1$ and 0.2 under AR-correlation structure with correlation parameter being 0.6*

| | | | $\rho = 0.25$ | | $\rho = 0.4$ | | $\rho = 0.6$ | | $\rho = 0.75$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n, d)$ | $s$ | $\alpha$: | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 |
| | | | | | | $m = 3$ | | | | |
| (50, 200) | 5 | | 9.6 | 20.3 | 9.8 | 20.4 | 10.1 | 20.5 | 9.7 | 19.4 |
| | 10 | | 9.4 | 20.3 | 9.5 | 20.8 | 10.4 | 19.5 | 9.8 | 20.9 |
| | 20 | | 10.8 | 19.5 | 10.6 | 21.2 | 10.0 | 19.3 | 9.6 | 18.8 |
| (100, 500) | 5 | | 10.3 | 20.6 | 10.5 | 20.5 | 9.5 | 20.1 | 10.3 | 19.8 |
| | 10 | | 10.4 | 19.5 | 9.6 | 20.3 | 9.5 | 20.8 | 10.4 | 21.0 |
| | 20 | | 8.9 | 20.8 | 9.3 | 19.2 | 8.8 | 21.0 | 8.7 | 20.9 |
| (100, 1000) | 5 | | 10.5 | 20.6 | 10.6 | 20.7 | 9.3 | 19.2 | 9.2 | 20.3 |
| | 10 | | 10.5 | 20.7 | 10.8 | 19.5 | 10.2 | 19.4 | 9.7 | 20.4 |
| | 20 | | 8.7 | 20.5 | 11.3 | 20.7 | 10.4 | 20.1 | 10.9 | 20.6 |
| | | | | | | $m = 5$ | | | | |
| (50, 200) | 5 | | 10.3 | 20.1 | 10.0 | 20.3 | 9.8 | 20.5 | 9.9 | 19.8 |
| | 10 | | 9.7 | 20.3 | 9.6 | 20.8 | 10.4 | 19.4 | 9.4 | 20.7 |
| | 20 | | 9.6 | 19.5 | 9.5 | 19.3 | 10.5 | 20.9 | 10.8 | 20.8 |
| (100, 500) | 5 | | 9.5 | 20.3 | 9.8 | 20.6 | 9.6 | 19.7 | 20.4 | 20.4 |
| | 10 | | 9.4 | 20.8 | 10.1 | 20.7 | 10.9 | 20.8 | 9.3 | 20.8 |
| | 20 | | 10.3 | 20.5 | 9.2 | 21.1 | 10.5 | 20.9 | 9.1 | 19.5 |
| (100, 1000) | 5 | | 10.7 | 20.8 | 10.5 | 19.5 | 9.8 | 20.0 | 9.6 | 19.5 |
| | 10 | | 10.6 | 20.5 | 10.8 | 20.9 | 9.2 | 20.8 | 8.9 | 21.1 |
| | 20 | | 11.0 | 21.2 | 9.3 | 19.6 | 9.5 | 20.3 | 8.8 | 18.7 |

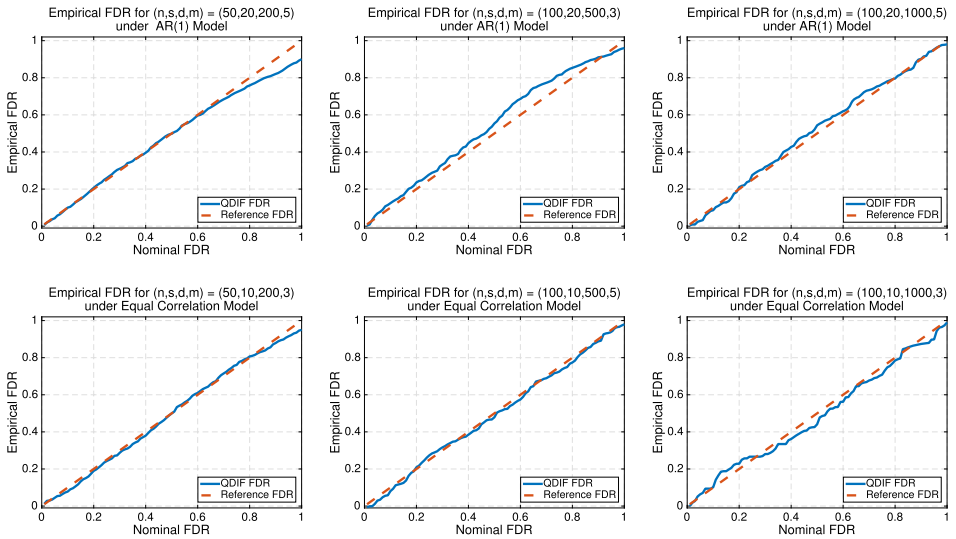


FIG. 1. *Empirical FDR of the proposed method in* AR(1) *and equal correlation models, where we take the correlation parameter as* 0.75.

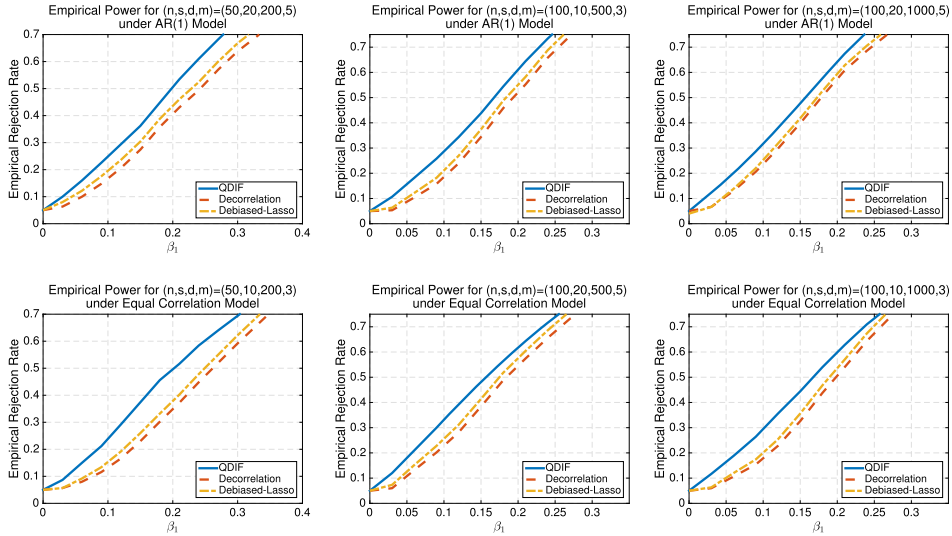


FIG. 2. *Empirical power for quadratic decorrelated inference (QDIF), debiased Lasso and decorrelation methods under* AR(1) *and equal correlation models, where we take the correlation parameter as* 0.75.

LEMMA 5.1. *Suppose that Assumptions 2.1–2.5 and* $\frac{d_0(\log n)^2 \log d_0}{n} = o(1)$ *hold. Then,*

$$
\begin{aligned}
&\max_{1\le k\le K} \big\|\nabla \bar{\mathbf{S}}_k^*(\boldsymbol{\theta}^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*)\big\|_2 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{d_0 \log d_0/n}\big),\\
&\max_{1\le k\le K} \left\| \frac{1}{n}\sum_{i=1}^n (\mathbf{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*)^T \boldsymbol{\Psi}_i (\mathbf{Z}_i - \boldsymbol{U}_i \mathbf{W}_k^*) - \mathbf{g}_{0k}(\boldsymbol{\theta}^*) \right\|_2 \\
&\quad = \mathcal{O}_{\mathbb{P}}\big(\sqrt{d_0 \log d_0/n}\big),\\
&\max_{1\le k\le K} \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i \boldsymbol{\Psi}_i \mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i \boldsymbol{\Psi}_i \mathbf{Z}_i) \right\|_2 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{d_0 \log d_0/n}\big),
\end{aligned}
\tag{5.1}
$$

*and*

$$
\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{S}_i^*(\boldsymbol{\theta}^*)\mathbf{S}_i^{*T}(\boldsymbol{\theta}^*) - \mathbf{C}^* \right\|_2 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{d_0 \log d_0 (\log n)^2/n}\big). \tag{5.2}
$$

LEMMA 5.2. *Recall that* $\widehat{\mathbf{g}}(\boldsymbol{\theta}) = \nabla \overline{\mathbf{S}}_n(\boldsymbol{\theta})$ *and* $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbb{E}\{\nabla \mathbf{S}_i^*(\boldsymbol{\theta})\}$. *Under the conditions in Theorem 2.7, we have*

$$
\begin{aligned}
\big\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}_0(\boldsymbol{\theta}^*)\big\|_2 &= \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{(s \vee s')d_0 \log d}{n}}\right),\\
\big\|\widehat{\mathbf{g}}(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)\big\|_2 &= \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{(s \vee s')d_0 \log d}{n}}\right).
\end{aligned}
$$

LEMMA 5.3. *Recall that* $\widehat{\mathbf{C}} = n^{-1}\sum_{i=1}^n \widehat{\mathbf{S}}_i(\widehat{\theta})\widehat{\mathbf{S}}_i^T(\widehat{\theta})$ *in (2.9), where* $\widehat{\mathbf{S}}_i(\theta) = \{\widehat{S}_{i1}(\theta), \ldots, \widehat{S}_{iK}(\theta)\}^T$ *and* $\mathbf{C}^* = \mathbb{E}\{\mathbf{S}_i^*(\theta^*)\mathbf{S}_i^*(\theta^*)^T\}$ *in (2.16), where* $\mathbf{S}_i^*(\theta^*) = (S_{i1}^*(\theta^*), \ldots, S_{iK}^*(\theta^*))^T$. *Under the conditions in Theorem 2.7, we have*

$$
\big\|\widehat{\mathbf{C}} - \mathbf{C}^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{d_0(s \vee s')\log d(\log n)^2}{n}}\right).
$$

LEMMA 5.4. *Under the same conditions as in Theorem 2.7, we have,*

$$\max_{1\le k\le K}\left\|\bar{\mathbf{S}}_{nk}(\boldsymbol{\theta}^*)-\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_{ik}^*(\boldsymbol{\theta}^*)\right\|_2=\mathcal{O}_{\mathbb{P}}\Big(\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big).$$

LEMMA 5.5. *Recall that* $Q_n^*(\boldsymbol{\theta})=\bar{\mathbf{S}}_n^*(\boldsymbol{\theta})^T\mathbf{C}^{*-1}\bar{\mathbf{S}}_n^*(\boldsymbol{\theta})$, *where* $\bar{\mathbf{S}}^*(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n\mathbf{S}_i^*(\boldsymbol{\theta})$. *Under the same conditions as in Theorem 2.7, we have*

$$\begin{aligned}&\|\nabla Q_n^*(\boldsymbol{\theta}^*)\|_2=\mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{d_0}{n}}\Big),\\&\|\nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)-\nabla Q_n^*(\boldsymbol{\theta}^*)\|_2\\&\quad=\mathcal{O}_{\mathbb{P}}\Big(\frac{d_0\{(s\vee s')\log d(\log n)^2\}^{1/2}}{n}+\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big).\end{aligned}$$

LEMMA 5.6. *Let $c$ be a small constant. Under the conditions in Theorem 2.7, uniformly over $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le cd_0^{-1/2}$ it holds that with probability tending to one*

$$\widetilde{Q}_n(\boldsymbol{\theta})-\widetilde{Q}_n(\boldsymbol{\theta}^*)-\nabla\widetilde{Q}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\ge C\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2$$

*for some positive constant $C$.*

LEMMA 5.7. *Suppose Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 hold for all $\beta_j^*$ for $j\in[d]$, and suppose that $\lambda\asymp\lambda'\asymp\sqrt{n^{-1}\log d}$ and $n^{-1/2}s(\log d)(\log n)=o(1/\log d)$. Let the "ideal" version of $\widetilde{\beta}_j$ be*

$$\frac{1}{n}\sum_{i=1}^{n}A_{ij}=\frac{1}{n}\sum_{i=1}^{n}\{\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1}\mathbf{g}_{0j}(\beta_j^*)\}^{-1}\mathbf{g}_{0j}(\beta_j^*)\mathbf{C}_j^{*-1}\cdot\mathbf{S}_{ij}^*(\beta_j^*),\tag{5.3}$$

*where $\mathbf{g}_{0j}^*$, $\mathbf{C}_j^*$ and $\mathbf{S}_{ij}^*(\beta_j^*)$ denote corresponding $\mathbf{g}_0$, $\mathbf{C}^*$ and $\mathbf{S}_i^*(\beta_j^*)$ in the previous section for $\beta_j^*$, and let the ideal version of $\widehat{T}_j$ be*

$$T_j=\sqrt{n}\sigma_j^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}A_{ij}\Big),$$

*where $\sigma_j$ is defined in (2.21). We have that $\sqrt{n}\widetilde{\beta}_j$ converges to $\frac{1}{\sqrt{n}}\sum_{i=1}^n A_{ij}$ such that as $n\to\infty$,*

$$\max_{j\in\mathcal{H}_0}\Big|\sqrt{n}\widetilde{\beta}_j-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_{ij}\Big|=\mathcal{O}_{\mathbb{P}}\big(n^{-1/2}s(\log d)(\log n)\big),$$

*and*

$$\max_j|\widehat{T}_j-T_j|=\mathcal{O}_{\mathbb{P}}\big(n^{-1/2}s(\log d)(\log n)\big).$$

LEMMA 5.8. *Suppose the conditions in Theorem 3.1 hold. Let $r_d$ be any sequence such that $r_d\to\infty$ as $d\to\infty$, and $r_d=o(|\mathcal{S}_1|)$. Then, we have*

$$\sup_{r_d/d\le u\le 1}\Big|\frac{\sum_{j\in\mathcal{S}_0}\mathbb{1}\{P_j\le u\}}{u\cdot|\mathcal{S}_0|}-1\Big|\to 0,$$

*in probability.*

5.2. *Proof of main results.*

PROOF OF THEOREM 2.7. Since $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_2 \lesssim \sqrt{s \log d/n} \le c d_0^{-1/2}$ under condition (2.18), we claim that $\boldsymbol{\theta}^*$ lies in $\Theta_n$ with probability tending to one. By the definition of $\widetilde{\boldsymbol{\theta}}$, it holds that $\widetilde{Q}_n(\widetilde{\boldsymbol{\theta}}) \le \widetilde{Q}_n(\boldsymbol{\theta}^*)$ which further implies

$$\widetilde{Q}_n(\widetilde{\boldsymbol{\theta}}) - \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla \widetilde{Q}_n(\boldsymbol{\theta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le -\nabla \widetilde{Q}_n(\boldsymbol{\theta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

By Lemma 5.6, the left hand side is lower bounded by $C\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$. Thus, by Cauchy–Schwarz inequality,

$$\begin{aligned} C\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 &\le \|\nabla \widetilde{Q}_n(\boldsymbol{\theta}^*)\|_2 \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \\ &\le (\|\nabla \widetilde{Q}_n(\boldsymbol{\theta}^*) - \nabla Q_n^*(\boldsymbol{\theta}^*)\|_2 + \|\nabla Q_n^*(\boldsymbol{\theta}^*)\|_2) \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2. \end{aligned}$$

Together with Lemma 5.5, we have $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{d_0/n})$ which completes the proof. □

PROOF OF THEOREM 2.8. We first note that

$$\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_2 \le \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{d_0}{n}} + \sqrt{\frac{s \log d}{n}} = o_{\mathbb{P}}\Big(\frac{1}{\sqrt{d_0}}\Big).$$

This implies $\widetilde{\boldsymbol{\theta}}$ belongs to the interior of $\Theta_n$. By the first order optimality condition, $\widetilde{\boldsymbol{\theta}}$ satisfies

$$\widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \overline{\mathbf{S}}_n(\widetilde{\boldsymbol{\theta}}) = 0 \quad \text{where } \widehat{\mathbf{g}}(\boldsymbol{\theta}) = \partial \overline{\mathbf{S}}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}.$$

By the mean-value theorem for vector valued functions, for each component of $\overline{\mathbf{S}}_n(\widetilde{\boldsymbol{\theta}})$, say $(\overline{\mathbf{S}}_n(\widetilde{\boldsymbol{\theta}}))_j$ there exists $\bar{\boldsymbol{\theta}}_j = \zeta_j \boldsymbol{\theta}^* + (1 - \zeta_j)\widetilde{\boldsymbol{\theta}}$ for some $\zeta_j \in [0, 1]$ such that $(\overline{\mathbf{S}}_n(\widetilde{\boldsymbol{\theta}}) - \overline{\mathbf{S}}_n(\boldsymbol{\theta}^*))_j = [\partial(\overline{\mathbf{S}}_n(\bar{\boldsymbol{\theta}}_j))_j/\partial \boldsymbol{\theta}]^T (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$. For notational simplicity we suppress the subscript $j$ in $\bar{\boldsymbol{\theta}}_j$ and write it as

$$\widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \{\overline{\mathbf{S}}_n(\boldsymbol{\theta}^*) + \widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})^T (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} = 0.$$

Thus, we have

$$\{\widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})\}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \overline{\mathbf{S}}_n(\boldsymbol{\theta}^*). \tag{5.4}$$

Define

$$\begin{aligned} \boldsymbol{T}_1 &= [\{\widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \widehat{\mathbf{g}}(\bar{\boldsymbol{\theta}})\}^{-1} - \{\mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1} \mathbf{g}_0(\boldsymbol{\theta}^*)\}^{-1}] \cdot \widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \overline{\mathbf{S}}_n(\boldsymbol{\theta}^*), \\ \boldsymbol{T}_2 &= \{\mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1} \mathbf{g}_0(\boldsymbol{\theta}^*)\}^{-1} \Big\{ \widehat{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})^T \widehat{\mathbf{C}}^{-1} \overline{\mathbf{S}}_n(\boldsymbol{\theta}^*) - \mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1} \cdot \frac{1}{n} \sum_{i=1}^n \mathbf{S}_i^*(\boldsymbol{\theta}^*) \Big\}, \\ \bar{\boldsymbol{\xi}} &= \frac{1}{n} \sum_{i=1}^n \boldsymbol{\xi}_i \quad \text{where } \boldsymbol{\xi}_i = -\{\mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1} \mathbf{g}_0(\boldsymbol{\theta}^*)\}^{-1} \mathbf{g}_0(\boldsymbol{\theta}^*)^T \mathbf{C}^{*-1} \mathbf{S}_i^*(\boldsymbol{\theta}^*). \end{aligned}$$

Then, it holds that $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = -\boldsymbol{T}_1 - \boldsymbol{T}_2 + \bar{\boldsymbol{\xi}}$. Putting together Lemmas 5.2, 5.3 and 5.4, we can show that

$$\|\boldsymbol{T}_1\|_2 = \mathcal{O}_{\mathbb{P}}\Big( \frac{d_0^{3/2}}{n} + \frac{d_0 \{(s \vee s') \log d (\log n)^2\}^{1/2}}{n} \Big)$$

with some tedious algebraic manipulation similar to the proof of Lemma 5.6. In addition, we can show that

$$\|\boldsymbol{T}_2\|_2 = \mathcal{O}_{\mathbb{P}}\Big( \frac{d_0^{3/2}}{n} + \frac{d_0 \{(s \vee s') \log d (\log n)^2\}^{1/2}}{n} + \frac{d_0^{1/2} (s \vee s') \log d \log n}{n} \Big).$$

Combining the above results, we have

$$\begin{aligned}&\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*-\bar{\boldsymbol{\xi}}\|_2\\&\quad=\mathcal{O}_{\mathbb{P}}\Big(\frac{d_0^{3/2}}{n}+\frac{d_0\{(s\vee s')\log d(\log n)^2\}^{1/2}}{n}+\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big).\end{aligned}$$

To show the limiting distribution of $n(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})$, we first note that

$$\begin{aligned}&d_0^{-1/2}n\big|(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)-\bar{\boldsymbol{\xi}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\bar{\boldsymbol{\xi}}\big|\\&\quad\le d_0^{-1/2}n\big[\big|(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*-\bar{\boldsymbol{\xi}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\big|+\big|\bar{\boldsymbol{\xi}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*-\bar{\boldsymbol{\xi}})\big|\big]\\&\quad\le d_0^{-1/2}n\big[\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*-\bar{\boldsymbol{\xi}}\|_2\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_2\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2+\|\bar{\boldsymbol{\xi}}\|_2\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_2\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*-\bar{\boldsymbol{\xi}}\|_2\big]\\&\quad\lesssim\frac{n}{d_0^{1/2}}\Big[\frac{d_0^{3/2}}{n}+\frac{d_0\{(s\vee s')\log d(\log n)^2\}^{1/2}}{n}+\frac{d_0^{1/2}(s\vee s')\log d\log n}{n}\Big]\frac{d_0^{1/2}}{n^{1/2}}\\&\quad=o_{\mathbb{P}}(1)\end{aligned}$$

by Theorem 2.7 and the assumed conditions. Thus, it suffices to show the limiting distribution of $\bar{\boldsymbol{\xi}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\bar{\boldsymbol{\xi}}$. Note that $\mathbb{E}\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\xi}_i\|_2^3\lesssim\mathbb{E}\|\mathbf{S}_i^*(\boldsymbol{\theta}^*)\|_2^3\lesssim d_0^{3/2}$. Theorem 1.1 in Bentkus (2003) implies

$$\sup_{B\in\mathcal{B}}\Big|\mathbb{P}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^n\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\xi}_i\in B\Big)-\mathbb{P}(\boldsymbol{N}\in B)\Big|\le\frac{Cd_0^{3/2}}{n^{1/2}},$$

where $\mathcal{B}$ is the set of all Euclidean balls in $\mathbb{R}^{d_0}$, $C$ is a positive constant and $\boldsymbol{N}=(N_1,\dots,N_{d_0})$ are $d_0$ independent $N(0,1)$. Finally, we obtain for any $t\in\mathbb{R}$,

$$\begin{aligned}&\mathbb{P}\Big(\frac{n\bar{\boldsymbol{\xi}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\bar{\boldsymbol{\xi}}-d_0}{\sqrt{2d_0}}\le t\Big)\\&\quad\le\mathbb{P}\Big(\frac{\|\boldsymbol{N}\|_2^2-d_0}{\sqrt{2d_0}}\le t\Big)+\frac{Cd_0^{3/2}}{n^{1/2}}\\&\quad=\mathbb{P}\Big(\frac{1}{\sqrt{d_0}}\sum_{i=1}^{d_0}\frac{N_i^2-1}{\sqrt{2}}\le t\Big)+\frac{Cd_0^{3/2}}{n^{1/2}}\le\Phi(t)+\frac{C'}{d_0^{1/2}}+\frac{Cd_0^{3/2}}{n^{1/2}},\end{aligned}$$

where $\Phi(\cdot)$ is the c.d.f. of a standard normal distribution and $C'$ is a absolute constant from the standard Berry–Esseen bound. The same probability can be lower bounded by using the same argument. Thus, as $d_0,n\to\infty$, $(n\bar{\boldsymbol{\xi}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\bar{\boldsymbol{\xi}}-d_0)/\sqrt{2d_0}$ converges weakly to $N(0,1)$. When $d_0$ is fixed, the Layponov condition for $\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}^T\boldsymbol{\xi}_i$ holds for any $\boldsymbol{v}\in\mathbb{R}^{d_0}$ and thus $\frac{1}{n^{1/2}}\sum_{i=1}^n\boldsymbol{v}^T\boldsymbol{\xi}_i\rightsquigarrow N(0,\boldsymbol{v}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{v})$. Finally, we obtain (2.22) by applying the Cramer–Wald device. □

PROOF OF THEOREM 3.1. For notational simplicity we suppress the dependence of $\widehat{u}_{\alpha,t}$ on $t$. By the definition of $\mathrm{FDP}(u)$ and $\widehat{\mathrm{FDP}}_t(u)$, we have

$$\mathrm{FDP}(\widehat{u}_\alpha)=\widehat{\mathrm{FDP}}_\lambda(\widehat{u}_\alpha)\cdot\frac{|\mathcal{S}_0|}{\pi(t)d}\cdot\frac{\sum_{j\in\mathcal{S}_0}\mathbb{1}\{P_j\le\widehat{u}_\alpha\}}{\widehat{u}_\alpha\cdot|\mathcal{S}_0|}.\tag{5.5}$$

We first show that $(u\cdot|\mathcal{S}_0|)^{-1}\sum_{j\in\mathcal{S}_0}\mathbb{1}\{P_j\le u\}\to1$ in probability. We prove a moderate deviation result for this quantity. By assumption, we have $|\mathcal{S}_1|\to\infty$. Let $r_d$ be a sequence

such that $r_d \to \infty$ as $d \to \infty$, and $r_d = o(|\mathcal{S}_1|)$. We first prove that $\mathbb{P}(\widehat{u}_\alpha \le r_d/d) \to 0$ as $n \to \infty$. Note that

$$1 - \chi_1^2(2\log d) = 2\big(1 - \Phi(\sqrt{2\log d})\big) \le \sqrt{2/\pi}\,(d\sqrt{2\log d})^{-1} \lesssim r_d/d.$$

Hence, for any $j \in \mathcal{S}_1$, we have

$$\mathbb{P}(P_j \le r_d/d) = \mathbb{P}\big(\Lambda_{nj} \ge 1 - \chi_1^{-2}(r_d/d)\big) \ge \mathbb{P}(\Lambda_{nj} \ge 2\log d). \tag{5.6}$$

Recall that $V_{nj} := \sqrt{n}(\widetilde{\beta}_j - \beta_j^*)/\widehat{\sigma}$. By extending the proof of Theorem 2.8, we get

$$\lim_{n\to\infty} \max_{j\in\{1,\dots,d\}} \sup_{x\in\mathbb{R}} \big|\mathbb{P}(V_{nj} \le x) - \Phi(x)\big| = 0. \tag{5.7}$$

For any $j \in \mathcal{S}_1$, we have

$$\Lambda_{nj} = n\widehat{\beta}_j^2/\widehat{\sigma}^2 = \{V_{nj} + \sqrt{n}\beta_j^*/\widehat{\sigma}\}^2.$$

Therefore, we have that

$$\begin{aligned}
\mathbb{P}(\Lambda_{nj} \le 2\log d) &\le \mathbb{P}\big(|V_{nj} + \sqrt{n}\beta_j^*/\widehat{\sigma}| \le \sqrt{2\log d}\big) \\
&\le \mathbb{P}\big(-|V_{nj}| + \sqrt{n}|\beta_j^*|/\widehat{\sigma} \le \sqrt{2\log d}\big) \\
&\le \mathbb{P}\big(CC'\sqrt{\log d} - (\log d)^{1/4} \le \sqrt{2\log d}\big) \\
&\quad + \mathbb{P}\big(|V_{nj}| \ge (\log d)^{1/4}\big),
\end{aligned}$$

where in the last inequality we used the condition that $|\beta_j^*| \ge C\sqrt{(\log d)/n}$ for $j \in \mathcal{S}_1$ and $1/\widehat{\sigma} \ge C'$. If $CC' > \sqrt{3}$, hence $CC'\sqrt{\log d} - (\log d)^{1/4} > \sqrt{2\log d}$ for large enough $d$. Moreover, by (5.7)

$$\begin{aligned}
&\max_{j\in\mathcal{S}_1} \mathbb{P}\big(|V_{nj}| \ge (\log d)^{1/4}\big) \\
&\quad \le \max_{j\in\mathcal{S}_1} \big|\mathbb{P}\big(|V_{nj}| \ge (\log d)^{1/4}\big) - 2\{1 - \Phi\big((\log d)^{1/4}\big)\}\big| + 2\{1 - \Phi\big((\log d)^{1/4}\big)\} \\
&\quad \to 0.
\end{aligned}$$

Hence, we get $\max_{j\in\mathcal{S}_1} \mathbb{P}(\Lambda_{nj} \le 2\log d) \to 0$. Therefore, by (5.6) we conclude that $\mathbb{P}(P_j \le r_d/d) \to 1$, uniformly, over $j \in \mathcal{S}_1$. Therefore, we have

$$\frac{1}{|\mathcal{S}_1|} \sum_{j\in\mathcal{S}_1} \mathbb{P}\bigg(P_j \le \frac{r_d}{d}\bigg) \to 1,$$

which implies that $1/|\mathcal{S}_1| \sum_{j\in\mathcal{S}_1} \mathbb{1}\{P_j \le r_d/d\} \to 1$ in $L_1$ and in probability. Hence, we have

$$\widehat{\mathrm{FDP}}_t(r_d/d) = \frac{\pi(t) \cdot r_d}{\sum_{j=1}^d \mathbb{1}\{P_j \le r_d/d\}} \le \pi(t) \cdot \frac{|\mathcal{S}_1|}{\sum_{j\in\mathcal{S}_1} \mathbb{1}\{P_j \le r_d/d\}} \cdot \frac{r_d}{|\mathcal{S}_1|}.$$

As $r_d = o(|\mathcal{S}_1|)$, we conclude that $\widehat{\mathrm{FDP}}_t(r_d/d) \to 0 < \alpha$ in probability, and hence by the definition of $\widehat{u}_\alpha$, we obtain

$$\mathbb{P}(\widehat{u}_\alpha \ge r_d/d) \to 1. \tag{5.8}$$

Hence, by (5.8) and Lemma 5.8 we conclude that $(\widehat{u}_\alpha \cdot |\mathcal{S}_0|)^{-1} \sum_{j\in\mathcal{S}_0} \mathbb{1}\{P_j \le \widehat{u}_\alpha\} \to 1$ in probability. Finally, by the definition of $\pi(t)$, we have

$$\frac{|\mathcal{S}_0|}{\pi(t)d} = \frac{|\mathcal{S}_0|}{\min(\sum_{j=1}^d \mathbb{1}\{P_j \ge t\}/(1-t), d)}.$$

If $\sum_{j=1}^{d} \mathbb{1}\{P_j \geq t\}/(1-t) \leq d$, we have

$$\frac{|\mathcal{S}_0|}{\pi(t)d} = \frac{|\mathcal{S}_0|(1-t)}{d - \sum_{j=1}^{d} \mathbb{1}\{P_j \leq t\}} = \frac{|\mathcal{S}_0|(1-t)}{d - (\sum_{j\in\mathcal{S}_0} + \sum_{j\notin\mathcal{S}_0})\mathbb{1}\{P_j \leq t\}}.$$

We have $|\mathcal{S}_0|/d \to 1$ and $(1/|\mathcal{S}_0|)\sum_{j\notin\mathcal{S}_0} \mathbb{1}\{P_j \leq t\} \leq (d - |\mathcal{S}_0|)/|\mathcal{S}_0| \to 0$ in probability. Moreover, given any $t \in [c, 1)$, for $d$ large enough, we have $t \geq r_d/d$. Hence, by Lemma (5.8) we have $(1/|\mathcal{S}_0|)\sum_{j\in\mathcal{S}_0} \mathbb{1}\{P_j \leq t\} \to t$ in probability. Hence, we conclude that $|\mathcal{S}_0|/(\pi(t) \cdot d) \to 1$ in probability. On the other hand, if $\sum_{j=1}^{d} \mathbb{1}\{P_j \geq t\}/(1-t) > d$, we have $|\mathcal{S}_0|/(\pi(t) \cdot d) = |\mathcal{S}_0|/d \to 1$.

Putting together the above results and by (5.5), we get $\mathrm{FDP}(\widehat{u}_\alpha) \leq \alpha$ and similarly $\mathrm{FDR}(\widehat{u}_\alpha) \leq \alpha$ in probability, hence concluding the proof. □

**Acknowledgments.** The authors are grateful to the Associate Editor and reviewers for their constructive comments, which led to a significant improvement of the earlier version of this paper.

Li is the corresponding author, and he also received partial travel support from NNSFC Grants 11690014 and 11690015 during his visit at Chinese Academy of Sciences and Nankai University.

The first author was supported by NIH Grant P50 DA039838 and NSF Grant DMS-1820702.

The second author was supported by NSF Grant DMS-1854637.

The third author was supported by NSF Grant DMS-1820702, NIH Grants P50 DA039838 and T32 LM012415.

## SUPPLEMENTARY MATERIAL

**Supplement to "Test of significance for high-dimensional longitudinal data"** (DOI: 10.1214/19-AOS1900SUPP; .pdf). The supplement consists of several technical lemmas along with their proofs, additional numerical results and numerical comparisons.

## REFERENCES

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BENTKUS, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** 385–402. MR1965117 https://doi.org/10.1016/S0378-3758(02)00094-0

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

DONG, C., WANG, S., LI, W.-D., LI, D., ZHAO, H. and PRICE, R. A. (2003). Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. *Am. J. Hum. Genet.* **72** 115–124.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385 https://doi.org/10.1214/17-AOS1568

FANG, E. X., NING, Y. and LI, R. (2020). Supplement to "Test of significance for high-dimensional longitudinal data." https://doi.org/10.1214/19-AOS1900SUPP.

FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 https://doi.org/10.1111/rssb.12224

GRAZIER G'SELL, M., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 423–444. MR3454203 https://doi.org/10.1111/rssb.12122

JAQUISH, C. E. (2007). The framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.* **8** 63. https://doi.org/10.1186/1471-2350-8-63

JAVANMARD, A. and MONTANARI, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems* 1187–1195.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453 https://doi.org/10.1214/13-AOS1169

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems* 476–484.

MA, S., SONG, Q. and WANG, L. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli* **19** 252–274. MR3019494 https://doi.org/10.3150/11-BEJ386

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836. MR1813977 https://doi.org/10.1093/biomet/87.4.823

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE GEER, S. and MÜLLER, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statist. Sci.* **27** 469–480. MR3025129 https://doi.org/10.1214/12-STS397

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

WANG, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.* **39** 389–417. MR2797851 https://doi.org/10.1214/10-AOS846

WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873 https://doi.org/10.1214/13-AOS1159

WANG, L. and QU, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 177–190. MR2655529 https://doi.org/10.1111/j.1467-9868.2008.00679.x

WANG, L., ZHOU, J. and QU, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68** 353–360. MR2959601 https://doi.org/10.1111/j.1541-0420.2011.01678.x

WANG, L., XUE, L., QU, A. and LIANG, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *Ann. Statist.* **42** 592–624. MR3210980 https://doi.org/10.1214/13-AOS1194

XUE, L., QU, A. and ZHOU, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *J. Amer. Statist. Assoc.* **105** 1518–1530. Supplementary materials available online. MR2796568 https://doi.org/10.1198/jasa.2010.tm10128

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHAO, T., LIU, H. and ZHANG, T. (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *Ann. Statist.* **46** 180–218. MR3766950 https://doi.org/10.1214/17-AOS1547