Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models

Rong Ma[1], T. Tony Cai[2] and Hongzhe Li[1]
Department of Biostatistics, Epidemiology and Informatics[1]
Department of Statistics[2]
University of Pennsylvania
Philadelphia, PA 19104

Abstract

High-dimensional logistic regression is widely used in analyzing data with binary outcomes. In this paper, global testing and large-scale multiple testing for the regression coefficients are considered in both single- and two-regression settings. A test statistic for testing the global null hypothesis is constructed using a generalized low-dimensional projection for bias correction and its asymptotic null distribution is derived. A lower bound for the global testing is established, which shows that the proposed test is asymptotically minimax optimal over some sparsity range. For testing the individual coefficients simultaneously, multiple testing procedures are proposed and shown to control the false discovery rate (FDR) and falsely discovered variables (FDV) asymptotically. Simulation studies are carried out to examine the numerical performance of the proposed tests and their superiority over existing methods. The testing procedures are also illustrated by analyzing a data set of a metabolomics study that investigates the association between fecal metabolites and pediatric Crohn's disease and the effects of treatment on such associations.

KEY WORDS: False discovery rate; Global testing; Large-scale multiple testing; Minimax lower bound.

1 INTRODUCTION

Logistic regression models have been applied widely in genetics, finance, and business analytics. In many modern applications, the number of covariates of interest usually grows with, and sometimes far exceeds, the number of observed samples. In such high-dimensional settings, statistical problems such as estimation, hypothesis testing, and construction of confidence intervals become much more challenging than those in the classical low-dimensional settings. The increasing technical difficulties usually emerge from the non-asymptotic analysis of both statistical models and the corresponding computational algorithms.

In this paper, we consider testing for high-dimensional logistic regression model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i^\top \beta, \qquad \text{for } i = 1, ..., n. \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. The observations are i.i.d. samples $Z_i = (y_i, X_i)$ for $i = 1, .., n$, and we assume $y_i|X_i \sim \text{Bernoulli}(\pi_i)$ independently for each $i = 1, ..., n$.

1.1 Global and Simultaneous Hypothesis Testing

It is important in high-dimensional logistic regression to determine 1) whether there are any associations between the covariates and the outcome and, if yes, 2) which covariates are associated with the outcome. The first question can be formulated as testing the global null hypothesis $H_0 : \beta = 0$; and the second question can be considered as simultaneously testing the null hypotheses $H_{0,i} : \beta_i = 0$ for $i = 1, ..., p$. Besides such single logistic regression problems, hypothesis testing involving two logistic regression models with regression coefficients $\beta^{(1)}$ and $\beta^{(2)}$ in $\mathbb{R}^p$ is also important. Specifically, one is interested in testing the global null hypothesis $H_0 : \beta^{(1)} = \beta^{(2)}$, or identifying the differentially associated covariates through simultaneously testing the null hypotheses $H_{0,i} : \beta_i^{(1)} = \beta_i^{(2)}$ for each $i = 1, ..., p$.

Estimation for high-dimensional logistic regression has been studied extensively. van de Geer (2008) considered high-dimensional generalized linear models (GLMs) with Lipschitz loss functions, and proved a non-asymptotic oracle inequality for the empirical risk minimizer with the Lasso penalty. Meier et al. (2008) studied the group Lasso for logistic regression and proposed an efficient algorithm that leads to statistically consistent estimates. Negahban et al. (2010) obtained the rate of convergence for the ℓ_1-regularized maximum likelihood estimator under GLMs using restricted strong convexity property. Bach (2010) extended tools from the convex optimization literature, namely self-concordant functions, to provide interesting extensions of theoretical results for the square loss to the logistic loss. Plan and Vershynin (2013) connected sparse logistic regression to one-bit compressed sensing and developed a unified theory for signal estimation with noisy observations.

In contrast, hypothesis testing and confidence intervals for high-dimensional logistic regression have only been recently addressed. van de Geer et al. (2014) considered constructing confidence intervals and statistical tests for single or low-dimensional components of the regression coefficients in high-dimensional GLMs. Mukherjee et al. (2015) studied the detection boundary for minimax hypothesis testing in high-dimensional sparse binary regression models when the design matrix is sparse. Belloni et al. (2016) considered estimating and constructing the confidence regions for a regression coefficient of primary interest in GLMs. More recently, Sur et al. (2017) and Sur and

Candès (2019) considered the likelihood ratio test for high-dimensional logistic regression under the setting that $p/n \to \kappa$ for some constant $\kappa < 1/2$, and showed that the asymptotic null distribution of the log-likelihood ratio statistic is a rescaled χ^2 distribution. Cai et al. (2017) proposed a global test and a multiple testing procedure for differential networks against sparse alternatives under the Markov random field model. Nevertheless, the problems of global testing and large-scale simultaneous testing for high-dimensional logistic regression models with $p \gtrsim n$ remain unsolved.

In this paper, we first consider global and multiple testing for a single high-dimensional logistic regression model. The global test statistic is constructed as the maximum of squared standardized statistics for individual coefficients, which are based on a two-step standardization procedure. The first step is to correct the bias of the logistic Lasso estimator using a generalized low-dimensional projection (LDP) method, and the second step is to normalize the resulting nearly unbiased estimators by their estimated standard errors. We show that the asymptotic null distribution of the test statistic is a Gumbel distribution and that the resulting test is minimax optimal under the Gaussian design by establishing the minimax separation distance between the null space and alternative space. For large-scale multiple testing, data-driven testing procedures are proposed and shown to control the false discovery rate (FDR) and falsely discovered variables (FDV) asymptotically. The framework for testing for single logistic regression is then extended to the setting of testing two logistic regression models.

The main contributions of the present paper are threefold.

1. We propose novel procedures for both the global testing and large-scale simultaneous testing for high dimensional logistic regressions. The dimension p is allowed to be much larger than the sample size n. Specifically, we require $\log p = O(n^{c_1})$ for the global test and $p = O(n^{c_2})$ for the multiple testing procedure, with some constant $c_1, c_2 > 0$. For the global alternatives characterized by the ℓ_∞ norm of the regression coefficients, the global test is shown to be minimax rate optimal with the optimal separation distance of order $\sqrt{\log p/n}$.

2. Following similar ideas in Ren et al. (2016) and Cai et al. (2017), our construction of the test statistics depends on a generalized version of the LDP method for bias correction. The original LDP method (Zhang and Zhang, 2014) relies on the linearity between the covariates and outcome variable. For logistic regression, the generalized approach first finds a linearization of the regression function, and the weighted LDP is then applied. Besides its usefulness in logistic regression, the generalized LDP method is flexible and can be applied to other nonlinear regression problems (see Section 7 for a detailed discussion).

3. The minimax lower bound is obtained for the global hypothesis testing under the Gaussian design. The lower bound depends on the calculation of the χ^2-divergence between two logistic regression models. To the best of our knowledge, this is the first lower bound result for high-dimensional logistic regression under the Gaussian design.

1.2 Other Related Work

We should note that a different but related problem, namely inference for high-dimensional linear regression, has been well studied in the literature. Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014a,b) considered confidence intervals and testing for low-dimensional parameters of the high-dimensional linear regression model and developed methods based on a two-stage debiased estimator that corrects the bias introduced at the first stage due to regularization. Cai and Guo (2017) studied minimaxity and adaptivity of confidence intervals for general linear functionals of the regression vector.

The problems of global testing and large-scale simultaneous testing for high-dimensional linear regression have been studied by Liu and Luo (2014), Ingster et al. (2010) and more recently by Xia et al. (2018) and Javanmard and Javadi (2019). However, due to the nonlinearity and the binary outcome, the approaches used in these works cannot be directly applied to logistic regression problems. In the Markov random field setting, Ren et al. (2016) and Cai et al. (2017) constructed pivotal/test statistics based on the debiased LDP estimators for node-wise logistic regressions with binary covariates. However, the results for sparse high-dimensional logistic regression models with general continuous covariates remain unknown.

Other related problems include joint testing and false discovery rate control for high-dimensional multivariate regression (Xia et al., 2018) and testing for high-dimensional precision matrices and Gaussian graphical models (Liu, 2013; Xia et al., 2015), where the inverse regression approach and de-biasing were carried out in the construction of the test statistics. Such statistics were then used for testing the global null with extreme value type asymptotic null distributions or to perform multiple testing that controls the false discovery rate.

1.3 Organization of the Paper and Notations

The rest of the paper is organized as follows. In Section 2, we propose the global test and establish its optimality. Some comparisons with existing works are made in detail. In Section 3, we present the multiple testing procedures and show that they control the FDR/FDP or FDV/FWER asymptotically. The framework is extended to the two-sample setting in Section 4. In Section 5, the numerical performance of the proposed tests are evaluated through extensive simulations. In Section 6, the methods are illustrated by an analysis of a metabolomics study. Further extensions and related problems are discussed in Section 7. In Section 8, some of the main theorems are proved. The proofs of other theorems as well as technical lemmas, and some further discussions are collected in the online Supplementary Materials.

Throughout our paper, for a vector $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$, we define the ℓ_p norm $\|\mathbf{a}\|_p = \big(\sum_{i=1}^n a_i^p\big)^{1/p}$, and the ℓ_∞ norm $\|\mathbf{a}\|_\infty = \max_{1\le j\le n} |a_i|$. $\mathbf{a}_{-j} \in \mathbb{R}^{n-1}$ stands for the subvector of $\mathbf{a}$ without the j the component. We denote $\text{diag}(a_1, ..., a_n)$ as the $n \times n$ diagonal matrix whose diagonal entries are $a_1, ..., a_n$. For a matrix $A \in \mathbb{R}^{p\times q}$, $\lambda_i(A)$ stands for the i-th largest singular value of A and $\lambda_{\max}(A)=\lambda_1(A)$, $\lambda_{\min}(A)=\lambda_{p\wedge q}(A)$. For a smooth function $f(x)$ defined on $\mathbb{R}$, we denote $\dot{f}(x)=df(x)/dx$ and $\ddot{f}(x)=d^2f(x)/dx^2$. Furthermore, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n=o(b_n)$ if $\lim_n a_n/b_n=0$, and write $a_n=O(b_n)$, $a_n\lesssim b_n$ or $b_n\gtrsim a_n$ if there exists a constant C such that $a_n\le Cb_n$ for all n. We write $a_n\asymp b_n$ if $a_n\lesssim b_n$ and $a_n\gtrsim b_n$. For a set A, we denote $|A|$ as its cardinality. Lastly, $C,C_0,C_1,...$ are constants that may vary from place to place.

2 GLOBAL HYPOTHESIS TESTING

In this section, we consider testing the global null hypotheses

$$H_0:\beta=0 \qquad \text{vs.} \qquad H_1:\beta\neq 0,$$

under the logistic regression model with random designs. The global testing problem corresponds to the detection of any associations between the covariates and the outcome.

Our construction of the global testing procedure begins with a bias-corrected estimator built upon a regularized estimator such as the ℓ_1-regularized M-estimator. For high-dimensional logistic regression, the ℓ_1-regularized M-estimator is defined as

$$\hat{\beta}=\arg\min_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[-y_i\beta^\top X_i+\log(1+e^{\beta^\top X_i})\right]+\lambda\|\beta\|_1\right\}, \tag{2.1}$$

which is the minimizer of a penalized log-likelihood function. Negahban et al. (2010) showed that, when X_i are i.i.d. sub-gaussian, under some mild regularity conditions, standard high-dimensional estimation error bounds for $\hat{\beta}$ under the ℓ_1 or ℓ_2 norm can be obtained by choosing $\lambda\asymp\sqrt{\log p/n}$. Once we obtain the initial estimator $\hat{\beta}$, our next step is to correct the bias of $\hat{\beta}$.

For technical reasons, we split the samples so that the initial estimation step and the bias correction step are conducted on separate and independent datasets. Without loss of generality, we assume there are $2n$ samples, divided into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, each with n independent samples. The initial estimator $\hat{\beta}$ is obtained from $\mathcal{D}_1$. In the following, we construct a nearly unbiased estimator $\check{\beta}$ based on $\hat{\beta}$ and the samples from $\mathcal{D}_2$, using the generalized LDP approach. Throughout the paper, the samples $Z_i=(X_i,Y_i)$, $i=1,...,n$, are from $\mathcal{D}_2$, which are independent of $\hat{\beta}$. We would like to emphasize that the sample splitting procedure is only used to simplify our theoretical analysis, which does not make it a restriction for practical applications. Numerically, as our simulations in Section 5 show, sample splitting is in fact not needed in order for our methods perform well (see further discussions in Section 7).

2.1 Construction of the Test Statistic via Generalized Low-Dimensional Projection

Let X be the design matrix whose i-th row is X_i. We rewrite the logistic regression model defined by (1.1) as

$$y_i = f(\beta^\top X_i) + \epsilon_i \tag{2.2}$$

where $f(u) = e^u/(1+e^u)$ and ϵ_i is error term. To correct the bias of the initial estimator $\hat{\beta}$, we consider the Taylor expansion of $f(u_i)$ at $\hat{u}_i$ for $u_i = \beta^\top X_i$ and $\hat{u}_i = \hat{\beta}^\top X_i$

$$f(u_i) = f(\hat{u}_i) + \dot{f}(\hat{u}_i)(u_i - \hat{u}_i) + Re_i$$

where Re_i is the reminder term. Plug this into the regression model (2.2), we have

$$y_i - f(\hat{u}_i) + \dot{f}(\hat{u}_i)X_i^\top \hat{\beta} = \dot{f}(\hat{u}_i)X_i^\top \beta + (Re_i + \epsilon_i). \tag{2.3}$$

By rewriting the logistic regression model as (2.3), we can treat $y_i - f(\hat{u}_i) + \dot{f}(\hat{u}_i)X_i^\top \hat{\beta}$ on the left hand side as the new response variable, whereas $\dot{f}(\hat{u}_i)X_i$ as the new covariates and $Re_i + \epsilon_i$ as the noise. Consequently, β can be considered as the regression coefficient of this approximate linear model.

The bias-corrected estimator, or, the generalized LDP estimator $\check{\beta}$ is defined as

$$\check{\beta}_j = \hat{\beta}_j + \frac{\sum_{i=1}^n v_{ij}(y_i - f(\hat{\beta}^\top X_i))}{\sum_{i=1}^n v_{ij}\dot{f}(\hat{\beta}^\top X_i)X_{ij}}, \quad j = 1, ..., p, \tag{2.4}$$

where X_{ij} is the j-th component of X_i and $v_j = (v_{1j}, v_{2j}, ..., v_{nj})^\top$ is the score vector that will be determined carefully (Ren et al., 2016; Cai et al., 2017). More specifically, we define the weighted inner product $\langle \cdot, \cdot \rangle_n$ for any $a, b \in \mathbb{R}^n$ as $\langle a, b \rangle_n = \sum_{i=1}^n \dot{f}(\hat{u}_i)a_i b_i$, and denote $\langle \cdot, \cdot \rangle$ as the ordinary inner product defined in Euclidean space. Combining (2.3) and (2.4), we can write

$$\check{\beta}_j - \beta_j = \frac{\langle v_j, \epsilon \rangle}{\langle v_j, \mathbf{x}_j \rangle_n} + \frac{\langle v_j, Re \rangle}{\langle v_j, \mathbf{x}_j \rangle_n} - \frac{\langle v_j, \mathbf{h}_{-j} \rangle_n}{\langle v_j, \mathbf{x}_j \rangle_n}, \tag{2.5}$$

where $\mathbf{x}_j \in \mathbb{R}^n$ denote the j-th column of X, $\mathbf{h}_{-j} = X_{-j}(\hat{\beta}_{-j} - \beta_{-j})$ where $X_{-j} \in \mathbb{R}^n \times \mathbb{R}^{p-1}$ is the submatrix of X without the j-th column, and $Re = (Re_1, ..., Re_n)^\top$ with $Re_i = f(u_i) - f(\hat{u}_i) - \dot{f}(\hat{u}_i)(u_i - \hat{u}_i)$. We will construct score vector v_j so that the first term on the right hand side of (2.5) is asymptotically normal, while the second and third terms, which together contribute to the bias of the generalized LDP estimator $\check{\beta}_j$, are negligible.

To determine the score vector v_j efficiently, we consider the following node-wise regression among the covariates

$$\mathbf{x}_j = X_{-j}\gamma_j + \eta_j, \qquad j = 1, ..., p, \tag{2.6}$$

where $\gamma_j = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \mathbb{E}[\|\mathbf{x}_j - X_{-j}\gamma\|_2^2]$ and η_j is the error term. Intuitively, if we set $v_j = \hat{W}^{-1}\eta_j$ for $\hat{W} = \mathrm{diag}(\dot{f}(\hat{u}_1), ..., \dot{f}(\hat{u}_n))$, then it should follow that

$$\langle v_j, \mathbf{h}_{-j}\rangle_n \leq \max_{k\neq j} |\langle v_j, \mathbf{x}_k\rangle_n| \cdot \|\hat{\beta} - \beta\|_1 = \max_{k\neq j} |\langle \eta_j, \mathbf{x}_k\rangle| \cdot \|\hat{\beta} - \beta\|_1 \approx 0.$$

In practice, we use the node-wise Lasso to obtain an estimate of η_j. For X from $\mathcal{D}_2$ and $\hat{\beta}$ obtained from $\mathcal{D}_1$, the score v_j is obtained by calibrating the Lasso-generated residue $\hat{\eta}_j$, i.e.

$$v_j(\lambda) = \hat{W}^{-1}\hat{\eta}_j(\lambda), \quad \hat{\eta}_j(\lambda) = \mathbf{x}_j - X_{-j}\hat{\gamma}_j(\lambda),$$

$$\hat{\gamma}_j(\lambda) = \arg\min_b \left\{ \frac{\|\mathbf{x}_j - X_{-j}b\|_2^2}{2n} + \lambda\|b\|_1 \right\}. \tag{2.7}$$

Clearly, $v_j(\lambda)$ depends on the tuning parameter λ. Define the following quantities

$$\zeta_j(\lambda) = \max_{k\neq j} \frac{|\langle v_j(\lambda), \mathbf{x}_k\rangle_n|}{\|v_j(\lambda)\|_n}, \qquad \tau_j(\lambda) = \frac{\|v_j(\lambda)\|_n}{|\langle v_j(\lambda), \mathbf{x}_j\rangle_n|}. \tag{2.8}$$

The tuning parameter λ can be determined through $\zeta_j(\lambda)$ and $\tau_j(\lambda)$ by the algorithm in Table 1, which is adapted from the algorithm in Zhang and Zhang (2014).

Table 1: Computation of v_j from the Lasso (2.7)

Input:	An upper bound ζ_j^* for ζ_j, with default value $\zeta^* = \sqrt{2\log p}$,
	tuning parameters $\kappa_0 \in [0, 1]$ and $\kappa_1 \in (0, 1]$;
Step 1:	If $\zeta_j(\lambda) > \zeta_j^*$ for all $\lambda > 0$, set $\zeta_j^* = (1 + \kappa_1)\inf_{\lambda>0} \zeta_j(\lambda)$;
	$\lambda \leftarrow \max\{\lambda : \zeta_j(\lambda) \leq \zeta_j^*\}, \zeta_j^* \leftarrow \zeta_j(\lambda), \tau_j^* \leftarrow \tau_j(\lambda)$;
Step 2:	$\lambda_j \leftarrow \min\{\lambda : \tau_j(\lambda) \leq (1 + \kappa_0)\tau_j^*\}$;
	$v_j \leftarrow v_j(\lambda_j), \tau_j \leftarrow \tau_j(\lambda_j), \zeta_j \leftarrow \zeta_j(\lambda_j)$
Output:	$\lambda_j, v_j, \tau_j, \zeta_j$

Once $\check{\beta}_j$ and τ_j are obtained, we define the standardized statistics

$$M_j = \check{\beta}_j / \tau_j,$$

for $j = 1, ..., p$. The global test statistic is then defined as

$$M_n = \max_{1\leq j\leq p} M_j^2. \tag{2.9}$$

2.2 Asymptotic Null Distribution

We now turn to the analysis of the properties of the global test statistic M_n defined in (A.1). For the random covariates, we consider both the Gaussian design and the bounded design. Under the

Gaussian design, the covariates are generated from a multivariate Gaussian distribution with an unknown covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$. In this case, we assume

($\mathbf{A1}$). $X_i \sim N(0, \mathbf{\Sigma})$ independently for each $i = 1, ..., n$.

In the case of bounded design, we assume instead

($\mathbf{A2}$). X_i for $i = 1, ..., n$ are i.i.d. random vectors satisfying $\mathbb{E}X_i = 0$ and $\max_{1 \le i \le n} \|X_i\|_\infty \le T$ for some constant $T > 0$.

Define the ℓ_1 ball

$$\mathcal{B}_1(k) = \left\{ \mathbf{\Omega} = (\omega_{ij}) \in \mathbb{R}^{p \times p} : \max_{1 \le i \le p} \sum_{j=1}^{p} \min\left(|\omega_{ij}| \sqrt{\frac{n}{\log p}}, 1 \right) \le k \right\}.$$

In general, $\mathcal{B}_1(k)$ includes any matrix $\mathbf{\Omega}$ whose rows ω_i are ℓ_0 sparse with $\|\omega_i\|_0 \le k$ or ℓ_1 sparse with $\|\omega_i\|_1 \le k\sqrt{\log p/n}$ for all $i = 1, ..., p$. The parameter space of the covariance matrix $\mathbf{\Sigma}$ and the regression vector β are defined as following.

($\mathbf{A3}$). The parameter space $\Theta(k)$ of $\theta = (\beta, \mathbf{\Sigma}) \in \mathbb{R}^p \times \mathbb{R}^{p \times p}$ satisfies

$$\Theta(k) = \left\{ (\beta, \mathbf{\Sigma}) : \quad \|\beta\|_0 \le k, M^{-1} \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le M, \mathbf{\Sigma}^{-1} \in \mathcal{B}_1(k) \right\},$$

for some constant $M \ge 1$. For convenience, we denote $\Theta_1(k) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le k\}$ and $\Theta_2(k) = \{\mathbf{\Sigma} \in \mathbb{R}^{p \times p} : M^{-1} \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le M, \mathbf{\Sigma}^{-1} \in \mathcal{B}_1(k)\}$, so that $\Theta(k) = \Theta_1(k) \times \Theta_2(k)$.

The following theorem states that the asymptotic null distribution of M_n under either the Gaussian or bounded design is a Gumbel distribution.

Theorem 1. *Let M_n be the test statistic defined in (A.1), D be the diagonal of $\mathbf{\Sigma}^{-1}$ and $(\xi_{ij}) = D^{-1/2}\mathbf{\Sigma}^{-1}D^{-1/2}$. Suppose $\max_{1 \le i < j \le p} |\xi_{ij}| \le c_0$ for some constant $0 < c_0 < 1$, $\log p = O(n^r)$ for some $0 < r < 1/5$, and*

1. *under the Gaussian design, we assume (A1) (A3) and $k = o\big(\sqrt{n}/\log^3 p\big)$; or*
2. *under the bounded design, we assume (A2) (A3) and $k = o\big(\sqrt{n}/\log^{5/2} p\big)$.*

Then under H_0, for any given $x \in \mathbb{R}$,

$$P_\theta\big(M_n - 2\log p + \log\log p \le x\big) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp(-x/2) \right), \quad \textit{as } (n, p) \to \infty.$$

The condition that $\log p = o(n^r)$ for some $0 < r < 1/5$ is consistent with those required for testing the global hypothesis in high-dimensional linear regression (Xia et al., 2018) and for testing two-sample covariance matrices (Cai et al., 2013). It allows the dimension p to be exponentially large comparing to the sample size n, which is much more flexible than the likelihood ratio test considered in Sur et al. (2017) and Sur and Candès (2019), where the dimension can only scale as

$p < n$. Under the Gaussian design, it is required that the sparsity k is $o\big(\sqrt{n}/\log^3 p\big)$ whereas for the bounded design, it suffices that the sparsity k to be $o\big(\sqrt{n}/\log^{5/2} p\big)$.

Remark 1. The analysis can be extended to testing $H_0 : \beta_G = 0$ versus $H_1 : \beta_G \neq 0$ for a given index set G. Specifically, we can construct the test statistic as $M_{G,n} = \max_{i\in G} M_j^2$ and obtain a similar Gumbel limiting distribution by replacing p by $|G|$, as $(n, |G|) \to \infty$. The sparsity condition thus should be forwarded to the set G.

Based on the limiting null distribution, the asymptotically α level test can be defined as

$$\Phi_\alpha(M_n) = I\{M_n \geq 2\log p - \log\log p + q_\alpha\},$$

where q_α is the $1-\alpha$ quantile of the Gumbel distribution with the cumulative distribution function $\exp\big(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\big)$, i.e.

$$q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}.$$

The null hypothesis H_0 is rejected if and only if $\Phi_\alpha(M_n) = 1$.

2.3 Minimax Separation Distance and Optimality

In this subsection, we answer the question: "What is the essential difficulty for testing the global hypothesis in logistic regression." To fix ideas, we begin with defining the minimax separation distance that measures such an essential difficulty for testing the global null hypothesis at a given level and type II error. In particular, we consider the alternative

$$H_1 : \beta \in \Big\{\beta \in \mathbb{R}^p : \|\beta\|_\infty \geq \rho, \|\beta\|_0 \leq k\Big\}$$

for some $\rho > 0$. This alternative concerns the detection of any discernible signals among the regression coefficients where the signals can be extremely sparse, which has interesting applications (see Xia et al. (2015)). Similar alternatives are also considered by Cai et al. (2013) and Cai et al. (2014).

By fixing a level $\alpha > 0$ and a type II error probability $\delta > 0$, we can define the δ-separation distance of a level α test procedure Φ_α for given design covariance $\mathbf{\Sigma}$ as

$$\begin{aligned}\rho(\Phi_\alpha, \delta, \mathbf{\Sigma}) &= \inf\Big\{\rho > 0 : \inf_{\beta\in\Theta_1(k):\|\beta\|_\infty\geq\rho} P_\theta(\Phi_\alpha = 1) \geq 1-\delta\Big\} \\ &= \inf\Big\{\rho > 0 : \sup_{\beta\in\Theta_1(k):\|\beta\|_\infty\geq\rho} P_\theta(\Phi_\alpha = 0) \leq \delta\Big\}. \qquad (2.10)\end{aligned}$$

The δ-separation distance $\rho(\Phi_\alpha, \delta, \Theta(k))$ over $\Theta(k)$ can thus be defined by taking the supremum over all the covariance matrices $\mathbf{\Sigma} \in \Theta_2(k)$, so that

$$\rho(\Phi_\alpha, \delta, \Theta(k)) = \sup_{\mathbf{\Sigma} \in \Theta_2(k)} \rho(\Phi_\alpha, \delta, \mathbf{\Sigma}),$$

which corresponds to the minimal ℓ_∞ distance such that the null hypothesis H_0 is well separated from the alternative H_1 by the test Φ_α. In general, δ-separation distance is an analogue of the statistical risk in estimation problems. It characterizes the performance of a specific α-level test with a guaranteed type II error δ. Consequently, we can define the (α, δ)-minimax separation distance over $\Theta(k)$ and all the α-level tests as

$$\rho^*(\alpha, \delta, \Theta(k)) = \inf_{\Phi_\alpha} \rho(\Phi_\alpha, \delta, \Theta(k)).$$

The definition of (α, δ)-minimax separation distance generalizes the ideas of Ingster (1993), Baraud (2002) and Verzelen (2012). The following theorem establishes the minimax lower bound of the (α, δ)-separation distance under the Gaussian design for testing the global null hypothesis over the parameter space $\Theta'(k) \subset \Theta(k)$ defined as

$$\Theta'(k) = \big(\Theta_1(k) \cap \{\beta \in \mathbb{R}^p : \|\beta\|_2 \lesssim (n^{1/4} \log p)^{-1}\}\big) \times \Theta_2(k).$$

Theorem 2. *Assume that $\alpha + \delta \leq 1$. Under the Gaussian design, if (A1) and (A3) hold, $(\beta, \mathbf{\Sigma}) \in \Theta'(k)$ and $k \lesssim \min\{p^\gamma, \sqrt{n}/\log^3 p\}$ for some $0 < \gamma < 1/2$, then the (α, δ)-minimax separation distance over $\Theta'(k)$ has the lower bound*

$$\rho^*(\alpha, \delta, \Theta'(k)) \geq c\sqrt{\frac{\log p}{n}} \tag{2.11}$$

for some constant $c > 0$.

In order to show the above lower bound is asymptotically sharp, we prove that it is actually attainable under certain circumstances, by our proposed global test Φ_α. In particular, for the bounded design, we make the following additional assumption.

(**A4**). It holds that $P_\theta(\max_{1 \leq i \leq n} |\beta^\top X_i| \geq C) = O(p^{-c})$ for some constant $C, c > 0$.

Theorem 3. *Suppose that $\log p = O(n^r)$ for some $0 < r < 1$. Under the alternative $H_1 : \|\beta\|_\infty \geq c_2\sqrt{\log p/n}$ for some $c_2 > 0$, and*

(i) under the Gaussian design, assume that (A1) and (A3) hold, $\|\beta\|_2 \leq C(\log\log p)/\sqrt{\log n}$ for $C \leq \min\{\sqrt{2/\lambda_{\max}(\mathbf{\Sigma})}, (2r\sqrt{2\lambda_{\max}(\mathbf{\Sigma})})^{-1}\}$, $\log p \gtrsim \log^{1+\delta} n$ for some $\delta > 0$ and $k = o(\sqrt{n}/\log^3 p)$; or

(ii) under the bounded design, assume that (A2), (A3), and (A4) hold, and $k = o(\sqrt{n}/\log^{5/2} p)$.

Then we have $P_\theta\big(\Phi_\alpha(M_n)=1\big)\to 1$ *as* $(n,p)\to\infty$.

In Theorem 3, (A4) is assumed for the bounded case and $\|\beta\|_2 = O(\log\log p/\sqrt{\log n})$ is required for the Gaussian case. In particular, since $\log p = O(n^r)$ for some $0<r<1$, the upper bound $\log\log p/\sqrt{\log n}$ for $\|\beta\|_2$ can be as large as $\sqrt{\log n}$. In Theorem 2, the minimax lower bound is established over $(\beta,\boldsymbol{\Sigma})\in\Theta'(k)$, so that the same lower bound holds over a larger set

$$(\beta,\boldsymbol{\Sigma})\in\big(\Theta_1(k)\cap\{\beta\in\mathbb{R}^p:\|\beta\|_2\le\log\log p/\sqrt{\log n}\}\big)\times\Theta_2(k), \tag{2.12}$$

since $\log\log p/\sqrt{\log n}\gtrsim(n^{1/4}\log p)^{-1}$. On the other hand, Theorem 3 (i) indicates an upper bound $\rho^*\lesssim\sqrt{\log p/n}$ attained by our proposed test under the Gaussian design over the set (2.12). These two results imply the minimax rate $\rho^*\asymp\sqrt{\log p/n}$ and the minimax optimality of our proposed test over the set (2.12).

2.4 Comparison with Existing Works

In this section, we make detailed comparisons and connections with some existing works concerning global hypothesis testing in the high-dimensional regression literature.

Ingster et al. (2010) addressed the detection boundary for high-dimensional sparse linear regression models, and more recently Mukherjee et al. (2015) studied the detection boundary for hypothesis testing in high-dimensional sparse binary regression models. However, although both works obtained the sharp detection boundary for the global testing problem $H_0:\beta=0$, their alternative hypotheses are different from ours. Specifically, Mukherjee et al. (2015) considered the alternative hypothesis $H_1:\beta\in\big\{\beta\in\mathbb{R}^p:\|\beta\|_0\ge k,\min\{|\beta_j|:\beta_k\neq 0\}\ge A\big\}$, which implies that β has at least k nonzero coefficients exceeding A in absolute values. Ingster et al. (2010) considered the alternative hypothesis $H_1:\beta\in\big\{\beta\in\mathbb{R}^p:\|\beta\|_0\le k,\|\beta\|_2\ge\rho\big\}$, which concerns k sparse β with ℓ_2 norm at least ρ. In fact, the proof of our Theorem 2 can be directly extended to such an alternative concerning the ℓ_2 norm, which amounts to obtaining a lower bound of order $\sqrt{\frac{k\log p}{n}}$ for high dimensional logistic regression. However, developing a minimax optimal test for such alternative is beyond the scope of the current paper.

Additionally, in contrast to the minimax separation distance considered in this paper, the papers by Ingster et al. (2010) and Mukherjee et al. (2015) considered the minimax risk (or the minimax total error probability) given by

$$\inf_{\Phi}\sup_{\boldsymbol{\Sigma}\in\Theta_2(k)}\mathrm{Risk}(\Phi,\boldsymbol{\Sigma})=\inf_{\Phi}\sup_{\boldsymbol{\Sigma}\in\Theta_2(k)}\left\{\max_{\beta\in H_0}P_\theta(\Phi=1)+\max_{\beta\in\Theta_1(k):\|\beta\|_\infty\ge\rho}P_\theta(\Phi=0)\right\}, \tag{2.13}$$

where the infimum is taken over all tests Φ. This minimax risk can be also written as

$$\inf_{\Phi}\sup_{\boldsymbol{\Sigma}\in\Theta_2(k)}\mathrm{Risk}(\Phi,\boldsymbol{\Sigma})=\inf_{\alpha\in(0,1)}\left\{\alpha+\inf_{\Phi_\alpha}\sup_{\boldsymbol{\Sigma}\in\Theta_2(k)}\sup_{\beta\in\Theta_1(k):\|\beta\|_\infty\ge\rho}P_\theta(\Phi_\alpha=0)\right\}. \tag{2.14}$$

A comparison of (2.10) and (2.14) yields the slight difference between the two criteria, as one depends on a given Type I error α and the other doesn't.

Moreover, these two papers considered different design scenarios from ours. In Ingster et al. (2010), only the isotropic Gaussian design was considered. As a result, the optimal tests proposed therein rely highly on the independence assumption. In Mukherjee et al. (2015), the general binary regression was studied under fixed sparse design matrices. In particular, the minimax lower and upper bounds were only derived in the special case of design matrices with binary entries and certain sparsity structures.

In comparison with the recent works of Sur et al. (2017), Candès and Sur (2018) and Sur and Candès (2019), besides the aforementioned difference in the asymptotics of (p, n), these two papers only considered the random Gaussian design, whereas our work also considered random bounded design as in van de Geer et al. (2014). In addition, Sur et al. (2017) and Sur and Candès (2019) developed the Likelihood Ratio (LLR) Test for testing the hypothesis $H_0 : \beta_{j_1} = \beta_{j_2} = ... = \beta_{j_k} = 0$ for any finite k. Intuitively, a valid test for the global null and $p/n \to \kappa \in (0, 1/2)$ can be adapted from the individual LLR tests using the Bonferroni procedure. However, as our simulations show (Section 5), such a test is less powerful compared to our proposed test.

Lastly, our minimax results focus on the highly sparse regime $k \lesssim p^{\gamma}$ where $\gamma \in (0, 1/2)$. As shown by Ingster et al. (2010) and Mukherjee et al. (2015), the problem under the dense regime where $\gamma \in (1/2, 1)$ can be very different from the sparse regime. Mostly likely, the fundamental difficulty of the testing problem changes in this situation so that different methods need to be carefully developed. We leave these interesting questions for future investigations.

3 LARGE-SCALE MULTIPLE TESTING

Denote by β the true coefficient vector in the model and denote $\mathcal{H}_0 = \{j : \beta_j = 0, j = 1, \cdots, p\}, \mathcal{H}_1 = \{j : \beta_j \neq 0, j = 1, \cdots, p\}$. In order to identify the indices in $\mathcal{H}_1$, we consider simultaneous testing of the following null hypotheses

$$H_{0,j} : \beta_j = 0 \quad \text{vs.} \quad H_{1,j} : \beta_j \neq 0, \quad 1 \leq j \leq p.$$

Apart from identifying as many nonzero β_j as possible, to obtain results of practical interest, we would like to control the false discovery rate (FDR) as well as the false discovery proportion (FDP), or the number of falsely discovered variables (FDV).

3.1 Construction of Multiple Testing Procedures

Recall that in Section 2, we define the standardized statistics $M_j = \check{\beta}_j / \tau_j$, for $j = 1, ..., p$. For a given threshold level $t > 0$, each individual hypothesis $H_{0,j} : \beta_j = 0$ is rejected if $|M_j| \geq t$.

Therefore for each t, we can define

$$\mathrm{FDP}_\theta(t) = \frac{\sum_{j\in\mathcal{H}_0} I\{|M_j| \geq t\}}{\max\big\{\sum_{j=1}^p I\{|M_j| \geq t\}, 1\big\}}, \qquad \mathrm{FDR}_\theta(t) = \mathbb{E}_\theta[\mathrm{FDP}(t)],$$

and the expected number of falsely discovered variables $\mathrm{FDV}_\theta(t) = \mathbb{E}_\theta\big[\sum_{j\in\mathcal{H}_0} I\{|M_j| \geq t\}\big]$.

Procedure Controlling FDR/FDP. In order to control the FDR/FDP at a pre-specified level $0 < \alpha < 1$, we can set the threshold level as

$$\tilde{t}_1 = \inf\left\{0 \leq t \leq b_p : \frac{\sum_{j\in\mathcal{H}_0} I\{|M_j| \geq t\}}{\max\big\{\sum_{j=1}^p I\{|M_j| \geq t\}, 1\big\}} \leq \alpha\right\}, \tag{3.1}$$

for some b_p to be determined later.

In general, the ideal choice $\tilde{t}_1$ is unknown and needs to be estimated because it depends on the knowledge of the true null $\mathcal{H}_0$. Let $G_0(t)$ be the proportion of the nulls falsely rejected by the procedure among all the true nulls at the threshold level t, namely, $G_0(t) = \frac{1}{p_0}\sum_{j\in\mathcal{H}_0} I\{|M_j| \geq t\}$, where $p_0 = |\mathcal{H}_0|$. In practice, it is reasonable to assume that the true alternatives are sparse. If the sample size is large, we can use the tails of normal distribution $G(t) = 2 - 2\Phi(t)$ to approximate $G_0(t)$. In fact, it will be shown that, for $b_p = \sqrt{2\log p - 2\log\log p}$, $\sup_{0\leq t\leq b_p}\big|\frac{G_0(t)}{G(t)} - 1\big| \to 0$ in probability as $(n,p) \to \infty$. To summarize, we have the following logistic multiple testing (LMT) procedure controlling the FDR and the FDP.

Procedure 1 (LMT)**.** *Let $0 < \alpha < 1$, $b_p = \sqrt{2\log p - 2\log\log p}$ and define*

$$\hat{t} = \inf\left\{0 \leq t \leq b_p : \frac{pG(t)}{\max\big\{\sum_{j=1}^p I\{|M_j| \geq t\}, 1\big\}} \leq \alpha\right\}. \tag{3.2}$$

If $\hat{t}$ in (3.2) does not exist, then let $\hat{t} = \sqrt{2\log p}$. We reject $H_{0,j}$ whenever $|M_j| \geq \hat{t}$.

Procedure Controlling FDV. For large-scale inference, it is sometimes of interest to directly control the number of falsely discovered variables (FDV) instead of the less stringent FDR/FDP, especially when the sample size is small (Liu and Luo, 2014). By definition, the FDV control, or equivalently, the per-family error rate control, provides an intuitive description of the Type I error (false positives) in variable selection. Moreover, controlling $\mathrm{FDV} = r$ for some $0 < r < 1$ is related to the family-wise error rate (FWER) control, which is the probability of at least one false positive. In fact, FDV control can be achieved by a suitable modification of the FDP controlling procedure introduced above. Specifically, we propose the following FDV (or FWER) controlling logistic multiple testing (LMT_V) procedure.

Procedure 2 (LMT_V)**.** *For a given tolerable number of falsely discovered variables $r < p$ (or a* *desired level of FWER $0 < r < 1$), let $\hat{t}_{FDV} = G^{-1}(r/p)$. $H_{0,j}$ is rejected whenever $|M_j| \geq \hat{t}_{FDV}$.*

3.2 Theoretical Properties for Multiple Testing Procedures

In this section we show that our proposed multiple testing procedures control the theoretical FDR/FDP or FDV asymptotically. For simplicity, our theoretical results are obtained under the bounded design scenario. For FDR/FDP control, we need an additional assumption on the interplay between the dimension p and the parameter space $\Theta(k)$.

Recall that $\eta_j = (\eta_{j1}, ..., \eta_{jn})^\top$ for $j = 1, ..., p$ defined in (2.6). We define $F_{jk} = \mathbb{E}_\theta[\eta_{ij}\eta_{ik}/\dot{f}(u_i)]$ for $1 \leq j, k \leq p$, and $\rho_{jk} = F_{jk}/\sqrt{F_{jj}F_{kk}}$. Denote $\mathcal{B}(\delta) = \{(j,k) : |\rho_{jk}| \geq \delta, i \neq j\}$ and $\mathcal{A}(\epsilon) = \mathcal{B}((\log p)^{-2-\epsilon})$.

(**A5**). Suppose that for some $\epsilon > 0$ and $q > 0$, $\sum_{(j,k)\in\mathcal{A}(\epsilon):j,k\in\mathcal{H}_0} p^{\frac{2|\rho_{jk}|}{1+|\rho_{jk}|}+q} = O(p^2/(\log p)^2)$.

The following proposition shows that M_j is asymptotically normal distributed and $G_0(t)$ is well approximated by $G(t)$.

Proposition 1. *Under (A2) (A3) and (A4), suppose $p = O(n^c)$ for some constant $c > 0$, $k = o(\sqrt{n}/\log^{5/2} p)$, then as $(n, p) \to \infty$,*

$$\sup_{j\in\mathcal{H}_0} \sup_{0\leq t\leq\sqrt{2\log p}} \left| \frac{P_\theta(|M_j| \geq t)}{2 - 2\Phi(t)} - 1 \right| \to 0. \tag{3.3}$$

If in addition we assume (A5), then

$$\sup_{0\leq t\leq b_p} \left| \frac{G_0(t)}{G(t)} - 1 \right| \to 0 \tag{3.4}$$

in probability, where Φ is the cumulative distribution function of the standard normal distribution and $b_p = \sqrt{2\log p - 2\log\log p}$.

The following theorem provides the asymptotic FDR and FDP control of our procedure.

Theorem 4. *Under the conditions of Proposition 1, for $\hat{t}$ defined in our LMT procedure, we have*

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}_\theta(\hat{t})}{\alpha p_0/p} \leq 1, \qquad \lim_{(n,p)\to\infty} P_\theta\left(\frac{\mathrm{FDP}_\theta(\hat{t})}{\alpha p_0/p} \leq 1+\epsilon\right) = 1 \tag{3.5}$$

for any $\epsilon > 0$.

For the FDV/FWER controlling procedure, we have the following theorem.

Theorem 5. *Under (A2) (A3) and (A4), assume $p = O(n^c)$ for some $c > 0$ and $k = o(\sqrt{n}/\log^{5/2} p)$. Let $r < p$ be the desired level of FDV. For $\hat{t}_{FDV}$ defined in our LMT$_V$ procedure, we have $\lim_{(n,p)\to\infty} \frac{\mathrm{FDV}_\theta(\hat{t}_{FDV})}{rp_0/p} \leq 1$. In addition, if $0 < r < 1$, we have $\lim_{(n,p)\to\infty} \frac{\mathrm{FWER}_\theta(\hat{t}_{FDV})}{rp_0/p} \leq 1$.*

The above theoretical results are obtained under the dimensionality condition $p = O(n^c)$, which is stronger than that of the global test. Essentially, the condition is needed to obtain the uniform convergence (3.3), whose form (as ratio) is stronger than the convergence in distribution in the ordinary sense (as direct difference).

4 TESTING FOR TWO LOGISTIC REGRESSION MODELS

In some applications, it is also interesting to consider hypothesis testing that involves two separate logistic regression models of the same dimension. Specifically, for $\ell = 1, 2$ and $i = 1, ..., n_\ell$, where $n_1 \asymp n_2$, $y_i^{(\ell)} = f(\beta^{(\ell)\top} X_i^{(\ell)}) + \epsilon_i^{(\ell)}$, where $f(u) = e^u/(1 + e^u)$, and $\epsilon_i^{(\ell)}$ is a binary random variable such that $y_i^{(\ell)} | X_i^{(\ell)} \sim \text{Bernoulli}(f(\beta^{(\ell)\top} X_i^{(\ell)}))$. The global null hypothesis $H_0 : \beta^{(1)} = \beta^{(2)}$ implies that there is overall no difference in association between covariates and the response. If this null hypothesis is rejected, we are interested in simultaneously testing the hypotheses $H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)}$ for each $j = 1, ..., p$.

To test the global null $H_0 : \beta^{(1)} = \beta^{(2)}$ against $H_1 : \beta^{(1)} \neq \beta^{(2)}$, we can first obtain $\breve{\beta}_j^{(\ell)}$ and $\tau_j^{(\ell)}$ for each model, and then calculate the coordinate-wise standardized statistics $T_j = \frac{\breve{\beta}_j^{(1)}}{\sqrt{2}\tau_j^{(1)}} - \frac{\breve{\beta}_j^{(2)}}{\sqrt{2}\tau_j^{(2)}}$, for $j = 1, ..., p$. Define the global test statistic as $T_n = \max_{1 \le j \le p} T_j^2$, it can be shown that the limiting null distribution is also a Gumbel distribution. The α level global test is thus defined as $\Phi_\alpha(T_n) = I\{T_n \ge 2\log p - \log\log p + q_\alpha\}$, where $q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}$. For multiple hypotheses testing of two regression vectors $H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)}$ for $j = 1, ..., p$, we consider the test statistics T_j defined above. The two-sample multiple testing procedure controlling FDR/FDP is given as follows.

Procedure 3. *Let* $0 < \alpha < 1$ *and define* $\hat{t} = \inf\left\{0 \le t \le b_p : \frac{pG(t)}{\max\left\{\sum_{j=1}^p I\{|T_j| \ge t\}, 1\right\}} \le \alpha\right\}$. *If the above* $\hat{t}$ *does not exist, let* $\hat{t} = \sqrt{2\log p}$. *We reject* $H_{0,j}$ *whenever* $|T_j| \ge \hat{t}$.

5 SIMULATION STUDIES

In this section we examine the numerical performance of the proposed tests. Due to the space limit, for both global and multiple testing problems, we only focus on the single regression setting, and report the results on two logistic regressions in the Supplementary Materials. Throughout our numerical studies, sample splitting was not used.

5.1 Global Hypothesis Testing

In the following simulations, we consider a variety of dimensions, sample sizes, and sparsity of the models. For alternative hypotheses, the dimension of the covariates p ranges from 100, 200, 300 to 400, and the sparsity k is set as 2 or 4. The sample sizes n are determined by the ratio $r = p/n$ that takes values of 0.2, 0.4 and 1.2. To generate the design matrix X, we consider the Gaussian design with the blockwise-correlated covariates so that $\mathbf{\Sigma} = \mathbf{\Sigma}_B$, where $\mathbf{\Sigma}_B$ is a $p \times p$ blockwise diagonal matrix including 10 equal-sized blocks, whose diagonal elements are 1's and off-diagonal elements are set as 0.7. Under the alternative, suppose $\mathcal{S}$ is the support of the regression coefficients β and $|\mathcal{S}| = k$, we set $|\beta_j| = \rho 1\{j \in \mathcal{S}\}$ for $j = 1, ..., p$ and $\rho = 0.75$ with equal proportions of ρ and $-\rho$. We set $\kappa_0 = 0$ and $\kappa_1 = 0.5$.

To assess the empirical performance of our proposed test ("Proposed"), we compare our test with (i) a Bonferroni procedure applied to the p-values from univariate screening using MLE statistic ("U-S"), and (ii) to the method of Sur et al. (2017); Sur and Candès (2019) ("LLR") in the setting where $r = 0.2$ and 0.4.

Table 2 shows the empirical type I errors of these tests at level $\alpha = 0.05$ based on 1000 simulations. Figure 1 shows the corresponding empirical powers under various settings. As we expected, our proposed method outperforms the other two alternatives in all the cases (including the moderate dimensional cases where $r = 0.2$ and 0.4), and the power increases as n or p grows. In the rather lower dimensional setting where $r = 0.2$, the LLR performs almost as well as our proposed method.

Table 2: Type I error with $\alpha = 0.05$ for the proposed method (Proposed), the Bonferroni corrected univariate screening method (U-S) and the Bonferroni corrected likelihood ratio based method of Sur and Candès (2019) (LLR), for different n, p and k.

p/n	$k=2$				$k=4$			
	$p=100$	200	300	400	$p=400$	600	800	1000
				Proposed				
0.2	0.052	0.066	0.042	0.054	0.058	0.050	0.046	0.070
0.4	0.038	0.054	0.062	0.054	0.046	0.050	0.060	0.074
1.2	0.026	0.044	0.042	0.045	0.014	0.044	0.054	0.054
				U-S				
0.2	0.040	0.032	0.024	0.018	0.018	0.022	0.028	0.034
0.4	0.050	0.032	0.024	0.020	0.028	0.028	0.032	0.046
1.2	0.028	0.038	0.024	0.020	0.032	0.018	0.034	0.014
				LLR				
0.2	0.050	0.050	0.068	0.040	0.058	0.044	0.046	0.034
0.4	0.084	0.070	0.048	0.056	0.062	0.042	0.058	0.064

5.2 Multiple Hypotheses Testing

FDR Control. In this case, we set $p = 800$ and let n vary from 600, 800, 1000, 1200 to 1400, so that all the cases are high-dimensional in the sense that $p > n/2$. The sparsity level k varies from 40, 50 to 60. For the true positives, given the support $\mathcal{S}$ such that $|\mathcal{S}| = k$, we set $|\beta_j| = \rho 1\{j \in \mathcal{S}\}$ for $j = 1, ..., p$ with equal proportions of ρ and $-\rho$. The design covariates X_i's are generated from

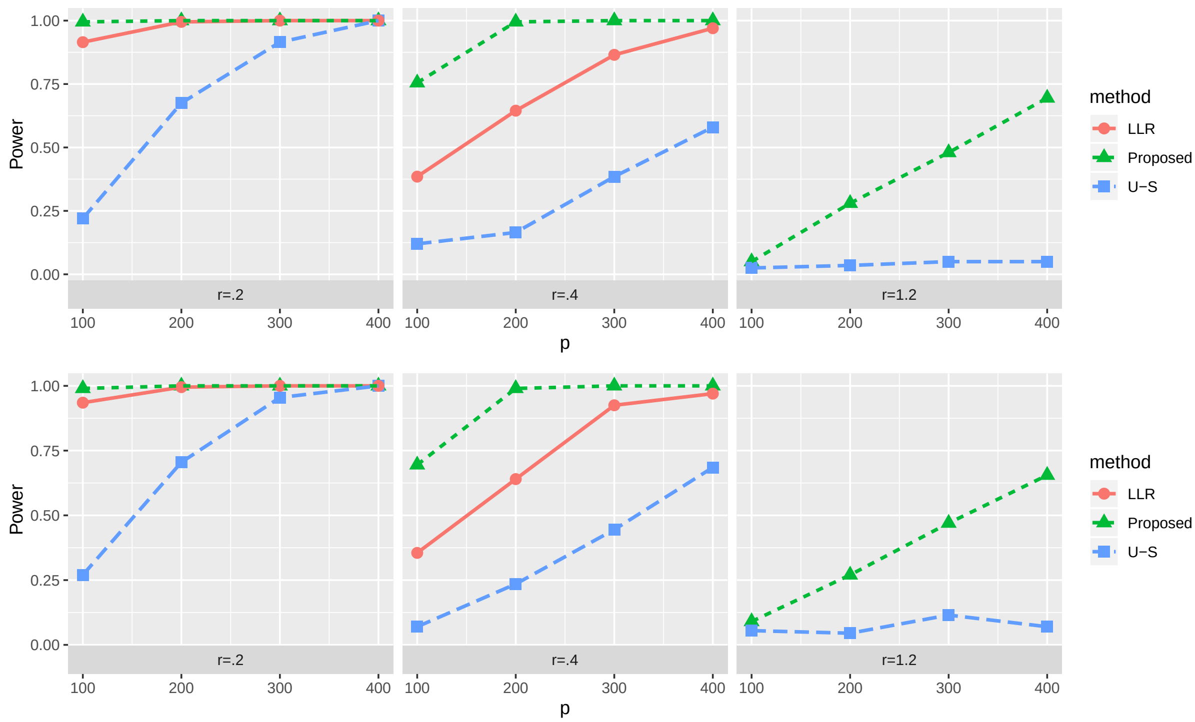

Figure 1: Empirical power with $\alpha = 0.05$ for the proposed method (Proposed), the Bonferroni corrected univariate screening method (U-S) and the Bonferroni corrected likelihood ratio based method of Sur and Candès (2019) (LLR). Top panel: $k = 2$; bottom panel: $k = 4$.

a ($|X_i^\top \beta| < 3$)-truncated multivariate Gaussian distribution with covariance matrix $\mathbf{\Sigma} = 0.01\mathbf{\Sigma}_M$, where $\mathbf{\Sigma}_M$ is a $p \times p$ blockwise diagonal matrix of 10 identical unit diagonal Toeplitz matrices whose off-diagonal entries descend from 0.1 to 0 (see Supplementary Material for the explicit form). The choice of κ_0 and κ_1 are the same as the global testing. Throughout, we set the desired FDR level as $\alpha = 0.2$.

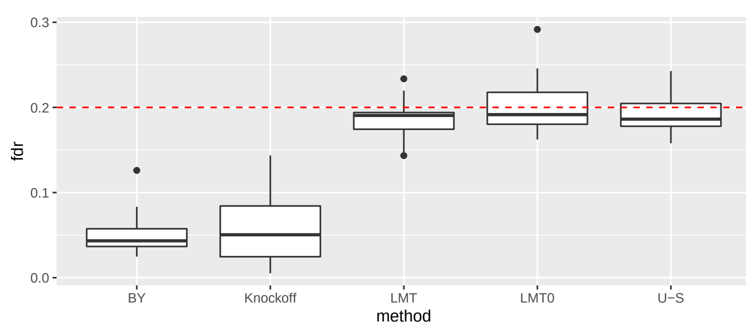

Figure 2: Boxplots of the empirical FDRs across all the settings for $\alpha = 0.2$.

We compare our proposed procedure (denoted as "LMT") with following methods: (i) the basic

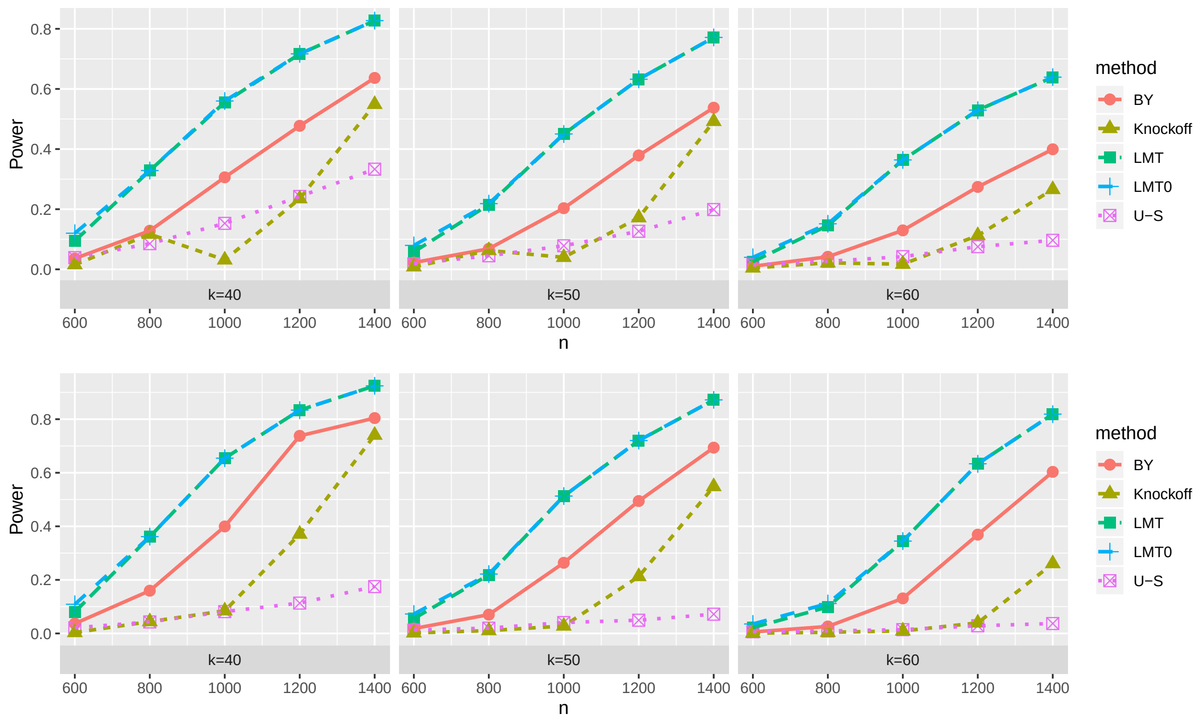

Figure 3: Empirical power under FDR $\alpha = 0.2$ for $\rho = 3$ (top) and $\rho = 4$ (bottom).

LMT procedure with b_p in (3.2) replaced by ∞ ("LMT0"), which is equivalent to applying the BH procedure (Benjamini and Hochberg, 1995) to our debiased statistics M_j, (ii) the BY procedure (Benjamini and Yekutieli, 2001) using our debiased statistics M_j ("BY"), implemented using the `R` function `p.adjust(...,method="BY")`, (iii) a BH procedure applied to the p-values from univariate screening using the MLE statistics ("U-S"), and (iv) the knockoff method of Candès et al. (2018) ("Knockoff"). Figure 2 shows boxplots of the pooled empirical FDRs (see Supplementary Material for the case-by-case FDRs) and Figure 4 shows the empirical powers of these methods based on 1000 replications. Here the power is defined as the number of correctly discovered variables divided by the number of truly associated variables. As a result, we find that LMT and LMT0 correctly control FDRs and have the greatest power among all the cases. In particular, the power of LMT and LMT0 are almost the same, which increases as the sparsity decreases, the signal magnitude ρ increases, or the sample size n increases, while LMT0 has slightly inflated FDRs. The U-S method, although correctly controls the FDRs, has poor power, which is largely due to the dependence among the covariates.

FDV Control. For our proposed test that controls FDV (denoted as LMT_V), by setting desired FDV level $r = 10$, we apply our method to various settings. Specifically, we set $\rho = 3$, $p \in \{800, 1000, 1200\}$, set $k \in \{40, 50, 60\}$, and let n vary from 400, 600, 800 to 1000. The design covariates are generated similarly as the previous part. The resulting empirical FDV and powers are summarized in Table 3. Our proposed LMT_V has the correct control of FDV in all the settings and the power increases as n grows, k decreases, or p decreases.

Table 3: Empirical performance of LMT_V with FDV level $r = 10$.

ρ	p	k	Empirical FDV				Empirical Power			
			$n = 400$	600	800	1000	400	600	800	1000
3	800	40	4.07	5.45	6.44	7.11	0.08	0.23	0.40	0.59
		50	4.30	6.29	7.27	8.26	0.06	0.16	0.32	0.49
		60	4.33	6.63	7.48	8.42	0.05	0.12	0.25	0.42
	1000	40	3.30	4.59	5.79	6.82	0.06	0.18	0.35	0.52
		50	3.49	5.42	6.43	7.03	0.05	0.13	0.26	0.43
		60	3.68	5.47	7.29	7.97	0.03	0.09	0.20	0.34
	1200	40	2.69	4.36	5.00	5.68	0.05	0.15	0.31	0.46
		50	2.97	4.22	5.73	6.43	0.03	0.11	0.21	0.36
		60	2.78	4.91	5.91	7.25	0.02	0.07	0.16	0.27

6 REAL DATA ANALYSIS

We illustrate our proposed methods by analyzing a dataset from the Pediatric Longitudinal Study of Elemental Diet and Stool Microbiome Composition (PLEASE) study, a prospective cohort study to investigate the effects of inflammation, antibiotics, and diet as environmental stressors on the gut microbiome in pediatric Crohn's disease (Lewis et al., 2015; Lee et al., 2015; Ni et al., 2017). The study considered the association between pediatric Crohn's disease and fecal metabolomics by collecting fecal samples of 90 pediatric patients with Crohn's disease at baseline, 1 week, and 8 weeks after initiation of either anti-tumor necrosis factor (TNF) or enteral diet therapy, as well as those from 25 healthy control children (Lewis et al., 2015). In details, an untargeted fecal metabolomic analysis was performed on these samples using liquid chromatography-mass spectrometry (LC-MS). Metabolites with more than 80% missing values across all samples were removed from the analysis. For each metabolite, samples with the missing values were imputed with its minimum abundance across samples. To avoid potential large outliers, for each sample, the metabolite abundances were further normalized by dividing 90% cumulative sum of the abundances of all metabolites. The normalized abundances were then log transformed and used in all analyses. The metabolomics annotation was obtained from Human Metabolome Database (Lee et al., 2015). In total, for each sample, abundances of 335 known metabolites were obtained and used in our analysis.

6.1 Association Between Metabolites and Crohn's Disease Before and After Treatment

We first test the overall association between 335 characterized metabolites and Crohn's disease by fitting a logistic regression using the data of 25 healthy controls and 90 Crohn's disease patients at the baseline. We obtain a global test statistic of 433.88 with a p-value < 0.001, indicating a strong association between Crohn's disease and fecal metabolites. At the FDR $< 5\%$, our multiple testing procedure selects four metabolites, including C14:0.sphingomyelin, C24:1.Ceramide.(d18:1) and 3-methyladipate/pimelate (see Table 4). Recent studies have demonstrated that sphingolipid metabolites, particularly ceramide and sphingosine-1-phosphate, are signaling molecules that regulate a diverse range of cellular processes that are important in immunity, inflammation and inflammatory disorders (Maceyka and Spiegel, 2014). In fact, ceramide acts to reduce tumor necrosis factor (TNF) release (Rozenova et al., 2010) and has important roles in the control of autophagy, a process strongly implicated in the pathogenesis of Crohn's disease (Barrett et al., 2008; Sewell et al., 2012).

We next investigate whether treatment of Crohn's disease alters the association between metabolites and Crohn's disease by fitting two separate logistic regressions using the metabolites measured one week or 8 weeks after the treatment. At each time point, a significant association is detected based on our global test (p-value < 0.001). One week after the treatment, we observe six metabolites associated with Crohn's disease, including all four identified at the baseline and two additional metabolites, beta-alanine and adipate (see Table 4). The beta-alanine and adipate associations are likely due to that beta-alanine and adipate are important ingredients of the enteral nutrition treatment of Crohn's disease. However, it is interesting that at 8 weeks after the treatment, valine, C16.carnitine and C18.carnitine are identified to be associated with Crohn's disease together with 3-methyladipate/pimelate and beta-alanine. It is known that carnitine plays an important role in Crohn's disease, which might be a consequence of the underlying functional association between Crohn's disease and mutations in the carnitine transporter genes (Peltekova et al., 2004; Fortin, 2011). Deficiency of carnitine can lead to severe gut atrophy, ulceration and inflammation in animal models of carnitine deficiency (Shekhawat et al., 2013). Our results may suggest that the treatment increases carnitine, leading to reduction of inflammation.

6.2 Comparison of Metabolite Associations Between Responders and Non-Responders

To compare the metabolic association with Crohn's disease for responders ($n = 47$) and non-responders ($n = 34$) eight weeks after treatment, we fit two logistic regression models, responder versus normal control and non-responder versus normal control. Our global test shows that there is an overall difference in regression coefficients for responders and for non-responders when compared to the normal controls (p-value < 0.001). We next apply our proposed multiple testing procedure to identify the metabolites that have different regression coefficients in these two different logis-

Table 4: Significant metabolites associated with Crohn's disease (coded as 1 in logistic regression) at the baseline, one week and 8 weeks after treatment with FDR $< 5\%$. The refitted regression coefficients show the direction of the association.

Disease Stage	HMDB ID	Synonyms	Refitted Coefficient
Baseline	00885	C16:0.cholesteryl ester	4.45
	12097	C14:0.sphingomyelin	1.74
	04953	C24:1.Ceramide.(d18:1)	4.25
	00555	3-methyladipate/pimelate	-12.82
Week 1	06726	C20:4.cholesteryl ester	2.17
	12097	C14:0.sphingomyelin	2.06
	04949	C16:0.Ceramide.(d18:1)	0.87
	00555	3-methyladipate/pimelate	-6.10
	00056	beta-alanine	2.95
	00448	adipate	-4.50
Week 8	00883	valine	1.40
	00222	C16.carnitine	0.58
	00848	C18.carnitine	0.39
	00555	3-methyladipate/pimelate	-5.95
	00056	beta-alanine	0.63

tic regression models. At the FDR < 0.05, our procedure identifies 9 metabolites with different regression coefficients (see Table 5). It is interesting that all these 9 metabolites have the same signs of the refitted coefficients, while the actual magnitudes of the associations between responders and non-responders when compared to the normal controls are different. Besides C24:4.cholesteryl ester, beta-alanine, valine, C18.carnitine and 3-methyladipate/pimelate that we observe in previous analyses, metabolites 5-hydroxytryptopha, nicotinate, and succinate also have differential associations between responders and non-responders when compared to the controls.

7 DISCUSSION

In this paper, for both global and multiple testing, the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ of the covariates is assumed to be sparse and unknown. Node-wise regression among the covariates is used to learn the covariance structure in constructing the debiased estimator. However, if the prior knowledge of $\mathbf{\Omega} = \mathbf{I}$ is available, the algorithm can be simplified greatly. Specifically, instead of incorporating the Lasso estimators as in (2.7), we let $v_j = \hat{W}^{-1}\mathbf{x}_j$ and $\tau_j = \|v_j\|_n / \langle v_j, \mathbf{x}_j \rangle$ for each $j = 1, ..., p$. The theoretical properties of the resulting global testing and multiple testing procedures still hold, while the computational efficiency is improved dramatically. However, from our theoretical analysis, even with the knowledge of $\mathbf{\Omega} = \mathbf{I}$, the theoretical requirement for the model sparsity ($k = o(\sqrt{n}/\log^3 p)$ in the Gaussian case and $k = o(\sqrt{n}/\log^{5/2} p)$ in the bounded case) cannot be relaxed due to the nonlinearity of the problem.

Table 5: Significant metabolites identified via logistic regression of responder vs normal control and non-responder vs normal control for FDR $\leq 5\%$.

HMDB ID	Synonyms	Refitted Coefficients	
		Responder vs. Normal	Non-Responder vs. Normal
06726	C20:4.cholesteryl ester	0.139	1.854
01043	Linoleic.acid	-0.686	-0.388
00472	5-hydroxytryptophan	1.000	1.034
00056	beta-alanine	0.503	2.298
00883	valine	0.628	0.530
00848	C18.carnitine	1.100	0.457
01488	nicotinate	-1.936	-4.312
00254	succinate	0.750	1.508
00555	3-methyladipate/pimelate	-1.989	-4.209

Sample splitting was used in this paper for theoretical purpose. This is different from other works on inference in high-dimensional linear/logistic regression models, including Ingster et al. (2010), van de Geer et al. (2014), Mukherjee et al. (2015) and Javanmard and Javadi (2019), where sample splitting is not needed. However, as we discussed throughout the paper, the assumptions and the alternatives that we considered are different from those previous papers. In the case of high-dimensional logistic regression model, a sample splitting procedure seems unavoidable under the current framework of our technical analysis without making additional strong structural assumptions such as the sparse inverse Hessian matrices used in van de Geer et al. (2014) or the weakly correlated design matrices used in Mukherjee et al. (2015). Our simulations showed that the sample splitting is actually not needed in order for our proposed methods to perform well. It is of interest to develop technical tools that can eliminate sample splitting in inference for high dimensional logistic regression models.

As mentioned in the introduction, the logistic regression model can be viewed as a special case of the single index model $y = f(\beta^\top x) + \epsilon$ where f is a known transformation function (Yang et al., 2015). Based on our analysis, it is clear that the theoretical results are not limited to the sigmoid transfer function. In fact, the proposed methods can be applied to a wide range of transformation functions satisfying the following conditions: (C1) f is continuous and for any $u \in \mathbb{R}$, $0 < f(u) < 1$; (C2) for any $u_1, u_2 \in \mathbb{R}$, there exists a constant $L > 0$ such that $|\dot{f}(u_1) - \dot{f}(u_2)| \leq L|u_1 - u_2|$; and (C3) for any constant $C > 0$, there exists $\delta > 0$ such that for any $|u| \leq C$, $\dot{f}(u) \geq \delta$. Examples include but are not limited to the following function classes

- *Cumulative density functions:* $f(x) = P(X \leq x)$ for some continuous random variable X supported on $\mathbb{R}$. In particular, when $X \sim N(0,1)$, the resulting model becomes the probit regression.

- *Affine hyperbolic tangent functions:* $f(x) = \frac{1}{2}\tanh(ax+b) + \frac{1}{2}$ for some parameter $a, b \in \mathbb{R}$. In particular, $(a,b) = (1,0)$ corresponds to $f(x) = e^x/(1+e^x)$.

- *Generalized logistic functions:* $f(x) = (1+e^{-x})^{-\alpha}$ for some $\alpha > 0$.

Besides the problems we considered in this paper, it is also of interest to construct confidence intervals for functionals of the regression coefficients, such as $\|\beta\|_1$, $\|\beta\|_2$, or $\theta^\top\beta$ for some given loading vector θ. In modern statistical machine learning, logistic regression is considered as an efficient classification method (Abramovich and Grinshtein, 2018). In practice, a predicted label with an uncertainty assessment is usually preferred. Therefore, another important problem is the construction of predictive intervals of the conditional probability π^* associated with a given predictor X^*. These problems are related to the current work and are left for future investigations.

8 PROOFS OF THE MAIN THEOREMS

In this section, we prove Theorems 1, Theorem 2 and Theorem 4 in the paper. The proofs of other results, including Theorems 3 and 5, Proposition 1 and the technical lemmas, are given in our Supplementary Materials.

Proof of Theorem 1 Define $F_{jj} = \mathbb{E}[\eta_{ij}^2/\dot{f}(u_i)]$. Under H_0, $F_{jj} = 4\mathbb{E}[\eta_{ij}^2] = 4/\omega_{jj}$, and by (A3), $c < F_{jj} < C$ for $j = 1, ..., p$ and some constant $C \geq c > 0$. Define statistics

$$\tilde{M}_j = \frac{\langle v_j, \epsilon\rangle}{\|v_j\|_n}, \quad \text{and} \quad \check{M}_j = \frac{\sum_{i=1}^n \eta_{ij}\epsilon_i/\dot{f}(u_i)}{\sqrt{nF_{jj}}}, \quad j = 1, ..., p.$$

and $\tilde{M}_n = \max_j \tilde{M}_j^2, \check{M}_n = \max_j \check{M}_j^2$. The following lemma shows that $\tilde{M}_n$ and therefore $\check{M}_n$ are good approximations of M_n.

Lemma 1. *Under the condition of Theorem 1, the following events*

$$B_1 = \left\{|\tilde{M}_n - \check{M}_n| = o(1)\right\}, \qquad B_2 = \left\{|\tilde{M}_n - M_n| = o\left(\frac{1}{\log p}\right)\right\},$$

hold with probability at least $1 - O(p^{-c})$ for some constant $c > 0$.

It follows that under the event $B_1 \cap B_2$, let $y_p = 2\log p - \log\log p + x$ and $\epsilon_n = o(1)$, we have

$$P_\theta\big(\check{M}_n \leq y_p - \epsilon_n\big) \leq P_\theta\big(M_n \leq y_p\big) \leq P_\theta\big(\check{M}_n \leq y_p + \epsilon_n\big)$$

Therefore it suffices to prove that for any $t \in \mathbb{R}$, as $(n,p) \to \infty$,

$$P_\theta\big(\check{M}_n \leq y_p\big) \to \exp\left(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\right). \tag{8.1}$$

Now define $\hat{M}_j = \frac{\sum_{i=1}^n \hat{Z}_{ij}}{\sqrt{nF_{jj}}}, j = 1, ..., p.$ where $\hat{Z}_{ij} = v_{ij}^0 \epsilon_i 1\{|v_{ij}^0 \epsilon_i| \le \tau_n\} - \mathbb{E}[v_{ij}^0 \epsilon_i 1\{|v_{ij}^0 \epsilon_i| \le \tau_n\}]$ for $\tau_n = \log(p+n)$, $v_{ij}^0 = \eta_{ij}/\dot{f}(u_i)$ and $\hat{M}_n = \max_j \hat{M}_j^2$. The following lemma states that $\hat{M}_n$ is close to $\check{M}_n$.

Lemma 2. *Under the condition of Theorem 1, $|\check{M}_n - \hat{M}_n| = o(1)$ with probability at least $1 - O(p^{-c})$ for some constant $c > 0$.*

By Lemma 2, it suffices to prove that for any $t \in \mathbb{R}$, as $(n, p) \to \infty$,

$$P_\theta\big(\hat{M}_n \le y_p\big) \to \exp\bigg(-\frac{1}{\sqrt{\pi}} \exp(-x/2)\bigg). \tag{8.2}$$

To prove this, we need the classical Bonferroni inequality.

Lemma 3. (Bonferroni inequality) *Let $B = \cup_{t=1}^p B_t$. For any integer $k < p/2$, we have*

$$\sum_{t=1}^{2k} (-1)^{t-1} A_t \le P(B) \le \sum_{t=1}^{2k-1} (-1)^{t-1} A_t, \tag{8.3}$$

where $A_t = \sum_{1 \le i_1 < ... < i_t \le p} P(B_{i_1} \cap ... \cap B_{i_t})$.

By Lemma 3, for any integer $0 < q < p/2$,

$$\begin{aligned}\sum_{d=1}^{2q} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\bigg(\bigcap_{k=1}^d A_{j_k}\bigg) &\le P_\theta\bigg(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p\bigg) \\ &\le \sum_{d=1}^{2p-1} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\bigg(\bigcap_{k=1}^d A_{j_k}\bigg),\end{aligned} \tag{8.4}$$

where $A_{j_k} = \{\hat{M}_{j_k}^2 \ge y_p\}$. Now let $w_{i_j} = \hat{Z}_{ij}/\sqrt{F_{jj}}$ for $j = 1, ..., p$, and $\mathbf{W}_i = (w_{i,j_1}, ..., w_{i,j_d})^\top$ for $1 \le i \le n$. Define $\|\mathbf{a}\|_{\min} = \min_{1 \le i \le d} |a_i|$ for any vector $\mathbf{a} \in \mathbb{R}^d$. Then we have

$$P_\theta\bigg(\bigcap_{k=1}^d A_{j_k}\bigg) = P_\theta\bigg(\bigg\|n^{-1/2} \sum_{i=1}^n \mathbf{W}_i\bigg\|_{\min} \ge y_p^{1/2}\bigg).$$

Then it follows from Theorem 1.1 in Zaitsev (1987) that

$$\begin{aligned}P_\theta\bigg(\bigg\|n^{-1/2} \sum_{i=1}^n \mathbf{W}_i\bigg\|_{\min} \ge y_p^{1/2}\bigg) &\le P_\theta\bigg(\|\mathbf{N}_d\|_{\min} \ge y_p^{1/2} - \epsilon_n (\log p)^{-1/2}\bigg) \\ &\quad + c_1 d^{5/2} \exp\bigg\{-\frac{n^{1/2}\epsilon_n}{c_2 d^3 \tau_n (\log p)^{1/2}}\bigg\},\end{aligned} \tag{8.5}$$

where $c_1 > 0$ and $c_2 > 0$ are constants, $\epsilon_n \to 0$ which will be specified later, and $\mathbf{N}_d = (N_{m_1}, ..., N_{m_d})$ is a normal random vector with $\mathbb{E}(\mathbf{N}_d) = 0$ and $\text{cov}(\mathbf{N}_d) = \text{cov}(\mathbf{W}_1)$. Here d is a fixed integer that does not depend on n, p. Because $\log p = o(n^{1/5})$, we can let $\epsilon_n \to 0$ sufficiently slow, say, $\epsilon_n = \sqrt{\log^5 p/n}$, so that for any large $c > 0$,

$$c_1 d^{5/2} \exp\left\{ - \frac{n^{1/2}\epsilon_n}{c_2 d^3 \tau_n (\log p)^{1/2}} \right\} = O(p^{-c}). \tag{8.6}$$

Combining (A.8), (A.9) and (A.10), we have

$$P_\theta\bigg(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p \bigg) \le \sum_{d=1}^{2p-1} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\bigg(\|\mathbf{N}_d\|_{\min} \ge y_p^{1/2} - \epsilon_n (\log p)^{-1/2} \bigg) + o(1). \tag{8.7}$$

Similarly, one can derive

$$P_\theta\bigg(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p \bigg) \ge \sum_{d=1}^{2p} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\bigg(\|\mathbf{N}_d\|_{\min} \ge y_p^{1/2} + \epsilon_n (\log p)^{-1/2} \bigg) + o(1). \tag{8.8}$$

Now we use the following lemma from Xia et al. (2018).

Lemma 4. *For any fixed integer $d \ge 1$ and real number $t \in \mathbb{R}$,*

$$\sum_{1 \le j_1 < ... < j_d \le p} P_\theta\bigg(\|\mathbf{N}_d\|_{\min} \ge y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2} \bigg) = \frac{1}{d!} \bigg(\frac{1}{\sqrt{\pi}} \exp(-t/2) \bigg)^d (1 + o(1)).$$

It then follows from the above lemma, (A.11) and (A.12) that

$$\begin{aligned}
\overline{\lim_{n,p\to\infty}} P_\theta\bigg(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p \bigg) &\le \sum_{d=1}^{2p} (-1)^{d-1} \frac{1}{d!} \bigg(\frac{1}{\sqrt{\pi}} \exp(-t/2) \bigg)^d, \\
\underline{\lim_{n,p\to\infty}} P_\theta\bigg(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p \bigg) &\ge \sum_{d=1}^{2p-1} (-1)^{d-1} \frac{1}{d!} \bigg(\frac{1}{\sqrt{\pi}} \exp(-t/2) \bigg)^d,
\end{aligned}$$

for any positive integer p. By letting $p \to \infty$, we obtain (A.7) and the proof is complete. □

Proof of Theorem 2. The proof essentially follows from the general Le Cam's method described in Section 7.1 of Baraud (2002). The key elements can be summarized as the following lemma that reduces the lower bound problem to calculation of the total variation distance between two posterior distributions.

Lemma 5. *Let $\mathcal{H}_1$ be some subset in an ℓ_2 bounded Hilbert space and ρ some positive number. Let μ_ρ be some probability measure on $\mathcal{H}_1 = \{\theta \in \Theta, \|\theta\| = \rho\}$. Set $P_{\mu_\rho} = \int P_\theta d\mu_\rho(\theta)$, P_0 as the (posterior) distribution at the null, and denote by Φ_α the level-α tests, we have*

$$\inf_{\Phi_\alpha} \sup_{\theta \in \mathcal{H}_1} P_\theta(\Phi_\alpha = 0) \geq \inf_{\Phi_\alpha} P_{\mu_\rho}(\Phi_\alpha = 0) \geq 1 - \alpha - TV(P_{\mu_\rho}, P_0),$$

where $TV(P_{\mu_\rho}, P_0)$ denotes the total variation distance between P_{μ_ρ} and P_0.

Now since by definition $\rho^*(\Phi_\alpha, \delta, \Theta(k)) \geq \rho^*(\Phi_\alpha, \delta, \mathbf{\Sigma})$ for any $\mathbf{\Sigma} \in \Theta_2(k)$, by Lemma 5, it suffices to construct the corresponding $\mathcal{H}_1$ for $\beta \in \Theta_\beta(k)$ and find a lower bound $\rho_1 = \rho(\eta)$ such that

$$\forall \rho \leq \rho_1 \qquad \inf_{\Phi_\alpha} P_{\mu_\rho}(\Phi_\alpha = 0) \geq 1 - \alpha - \eta = \delta. \tag{8.9}$$

for fixed covariance $\mathbf{\Sigma} = \mathbf{I}$. In this case, an upper bound for the χ^2-divergence between P_{μ_ρ} and P_0, defined as $\chi^2(P_{\mu_\rho}, P_0) = \int \frac{(dP_{\mu_\rho})^2}{dP_0} - 1$, can be obtained by carefully constructing the alternative space $\mathcal{H}_1$. Since $TV(f, g) \leq \sqrt{\chi^2(f, g)}$ (see p.90 of Tsybakov (2009)), it follows that $\inf_{\Phi_\alpha} P_{\mu_\rho}(\Phi_\alpha = 0) \geq 1 - \alpha - \sqrt{\chi^2(P_{\mu_\rho}, P_0)}$. By choosing $\rho_1 = \rho(\eta)$ such that for any $\rho \leq \rho_1$, $\chi^2(P_{\mu_\rho}, P_0) \leq \eta^2 = (1 - \alpha - \delta)^2$, we have (8.9) holds. In the following, we will construct the alternative space $\mathcal{H}_1$ and derive an upper bound of $\chi^2(P_{\mu_\rho}, P_0)$ where P_0 corresponds to the null space $\mathcal{H}_0$ defined at a single point $\beta = 0$. We divide the proofs into two parts. Throughout, the design covariance matrix is chosen as $\mathbf{\Sigma} = \mathbf{I}$.

Step 1: Construction of $\mathcal{H}_1$. Firstly, for a set M, we define $\ell(M, n)$ as the set of all the n-element subsets of M. Let $[1 : p] \equiv \{1, ..., p\}$, so $\ell([1 : p], k)$ contains all the k-element subsets of $[1 : p]$. We define the alternative parameter space $\mathcal{H}_1 = \big\{\beta \in \mathbb{R}^p : \beta_j = \rho 1\{j \in I\}$ for $I \in \ell([1 : p], k)\big\}$. In other words, $\mathcal{H}_1$ contains all the k-sparse vectors $\beta(I)$ whose nonzero components ρ are indexed by I. Apparently, for any $\beta \in \mathcal{H}_1$, it follows $\|\beta\|_\infty = \rho$ and $\mathcal{H}_1 \subseteq \Theta_1(k)$.

Step 2: Control of $\chi^2(P_{\pi_{\mathcal{H}_1}}, P_0)$. Let π denote the uniform prior of the random index set I over $\ell([1 : p], k)$. This prior induces a prior distribution $\pi_{\mathcal{H}_1}$ over the parameter space $\mathcal{H}_1$. For $\{\mathbf{0}_p\} = \mathcal{H}_0$, the corresponding joint distribution of the data $\{(X_i, y_i)\}_{i=1}^n$ is

$$f = \prod_{i=1}^n p(X_i, y_i) = \frac{1}{(2\pi)^{np/2}} \prod_{i=1}^n \frac{1}{2} e^{-\|X_i\|_2^2/2}.$$

Similarly, the posterior distribution of the samples over the prior $\pi_{\mathcal{H}_1}$ is denoted as

$$g = \prod_{i=1}^n \int_{\mathcal{H}_1} p(X_i, y_i; \beta) \pi_{\mathcal{H}_1} = \frac{1}{\binom{p}{k}} \sum_{\beta \in \mathcal{H}_1} \prod_{i=1}^n p(X_i, y_i; \beta).$$

As a result, we have the following lemma controlling $\chi^2(P_{\pi_{\mathcal{H}_1}}, P_0) = \chi^2(g, f)$.

Lemma 6. *Let $\rho^2 = \frac{1}{n}\log\big(1 + \frac{p}{h(\eta)k^2}\big)$ where $h(\eta) = [\log(\eta^2+1)]^{-1}$ and $\eta = 1-\alpha-\delta$, then we have $\chi^2(g,f) \le (1-\alpha-\delta)^2$.*

Combining Lemma 5 and Lemma 6, we know that for $\alpha, \delta > 0$ and $\alpha + \delta < 1$, if $\rho = \sqrt{\frac{1}{n}\log\big(1+\frac{p}{h(\eta)k^2}\big)}$, then $\forall \rho' \le \rho, \inf_{\Phi_\alpha}\sup_{\beta\in\Theta(k):\|\beta\|_\infty \ge \rho'} P_\theta(\Phi_\alpha = 0) \ge \delta$. Therefore, it follows that

$$\rho^*(\alpha,\delta,\Theta(k)) \ge \rho^*(\alpha,\delta,\mathbf{I}) \gtrsim \sqrt{\frac{1}{n}\log\left(1+\frac{p}{k^2}\right)}. \tag{8.10}$$

Lastly, note that for the above chosen ρ, $\mathcal{H}_1 \subset \Theta_1(k) \cap \{\beta \in \mathbb{R}^p : \|\beta\|_2 \lesssim (n^{1/4}\log p)^{-1}\}$ when $k \lesssim \min\{p^\gamma, \sqrt{n}/\log^3 p\}$ for some $0 < \gamma < 1/2$. This completes the proof. $\square$

Proof of Theorem 4. The proof follows similar arguments of the proof of Theorem 3.1 in Javanmard and Javadi (2019). We first consider the case when $\hat{t}$, given by (3.2), does not exist. In this case, $\hat{t} = \sqrt{2\log p}$ and we consider the event $\Omega_0 = \{\sum_{j\in\mathcal{H}_0} I\big(|M_j| \ge \sqrt{2\log p}\big) \ge 1\}$ that there are at least one false positive. In order to show the FDR/FDP can be controlled in this case, we show that

$$P_\theta(\Omega_0) \to 0, \quad \text{as } (n,p)\to\infty. \tag{8.11}$$

Note that for $j \in \mathcal{H}_0$, we have $M_j = \frac{\check{\beta}_j}{\tau_j} = \frac{\langle v_j, \epsilon\rangle}{\|v_j\|_n} + \frac{\langle v_j, Re\rangle}{\|v_j\|_n} - \frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n}$. Then

$$\begin{aligned} P_\theta(\Omega_0) \le P_\theta\bigg(&\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle v_j,\epsilon\rangle}{\|v_j\|_n} + \frac{\langle v_j, Re\rangle}{\|v_j\|_n} - \frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} \ge \sqrt{2\log p}\bigg) \ge 1\bigg) \\ &+ P_\theta\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle v_j,\epsilon\rangle}{\|v_j\|_n} + \frac{\langle v_j, Re\rangle}{\|v_j\|_n} - \frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} \le -\sqrt{2\log p}\bigg) \ge 1\bigg). \end{aligned} \tag{8.12}$$

For any $\epsilon > 0$, we can bound the first term by

$$\begin{aligned} &P_\theta\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle v_j,\epsilon\rangle}{\|v_j\|_n} + \frac{\langle v_j, Re\rangle}{\|v_j\|_n} - \frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} \ge \sqrt{2\log p}\bigg) \ge 1\bigg) \\ &= P_\theta\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\tilde{M}_j \ge \sqrt{2\log p} + \frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} - \frac{\langle v_j, Re\rangle}{\|v_j\|_n}\bigg) \ge 1\bigg) \\ &\le P_\theta\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\tilde{M}_j \ge \sqrt{2\log p} - \epsilon\bigg) \ge 1\bigg) + P_\theta\bigg(\max_{j\in\mathcal{H}_0}\bigg|\frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} - \frac{\langle v_j, Re\rangle}{\|v_j\|_n}\bigg| \ge \epsilon\bigg) \\ &\le p\max_{j\in\mathcal{H}_0} P_\theta\bigg(\tilde{M}_j \ge \sqrt{2\log p} - \epsilon\bigg) + P_\theta\bigg(\max_{j\in\mathcal{H}_0}\bigg|\frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} - \frac{\langle v_j, Re\rangle}{\|v_j\|_n}\bigg| \ge \epsilon\bigg) \end{aligned}$$

By the proof of Lemma 1, we know that $P_\theta\big(\max_{j\in\mathcal{H}_0}\big|\frac{\langle v_j, \mathbf{h_{-j}}\rangle_\mathbf{n}}{\|v_j\|_n} - \frac{\langle v_j, Re\rangle}{\|v_j\|_n}\big| \ge \epsilon\big) \to 0$. In addition, for $j \in \mathcal{H}_0$, $P_\theta\big(\tilde{M}_j \ge \sqrt{2\log p} - \epsilon\big) \le P_\theta\big(\check{M}_j \ge \sqrt{2\log p} - 2\epsilon\big) + P_\theta(|\tilde{M}_j - \check{M}_j| \ge \epsilon)$, where $\max_{j\in\mathcal{H}_0} P_\theta(|\tilde{M}_j - \check{M}_j| \geq \epsilon) = O(p^{-c})$ for some sufficiently large $c > 0$. Now since $\check{M}_j = \frac{\sum_{i=1}^n \eta_{ij}\epsilon_i/\dot{f}(u_i)}{\sqrt{nF_{jj}}}$ where $\mathbb{E}\frac{\eta_{ij}\epsilon_i/\dot{f}(u_i)}{\sqrt{F_{jj}}} = 0$ and $\mathrm{Var}(\frac{\eta_{ij}\epsilon_i/\dot{f}(u_i)}{\sqrt{F_{jj}}}) = 1$, by Lemma 6.1 of Liu (2013), we have $\sup_{0\leq t\leq 4\sqrt{\log p}}\left|\frac{P_\theta(|\check{M}_j|\geq t)}{G(t)} - 1\right| \leq C(\log p)^{-1}$. Now let $t = \sqrt{2\log p} - 2\epsilon$, we have

$$P_\theta\bigg(\check{M}_j \geq \sqrt{2\log p} - 2\epsilon\bigg) \leq G(\sqrt{2\log p} - 2\epsilon) + C\frac{G(\sqrt{2\log p} - 2\epsilon)}{\log p}.$$

Hence $p\max_{j\in\mathcal{H}_0} P_\theta\bigg(\tilde{M}_j \geq \sqrt{2\log p} - \epsilon\bigg) \leq CpG(\sqrt{2\log p} - 2\epsilon) + O(p^{-c})$, which goes to zero as $(n,p)\to\infty$. By symmetry, we know that the second term in (8.12) also goes to 0. Therefore we have proved (8.11).

Now consider the case when $0 \leq \hat{t} \leq b_p$ holds. We have

$$\mathrm{FDP}_\theta(\hat{t}) = \frac{\sum_{j\in\mathcal{H}_0} I\{|M_j| \geq \hat{t}\}}{\max\big\{\sum_{j=1}^p I\{|M_j| \geq \hat{t}\}, 1\big\}} \leq \frac{p_0 G(\hat{t})}{\max\big\{\sum_{j=1}^p I\{|M_j| \geq \hat{t}\}, 1\big\}}(1 + A_p),$$

where $A_p = \sup_{0\leq t\leq b_p}\left|\frac{\sum_{j\in\mathcal{H}_0} I\{|M_j|\geq t\}}{p_0 G(t)} - 1\right|$. Note that by definition $\frac{p_0 G(\hat{t})}{\max\big\{\sum_{j=1}^p I\{|M_j|\geq\hat{t}\},1\big\}} \leq \frac{p_0\alpha}{p}$. The proof is complete if $A_p \to 0$ in probability, which has been shown by Proposition 1. □

FUNDING

This research was supported by NIH grants R01CA127334 and R01GM123056.

SUPPLEMENTARY MATERIALS

In the online Supplemental Materials, we prove Theorem 3, 5, Proposition 1, and the technical lemmas. The technical results and simulations concerning the two-sample tests discussed in Section 4 are also included.

References

Abramovich, F. and V. Grinshtein (2018). High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory 65*(5), 3068–3079.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII 1983*, pp. 1–198. Springer.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics 4*, 384–414.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli 8*(5), 577–606.

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature Genetics 40*(8), 955.

Belloni, A., V. Chernozhukov, and Y. Wei (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics 34*(4), 606–619.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics 45*(2), 615–646.

Cai, T. T., H. Li, J. Ma, and Y. Xia (2017). Differential markov random field analysis with applications to detecting differential microbial community structures. *Unpublished Manuscript*.

Cai, T. T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*(501), 265–277.

Cai, T. T., W. Liu, and Y. Xia (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(2), 349–372.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3), 551–577.

Candès, E. J. and P. Sur (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*.

Fortin, G. (2011). L-carnitine and intestinal inflammation. In *Vitamins & Hormones*, Volume 86, pp. 353–366. Elsevier.

Gradshteyn, I. S. and I. M. Ryzhik (2014). *Table of Integrals, Series, and Products*. Academic press.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods of Statiststics 2*(2), 85–114.

Ingster, Y. I., A. B. Tsybakov, and N. Verzelen (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics 4*, 1476–1526.

Javanmard, A. and H. Javadi (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics 13*(1), 1212–1253.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Kolokoltsov, V. N. (2011). *Markov Processes, Semigroups, and Generators*, Volume 38. Walter de Gruyter.

Lee, D., R. N. Baldassano, A. R. Otley, L. Albenberg, A. M. Griffiths, C. Compher, E. Z. Chen, H. Li, E. Gilroy, L. Nessel, et al. (2015). Comparative effectiveness of nutritional and biological therapy in north american children with active crohn's disease. *Inflammatory Bowel Diseases 21*(8), 1786–1793.

Lewis, J. D., E. Z. Chen, R. N. Baldassano, A. R. Otley, A. M. Griffiths, D. Lee, K. Bittinger, A. Bailey, E. S. Friedman, C. Hoffmann, et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease. *Cell Host & Microbe 18*(4), 489–500.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics 41*(6), 2948–2978.

Liu, W. and S. Luo (2014). Hypothesis testing for high-dimensional regression models. Technical report.

Maceyka, M. and S. Spiegel (2014). Sphingolipid metabolites in inflammatory disease. *Nature 510*(7503), 58.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 53–71.

Mukherjee, R., N. S. Pillai, and X. Lin (2015). Hypothesis testing for high-dimensional sparse binary regression. *Annals of Statistics 43*(1), 352–381.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2010). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Technical Report Number 979*.

Ni, J., T.-C. D. Shen, E. Z. Chen, K. Bittinger, A. Bailey, M. Roggiani, A. Sirota-Madi, E. S. Friedman, L. Chau, A. Lin, et al. (2017). A role for bacterial urease in gut dysbiosis and crohn's disease. *Science Translational Medicine 9*(416), eaah6888.

Peltekova, V. D., R. F. Wintle, L. A. Rubin, C. I. Amos, Q. Huang, X. Gu, B. Newman, M. Van Oene, D. Cescon, G. Greenberg, et al. (2004). Functional variants of octn cation transporter genes are associated with crohn's disease. *Nature Genetics 36*(5), 471.

Plan, Y. and R. Vershynin (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory 59*(1), 482–494.

Ren, Z., C.-H. Zhang, and H. H. Zhou (2016). Asymptotic normality in estimation of large ising graphical model. *Unpublished Manuscript*.

Rozenova, K. A., G. M. Deevska, A. A. Karakashian, and M. N. Nikolova-Karakashian (2010). Studies on the role of acid sphingomyelinase and ceramide in the regulation of tumor necrosis factor α (tnfα)-converting enzyme activity and tnfα secretion in macrophages. *Journal of Biological Chemistry 285*(27), 21103–21113.

Sewell, G. W., Y. A. Hannun, X. Han, G. Koster, J. Bielawski, V. Goss, P. J. Smith, F. Z. Rahman, R. Vega, S. L. Bloom, et al. (2012). Lipidomic profiling in crohn's disease: abnormalities in phosphatidylinositols, with preservation of ceramide, phosphatidylcholine and phosphatidylserine composition. *The International Journal of Biochemistry & Cell Biology 44*(11), 1839–1846.

Shekhawat, P. S., S. Sonne, A. L. Carter, D. Matern, and V. Ganapathy (2013). Enzymes involved in l-carnitine biosynthesis are expressed by small intestinal enterocytes in mice: Implications for gut health. *Journal of Crohn's & Colitis 7*, e197–e205.

Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*(29), 14516–14525.

Sur, P., Y. Chen, and E. J. Candès (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 1–72.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*(2), 614–645.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics 6*, 38–90.

Xia, Y., T. Cai, and T. T. Cai (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika 102*(2), 247–266.

Xia, Y., T. Cai, and T. T. Cai (2018). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statistica Sinica 28*, 63–92.

Xia, Y., T. T. Cai, and H. Li (2018). Joint testing and false discovery rate control in high-dimensional multivariate regression. *Biometrika 105*(2), 249–269.

Yang, Z., Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang (2015). Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv preprint arXiv:1511.04514*.

Zaitsev, A. Y. (1987). On the gaussian approximation of convolutions under multidimensional analogues of s.n. bernstein's inequality conditions. *Probability Theory and Related Fields 74*(4), 535–566.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.

Supplement to "Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models"

Rong Ma[1], T. Tony Cai[2] and Hongzhe Li[1]
Department of Biostatistics, Epidemiology and Informatics[1]
Department of Statistics[2]
University of Pennsylvania
Philadelphia, PA 19104

Abstract

In this Supplementary Material we prove Theorem 3, 5 and Proposition 1 in the main paper and the technical lemmas. The technical and simulation results of the two-sample tests discussed in Section 4 of the main paper are included in the appendix.

1 Proofs of Main Results

1.1 Proof of Proposition 1

By similar argument as in Lemma 1, we can prove the following lemma.

Lemma 7. *Assume (A2) (A3) and (A4), $k = o(\sqrt{n}/\log^{5/2} p)$, then*

$$\max_{j\in\mathcal{H}_0} |\tilde{M}_j - \check{M}_j| = o\left(\frac{1}{\sqrt{\log p}}\right), \quad \max_{j\in\mathcal{H}_0} |\tilde{M}_j - M_j| = o\left(\frac{1}{\sqrt{\log p}}\right),$$

hold with probability at least $1 - O(p^{-c})$ for some constant $c > 0$.

For (20), by Lemma 6.1 in Liu (2013), we have

$$\max_{1\le j\le p} \sup_{0\le t\le 4\sqrt{\log p}} \left| \frac{P_\theta(|\check{M}_j| \ge t)}{G(t)} - 1 \right| \le C(\log p)^{-2-\gamma_1} \tag{1.1}$$

for some constant $0 < \gamma_1 < 1/2$. So (20) follows from Lemma 7, and the fact that $G(t + o(1/\sqrt{\log p}))/G(t) = 1 + o(1)$ uniformly in $0 \le t \le \sqrt{2\log p}$.

For (21), it suffices to show that

$$\sup_{0\le t\le b_p}\left|\frac{\sum_{j\in\mathcal{H}_0} I\{|\check{M}_j|\ge t\}}{p_0 G(t)}-1\right|\to 0 \quad \text{in probability.} \tag{1.2}$$

Let $z_0 < z_1 < \ldots < z_{d_p} \le 1$ and $t_i = G^{-1}(z_i)$, where $z_0 = G(b_p)$, $z_i = c_p/p + c_p^{2/3}e^{i^\delta}/p$ with $c_p = pG(b_p)$, and $d_p = [\log((p-c_p)/c_p^{2/3})]^{1/\delta}$ and $0<\delta<1$, which will be specified later. We have $G(t_i)/G(t_{i+1}) = 1+o(1)$ uniformly in i, and $t_0/\sqrt{2\log(p/c_p)} = 1+o(1)$. Note that uniformly for $1\le j\le m$, $G(t_i)/G(t_{i-1})\to 1$ as $p\to\infty$. The proof of (1.2) reduces to show that

$$\max_{0\le i\le d_p}\left|\frac{\sum_{j\in\mathcal{H}_0} I\{|\check{M}_j|\ge t_i\}}{p_0 G(t_i)}-1\right|\to 0 \tag{1.3}$$

in probability. In fact, for each $\epsilon>0$, we have

$$P_\theta\left(\max_{0\le i\le d_p}\left|\frac{\sum_{j\in\mathcal{H}_0}[I\{|\check{M}_j|\ge t_i\}-G(t_i)]}{p_0G(t_i)}\right|\ge\epsilon\right)\le\sum_{j=0}^{d_p}P_\theta\left(\left|\frac{\sum_{j\in\mathcal{H}_0}[I\{|\check{M}_j|\ge t_i\}-G(t_i)]}{p_0G(t_i)}\right|\ge\epsilon/2\right).$$

Set $I(t)=\frac{\sum_{j\in\mathcal{H}_0}[I\{|\check{M}_j|\ge t\}-P_\theta(|\check{M}_j|\ge t)]}{p_0G(t)}$. By Markov's inequality $P_\theta(|I(t_i)|\ge\epsilon/2)\le\frac{\mathbb{E}[I(t_i)]^2}{\epsilon^2/4}$, and it suffices to show $\sum_{j=0}^{d_p}\mathbb{E}[I(t_i)]^2=o(1)$. To see this, by (1.1),

$$\begin{aligned}
\mathbb{E}I^2(t) &= \frac{\sum_{j\in\mathcal{H}_0}[P_\theta(|\check{M}_j|\ge t)-P_\theta^2(|\check{M}_j|\ge t)]}{p_0^2G^2(t)} \\
&\quad+\frac{\sum_{j,k\in\mathcal{H}_0,k\ne j}[P_\theta(|\check{M}_j|\ge t,|\check{M}_k|\ge t)-P_\theta(|\check{M}_j|\ge t)P_\theta(|\check{M}_k|\ge t)]}{p_0^2G^2(t)} \\
&\le \frac{C}{p_0G(t)}+\frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon):j,k\in\mathcal{H}_0}\frac{P_\theta(|\check{M}_j|\ge t,|\check{M}_k|\ge t)}{G^2(t)} \\
&\quad+\frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon)^c:j,k\in\mathcal{H}_0}\left[\frac{P_\theta(|\check{M}_j|\ge t,|\check{M}_k|\ge t)}{G^2(t)}-1\right] \\
&= \frac{C}{p_0G(t)}+I_{11}(t)+I_{12}(t).
\end{aligned}$$

For $(j,k)\in\mathcal{A}(\epsilon)^c$ with $j,k\in\mathcal{H}_0$, applying Lemma 6.1 in Liu (2013), we have $I_{12}(t)\le C(\log p)^{-1-\xi}$ for some $\xi>0$ uniformly in $0<t<\sqrt{2\log p}$. By Lemma 6.2 in Liu (2013), for $(j,k)\in\mathcal{A}(\epsilon)$ with $j,k\in\mathcal{H}_0$, we have

$$P_\theta(|\check{M}_j|\ge t,|\check{M}_k|\ge t)\le C(t+1)^{-2}\exp\left(-\frac{t^2}{1+|\rho_{jk}|}\right).$$

So that

$$I_{11}(t) \le C\frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon):j,k\in\mathcal{H}_0}(t+1)^{-2}\exp\left(-\frac{t^2}{1+|\rho_{jk}|}\right)G^{-2}(t) \le C\frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon):j,k\in\mathcal{H}_0}[G(t)]^{-\frac{2|\rho_{jk}|}{1+|\rho_{jk}|}}.$$

Note that for $0 \le t \le b_p$, we have $G(t) \ge G(b_p) = c_p/p$, so that by assumption (A5) it follows that for some $\epsilon, q > 0$,

$$I_{11}(t) \le C\sum_{(j,k)\in\mathcal{A}(\epsilon):j,k\in\mathcal{H}_0} p^{\frac{2|\rho_{jk}|}{1+|\rho_{jk}|}+q-2} = O(1/(\log p)^2).$$

By the above inequalities, we can prove (1.3) by choosing $0 < \delta < 1$ so that

$$\begin{aligned}\sum_{i=0}^{d_p}\mathbb{E}[I(t_i)]^2 &\le C\sum_{i=0}^{d_p}(pG(t_i))^{-1} + Cd_p[(\log p)^{-1-\delta} + (\log p)^{-2}] \\ &\le C\sum_{i=0}^{d_p}\frac{1}{c_p + c_p^{2/3}e^{i^\delta}} + o(1) \\ &= o(1).\end{aligned}$$

□

1.2 Proof of Theorem 3

Define $M_j' = \tau_j^{-1}(\check{\beta}_j - \beta_j)$, and $M_n' = \max_j (M_j')^2$, we have $-\beta_j/\tau_j = M_j' - M_j$. Thus

$$\beta_j^2/\tau_j^2 \le 2(M_j')^2 + 2M_j^2, \quad \text{for all } j, \tag{1.4}$$

and

$$\max_j \beta_j^2/\tau_j^2 \le 2M_n' + 2M_n. \tag{1.5}$$

The main idea for proving Theorem 3 is that, in order to show that M_n is "large", we show that M_n' is "small" while $\max_j \beta_j^2/\tau_j^2$ is "large" under the condition of Theorem 3. In the following, we consider the Gaussian design and the bounded design separately. For the Gaussian design, we divide the proof into two parts.

Gaussian Design, Case 1. $\|\beta\|_2 \lesssim (\log p)^{-1/2}$**.** In this case, $\beta^\top X_i$ are i.i.d. $N(0, \beta^\top\mathbf{\Sigma}\beta)$. By Lemma 6 in Cai et al. (2014), we have

$$P_\theta\left(\max_{1\le i\le n}|\beta^\top X_i| \ge \|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log p}\right) = O(p^{-c}), \tag{1.6}$$

then (A4), or $P_\theta(\max_{1\le i\le n}|\beta^\top X_i|\le c)\to 1$ for some constant $c>0$, holds. Consequently, the following lemma can be established by similar arguments as the proof of Lemma 1.

Lemma 8. *Under the condition of Theorem 3, suppose (A4) hold, then*

$$P_\theta\big(|M_j'|\ge \sqrt{C_0\log p}\big)=O(p^{-c}) \tag{1.7}$$

for some constants $C_0, c>0$.

By Lemma 8, we have

$$P_\theta\big(M_n'\ge C_0\log p\big)=O(p^{-c}) \tag{1.8}$$

for some $C_0,c>0$. On the other hand, to bound τ_j, we start with the inequality

$$\frac{\|\hat\eta_j\|_2}{\langle\hat\eta_j,\mathbf{x}_j\rangle}\le\frac{C_2}{\sqrt{n}}$$

obtained as (2.10) in the proof of Lemma 1. By (A4), there exists some constant $0<\kappa<1$ such that $\kappa<|f(u_i)|<1-\kappa$ with high probability. Then it follows that

$$1-\dot f(\hat u_i)\le\xi\dot f(\hat u_i),\qquad \text{where } \xi_1=\frac{1-\kappa+\kappa^2}{\kappa(1-\kappa)}.$$

Thus, since

$$\|v_j\|_n-\|\hat\eta_j\|_2\le\sqrt{\sum_{i=1}^n(\dot f(\hat u_i)-\dot f^2(\hat u_i))v_{ij}^2}\le\sqrt{\xi_1\sum_{i=1}^n\dot f^2(\hat u_i)v_{ij}^2}=\xi_1^{1/2}\|\hat\eta_j\|_2,$$

we have, with probability at least $1-O(p^{-c})$,

$$\tau_j=\frac{\|v_j\|_n}{|\langle v_j,\mathbf{x}_j\rangle_n|}\le(1+\xi_1^{1/2})\frac{\|\hat\eta_j\|_2}{|\langle\hat\eta_j,\mathbf{x}_j\rangle|}\le C_2\frac{1+\xi_1^{1/2}}{\sqrt{n}}=\frac{C_3}{\sqrt{n}}, \tag{1.9}$$

for some constant $C_3>0$. Therefore, since $\|\beta\|_\infty\ge c_2\sqrt{\log p/n}$,

$$\max_j\beta_j^2/\tau_j^2\ge c_2^2\frac{\log p}{n}\cdot C_3^{-2}n=C_4\log p \tag{1.10}$$

with probability converging to 1. In particular, when c_2 is chosen such that the constant $C_4-2C_0\ge 4$, then under H_1, combining (1.5) (1.8) and (1.10), we have $P_\theta\big(\Phi_\alpha(M_n)=1\big)\to 1$ as $(n,p)\to\infty$.

Gaussian Design, Case 2. $\|\beta\|_2\gtrsim(\log p)^{-1/2}$. In this case, we have

$$\|\beta\|_\infty\ge\sqrt{\|\beta\|_2^2/k}\gtrsim(k\log p)^{-1/2}. \tag{1.11}$$

By (1.6), with probability at least $1 - O(n^{-c})$,

$$\min_{1\le i\le n} \dot{f}(u_i) \ge \frac{\exp(\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n})}{(1+\exp(\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n}))^2} \ge \frac{1}{4e^{\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n}}}. \tag{1.12}$$

Let

$$L(n) = e^{-\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n}}/4,$$

it follows that with probability at least $1 - O(n^{-c})$,

$$1 - \dot{f}(\hat{u}_i) \le \xi_2 \dot{f}(\hat{u}_i), \qquad \text{where } \xi_2 = \frac{1-L(n)}{L(n)}.$$

Thus, with probability at least $1 - O(n^{-c})$

$$\tau_j = \frac{\|v_j\|_n}{|\langle v_j, \mathbf{x}_j\rangle_n|} \le (1+\xi_2^{1/2})\frac{\|\hat{\eta}_j\|_2}{|\langle\hat{\eta}_j, \mathbf{x}_j\rangle|} \le C_2\frac{1+\xi_2^{1/2}}{\sqrt{n}} \le \frac{C_3 e^{\|\beta\|_2\sqrt{0.5\lambda_{\max}(\mathbf{\Sigma})\log n}}}{\sqrt{n}}, \tag{1.13}$$

for some constant $C_3 > 0$. Therefore, for $j = \arg\max|\beta_j|$, plug in (1.11) and $k = o(\sqrt{n}/\log^3 p)$, we have

$$\beta_j^2/\tau_j^2 \gtrsim \frac{n}{k\log p}e^{-\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n}} \ge C_4\sqrt{n}\log^2 p e^{-\|\beta\|_2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})\log n}} \tag{1.14}$$

with probability at least $1-O(n^{-c})$. Observe that as long as $\|\beta\|_2 \le C'\sqrt{\log n}$ for $C' = (2\sqrt{2\lambda_{\max}(\mathbf{\Sigma})})^{-1}$ (which is true since by assumption $\log\log p \lesssim r\log n$ and $\|\beta\|_2 \le C\log\log p/\sqrt{\log n}$ for some $C \le (2r\sqrt{2\lambda_{\max}(\mathbf{\Sigma})})^{-1}$), we have

$$\beta_j^2/\tau_j^2 \ge C_4 \log^2 p \tag{1.15}$$

with probability at least $1 - O(n^{-c})$.

Now we show that for the same $j = \arg\max|\beta_j|$,

$$P_\theta\big((M_j')^2 \ge C_0\log p\big) = O(n^{-c}) \tag{1.16}$$

for some $C_0, c > 0$. This can be established by the following lemma.

Lemma 9. *Under the condition of Theorem 3, if $\|\beta\|_2 \gtrsim (\log p)^{-1/2}$, then for any $j = 1, ..., p$,*

$$P_\theta\big(M_j' \ge C_1\sqrt{\log p}\big) = O(n^{-c}) \tag{1.17}$$

for some constants $C_1, c > 0$.

Therefore, by (1.4) (1.15) and (1.16), we have

$$M_n \ge M_j^2 \ge \frac{1}{2}C_4\log^2 p - C_0\log p$$

with probability at least $1 - O(n^{-c})$. Thus $P_\theta\big(\Phi_\alpha(M_n) = 1\big) \to 1$ as $n \to \infty$.

Bounded Design. The proof under the bounded design follows the same argument as the Case 1 of the Gaussian design, thus is omitted. □

1.3 Proof of Theorem 5

By (20) in Proposition 1, let $t = \hat{t}_{FDV}$, it follows that as $(n,p) \to \infty$,

$$\sup_{j \in \mathcal{H}_0} \left| \frac{P_\theta(|M_j| \geq \hat{t}_{FDV})}{G(\hat{t}_{FDV})} - 1 \right| \to 0, \tag{1.18}$$

So that by noting that $G(\hat{t}_{FDV}) = r/p$, we have as $(n,p) \to \infty$,

$$\left| \frac{\sum_{j \in \mathcal{H}_0} P_\theta(|M_j| \geq \hat{t}_{FDV})}{r/p} - p_0 \right| \to 0, \tag{1.19}$$

which completes the proof of (23). To prove (24), it suffices to note that

$$\mathrm{FWER}_\theta(t) = P_\theta\bigg(\sum_{j \in \mathcal{H}_0} I(|M_j| \geq t) \geq 1 \bigg) = P_\theta\bigg(\bigcup_{j \in \mathcal{H}_0} \{|M_j| \geq t\} \bigg) \leq \sum_{j \in \mathcal{H}_0} P_\theta(|M_j| \geq t),$$

and the final result follows from (1.19). □

2 Proofs of Technical Lemmas

Proof of Lemma 1. We start with the following lemma. In general, we will prove Lemma 1 under more general conditions posed in this lemma.

Lemma 10. *If one of the following two conditions holds,*

(C1) under Gaussian design, assume (A1) (A3) hold, $k = o(\sqrt{n}/\log^3 p)$, *and* $\|X\beta\|_\infty \leq c_2$ *for some constant* $c_2 > 0$*;*

(C2) under the bounded design, assume (A2) (A3) (A4) hold, and $k = o(\sqrt{n}/\log^{5/2} p)$*,*

then

$$\max_{1 \leq j \leq p} \left| \frac{\|v_j\|_n}{\sqrt{n}} - F_{jj}^{1/2} \right| = o\bigg(\frac{1}{\log p} \bigg) \tag{2.1}$$

in probability.

Lemma 10 can be established by combining results from Lemma 11 and Lemma 12 below, which provide some high probability bounds under the Gaussian and the bounded design, respectively.

Lemma 11. *Under the Gaussian design, assume (A1) and (A3) hold, the following events*

$$
\begin{aligned}
A_0 &= \left\{ \|\hat{\beta} - \beta\|_1 = O\left(k\sqrt{\frac{\log p}{n}} \right) \right\}, \\
A_1 &= \left\{ \max_{1 \le j \le p} \frac{1}{n} \|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_2^2 = O\left(k \frac{\log p}{n} \right) \right\}, \\
A_2 &= \left\{ \max_{1 \le j \le p} \|\hat{\gamma}_j - \gamma_j\|_1 = O\left(k\sqrt{\frac{\log p}{n}} \right) \right\}, \\
A_3 &= \left\{ \max_{i,j} \left| \hat{\eta}_{ij} - \eta_{ij} \right| = O\left(\frac{k \log p}{\sqrt{n}} \right) \right\},
\end{aligned}
$$

hold with probability at least $1 - O(p^{-c})$ *for some constant* $c > 0$*. In addition, if* $\|X\beta\|_\infty \le c_1$ *for some constant* $c_1 > 0$ *and* $k = o(n)$*, the following events*

$$
\begin{aligned}
A_4 &= \left\{ \max_i \left| \frac{1}{\dot{f}(\hat{u}_i)} - \frac{1}{\dot{f}(u_i)} \right| = O\left(\frac{k \log p}{\sqrt{n}} \right) \right\}, \\
A_5 &= \left\{ \max_{1 \le j \le p} \left| \frac{\|v_j\|_n}{\sqrt{n}} - F_{jj}^{1/2} \right| = O\left(\frac{\sqrt{k \log p}}{n^{1/4}} \right) \right\},
\end{aligned}
$$

hold with probability at least $1 - O(p^{-c})$ *for some constant* $c > 0$*.*

In particular, in (C1) of Lemma 10, we assume that $k = o(\sqrt{n}/\log^3 p)$, so A_5 in Lemma 11 implies Lemma 10 under (C1). On the other hand, under the bounded design, we have the following lemma.

Lemma 12. *Under the bounded design, assume (A2) (A3) and (A4) hold,* $k = o(n/\log p)$*, then events* A_0, A_1, A_2 *(in Lemma 11) and*

$$
\begin{aligned}
A_3' &= \left\{ \max_{i,j} \left| \hat{\eta}_{ij} - \eta_{ij} \right| = O\left(k\sqrt{\frac{\log p}{n}} \right) \right\}, \\
A_4' &= \left\{ \max_i \left| \frac{1}{\dot{f}(\hat{u}_i)} - \frac{1}{\dot{f}(u_i)} \right| = O\left(k\sqrt{\frac{\log p}{n}} \right) \right\}, \\
A_5' &= \left\{ \max_{1 \le j \le p} \left| \frac{\|v_j\|_n}{\sqrt{n}} - F_{jj}^{1/2} \right| = O\left(\frac{\sqrt{k} \log^{1/4} p}{n^{1/4}} \right) \right\},
\end{aligned}
$$

hold with probability at least $1 - O(p^{-c})$ *for some constant* $c > 0$*.*

In (C2) of Lemma 10, we assume that $k = o(\sqrt{n}/\log^{5/2} p)$, so event A_5' in Lemma 12 implies Lemma 10 under (C2). Now we proceed to prove Lemma 1.

For event B_1, we first show that

$$
\max_j |\check{M}_j - \tilde{M}_j| = o\left(\frac{1}{\sqrt{\log p}} \right), \tag{2.2}
$$

holds in probability. To see this, note that for any j,

$$
\begin{aligned}
|\check{M}_j - \tilde{M}_j| &\leq \left| \frac{\langle v_j, \epsilon\rangle}{\|v_j\|_n} - \frac{\langle v_j, \epsilon\rangle}{\sqrt{nF_{jj}}} \right| + \left| \frac{\langle v_j, \epsilon\rangle}{\sqrt{nF_{jj}}} - \frac{\sum_{i=1}^n \eta_{ij}\epsilon_i/\dot{f}(u_i)}{\sqrt{nF_{jj}}} \right| \\
&= T_1 + T_2.
\end{aligned}
$$

It follows that

$$
T_1 \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n v_{ij}\epsilon_i \right| \cdot \left| \frac{\sqrt{n}}{\|v_j\|_n} - \frac{1}{\sqrt{F_{jj}}} \right|. \tag{2.3}
$$

To bound T_1, by Lemma 10, we only need to obtain an upper bound of $\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n v_{ij}\epsilon_i\right|$. Note that conditional on X and $\hat{\beta}$, v_{ij} is fixed and $v_{ij}\epsilon_i$ are conditional independent sub-gaussian random variables. In particular, we have $\mathbb{E}[v_{ij}\epsilon_i|X,\hat{\beta}] = 0$ and $\mathbb{E}[v_{ij}^2\epsilon_i^2|X,\hat{\beta}] \leq v_{ij}^2$. Thus, by concentration of independent sub-gaussian random variables, for any $t \geq 0$

$$
P_\theta\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\epsilon_i \geq t \middle| X, \hat{\beta}\right) \leq \exp\left(-\frac{t^2 n^2}{2\sum_{i=1}^n v_{ij}^2}\right).
$$

It then follows that

$$
P_\theta\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\epsilon_i \geq t\right) = \int P_\theta\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\epsilon_i \geq t \middle| X, \hat{\beta}\right) dP_{X,\hat{\beta}} \leq \mathbb{E}\exp\left(-\frac{t^2 n^2}{2\sum_{i=1}^n v_{ij}^2}\right).
$$

Let $t = C\sqrt{\log p/n}$, we have

$$
P_\theta\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\epsilon_i \geq C\sqrt{\frac{\log p}{n}}\right) \leq \mathbb{E}\exp\left(-\frac{c\log p}{2\sum_{i=1}^n v_{ij}^2/n}\right). \tag{2.4}
$$

Now under either (C1) or (C2), we have

$$
\left| \frac{1}{n}\sum_{i=1}^n v_{ij}^2 - \frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}^2(u_i) \right| \leq \max_i |\hat{\eta}_{ij}^2/\dot{f}(\hat{u}_i)^2 - \eta_{ij}^2/\dot{f}^2(u_i)| = o_P(1).
$$

To see this, by Lemma 11 and Lemma 12, we have

$$
\begin{aligned}
\max_i |\hat{\eta}_{ij}^2/\dot{f}^2(\hat{u}_i) - \eta_{ij}^2/\dot{f}^2(u_i)| &\leq \max_i \frac{|\eta_{ij}^2\dot{f}^2(\hat{u}_i) - \hat{\eta}_{ij}^2\dot{f}^2(u_i)|}{r^2(r^2 - o(1))} \\
&\leq \max_i \frac{\eta_{ij}^2|\dot{f}^2(\hat{u}_i) - \dot{f}^2(\hat{u}_i)| + \dot{f}^2(u_i)|\hat{\eta}_{ij}^2 - \eta_{ij}^2|}{r^2(r^2 - o(1))} \\
&= \begin{cases} O(k\log^2 p/\sqrt{n}) & \text{under (C1)} \\ O(k\log^{1/2} p/\sqrt{n}) & \text{under (C2)} \end{cases}
\end{aligned}
$$

with probability at least $1 - O(p^{-c})$. By concentration inequality for sub-exponential random variables $\eta_{ij}^2/\dot{f}^2(u_i)$ (see the arguments following (2.26) in the proof of Lemma 10 for more details), we have

$$P_\theta\left(\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}^2(u_i) > C + \sqrt{\frac{\log p}{n}}\right) = O(p^{-c})$$

for some $C, c > 0$. Thus it follows that

$$P_\theta\left(\frac{1}{n}\sum_{i=1}^n v_{ij}^2 > C\right) = O(p^{-c}).$$

for some $C, c > 0$. Now notice that

$$\begin{aligned}
\mathbb{E}\exp\left(-\frac{c\log p}{2\sum_{i=1}^n v_{ij}^2/n}\right) &\le \mathbb{E}\left[\exp\left(-\frac{c\log p}{2\sum_{i=1}^n v_{ij}^2/n}\right)1\left\{\frac{1}{n}\sum_{i=1}^n v_{ij}^2 \le C\right\}\right]\\
&\quad + \mathbb{E}\left[\exp\left(-\frac{c\log p}{2\sum_{i=1}^n v_{ij}^2/n}\right)1\left\{\frac{1}{n}\sum_{i=1}^n v_{ij}^2 > C\right\}\right]\\
&\le p^{-1/2C} + O(p^{-c'})\\
&= O(p^{-c}),
\end{aligned}$$

by (2.18), we have

$$P_\theta\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n v_{ij}\epsilon_i \ge C\sqrt{\log p}\right) = O(p^{-c}). \tag{2.5}$$

Thus, combining with Lemma 10, we have

$$T_1 \le C\sqrt{\log p}\cdot o\left(\frac{1}{\log p}\right) = o\left(\frac{1}{\sqrt{\log p}}\right),$$

with probability at least $1 - O(p^{-c})$. On the other hand,

$$\begin{aligned}
T_2 &\le F_{jj}^{-1/2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n v_{ij}\epsilon_i - \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{ij}\epsilon_i/\dot{f}(u_i)\right|\\
&= F_{jj}^{-1/2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\left[\frac{\hat{\eta}_{ij}}{\dot{f}(\hat{u}_i)} - \frac{\eta_{ij}}{\dot{f}(u_i)}\right]\right|.
\end{aligned}$$

Following the same conditional argument as (2.18), we have

$$P_\theta\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\left[\frac{\hat{\eta}_{ij}}{\dot{f}(\hat{u}_i)} - \frac{\eta_{ij}}{\dot{f}(u_i)}\right] \ge t\right) \le \mathbb{E}\exp\left(-\frac{t^2}{2\sum_{i=1}^n \alpha_{ij}^2/n}\right)$$

where $\alpha_{ij} = \frac{\hat{\eta}_{ij}}{\dot{f}(\hat{u}_i)} - \frac{\eta_{ij}}{\dot{f}(u_i)}$. Under (C2), we have $\alpha_{ij}^2 = O\big(\frac{k^2 \log p}{n}\big)$. Then

$$P_\theta\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\bigg[\frac{\hat{\eta}_{ij}}{\dot{f}(\hat{u}_i)} - \frac{\eta_{ij}}{\dot{f}(u_i)}\bigg] \geq t\bigg) \leq \exp\bigg(-\frac{nt^2}{2k^2 \log p}\bigg) + O(p^{-c}).$$

Let $t = k\log p/\sqrt{n}$, we have

$$P_\theta\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\bigg[\frac{\hat{\eta}_{ij}}{\dot{f}(\hat{u}_i)} - \frac{\eta_{ij}}{\dot{f}(u_i)}\bigg] \geq \frac{k\log p}{\sqrt{n}}\bigg) = O(p^{-c}).$$

Therefore $T_2 = O\big(\frac{k\log p}{\sqrt{n}}\big) = o(1/\sqrt{\log p})$ with probability at least $1 - O(p^{-c})$ as long as $k = o(\sqrt{n}/\log^{3/2} p)$. Under (C1), similar argument yields $T_2 = o(1/\sqrt{\log p})$ with probability at least $1 - O(p^{-c})$ as long as $k = o(\sqrt{n}/\log^2 p)$. Using a union bound argument across $j = 1, ..., p$, we prove that (2.2) holds in probability. Using the same argument, we can prove

$$P_\theta\bigg(\max_j |\check{M}_j| > C\sqrt{\log p}\bigg) = O(p^{-c}). \tag{2.6}$$

Therefore, we have

$$|\check{M}_n - \tilde{M}_n| \leq \max_j |\check{M}_j^2 - \tilde{M}_j^2| \leq C(\max_j |\tilde{M}_j|) \cdot \max_j |\check{M}_j - \tilde{M}_j| = o(1)$$

with probability at least $1 - O(p^{-c})$. This completes the proof of event B_1.

For event B_2, note that

$$|\tilde{M}_n - M_n| \leq \max_j |\tilde{M}_j^2 - M_j^2| \leq C(\max_j |\check{M}_j|) \cdot \max_j \bigg(\frac{|\langle v_j, Re_i\rangle|}{\|v_j\|_n} + \frac{|\langle v_j, \mathbf{h}_{-j}\rangle|}{\|v_j\|_n}\bigg). \tag{2.7}$$

To bound $\max_j |\langle v_j, Re\rangle|/\|v_j\|_n$, by Lemma 10 and mean value theorem,

$$\frac{|\langle v_j, Re\rangle|}{\|v_j\|_n} \leq \frac{\big|\sum_{i=1}^n v_{ij}(\dot{f}(\hat{u}_i) - \dot{f}(u_i^*))(\hat{u}_i - u_i)\big|}{\sqrt{n}(F_{jj}^{1/2} - o_P(1))}$$

Under (C1), $\max_{i,j} |v_{ij}| = O_P(\sqrt{\log p})$ and thereby

$$\begin{aligned}&\bigg|\sum_{i=1}^n v_{ij}(\dot{f}(\hat{u}_i) - \dot{f}(u_i^*))(\hat{u}_i - u_i)\bigg| \leq \sum_{i=1}^n (\hat{u}_i - u_i)^2 \cdot \max_{i,j} |v_{ij}| = \|X(\hat{\beta} - \beta)\|_2^2 \cdot O(\sqrt{\log p})\\ &= O(k\log^{3/2} p)\end{aligned}$$

with probability at least $1-O(p^{-c})$. Thus

$$\max_j \frac{|\langle v_j, Re\rangle|}{\|v_j\|_n} = O\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right)$$

in probability. Under (C2), $\max_{i,j}|v_{ij}| = O_P(1)$ and thereby

$$\begin{aligned}\left|\sum_{i=1}^n v_{ij}(\dot f(\hat u_i) - \dot f(u_i^*))(\hat u_i - u_i)\right| &\le \sum_{i=1}^n (\hat u_i - u_i)^2 \cdot \max_{i,j}|v_{ij}| = \|X(\hat\beta-\beta)\|_2^2 \cdot O(1) \\ &= O(k\log p)\end{aligned}$$

with probability at least $1-O(p^{-c})$. Thus

$$\max_j \frac{|\langle v_j, Re\rangle|}{\|v_j\|_n} = O\left(\frac{k\log p}{\sqrt{n}}\right) \tag{2.8}$$

In general, either (C1) or (C2) implies that

$$\max_j |\langle v_j, Re\rangle|/\|v_j\|_n = o(\log^{-3/2}p) \tag{2.9}$$

with probability at least $1-O(p^{-c})$. On the other hand, to bound $\max_j|\langle v_j, \mathbf{h}_{-j}\rangle|/\|v_j\|_n$, by Proposition 1 (ii) in Zhang and Zhang (2014), we know that if we choose $\lambda = C\sqrt{\log p/n}$, then under (C1) or (C2)

$$\max_{k\neq j}\frac{\langle\hat\eta_j, \mathbf{x}_k\rangle}{\|\hat\eta_j\|_2} \le C_1\sqrt{2\log p}, \qquad \frac{\|\hat\eta_j\|_2}{\langle\hat\eta_j, \mathbf{x}_j\rangle} \le \frac{C_2}{\sqrt{n}} \tag{2.10}$$

with probability at least $1-O(p^{-c})$. Note that

$$\|\hat\eta_j\|_2 = \sqrt{\sum_{i=1}^n \hat\eta_{ij}^2} = \sqrt{\sum_{i=1}^n \dot f^2(\hat u_i)v_{ij}^2} \le \sqrt{\sum_{i=1}^n \dot f(\hat u_i)v_{ij}^2} = \|v_j\|_n,$$

we have

$$\eta_j = \max_{k\neq j}\frac{\langle v_j, \mathbf{x}_k\rangle_n}{\|v_j\|_n} \le C_1\sqrt{2\log p} \tag{2.11}$$

in probability. Therefore under either (C1) or (C2)

$$\begin{aligned}\frac{|\langle v_j, \mathbf{h}_{-j}\rangle|}{\|v_j\|_n} &\le \|v_j\|_n^{-1}\left|\sum_{i=1}^n v_{ij}\dot f(\hat u_i)X_{i,-j}^\top(\hat\beta_{-j} - \beta_{-j})\right| \le \max_{k\neq j}\frac{|\langle v_j, \mathbf{x}_k\rangle_n|}{\|v_j\|_n}\cdot\|\hat\beta-\beta\|_1 \\ &= \eta_j\|\hat\beta-\beta\|_1 = O\left(\frac{k\log p}{\sqrt{n}}\right)\end{aligned} \tag{2.12}$$

with probability at least $1-O(p^{-c})$. Back to (2.7), note that $\max_j|\tilde M_n| \le \max_j|\check M_n| + o_P(1) =$

$O_P(\sqrt{\log p})$, we have

$$|\tilde{M}_n - M_n| = o\bigg(\frac{1}{\log p}\bigg)$$

with probability at least $1 - O(p^{-c})$. $\square$

Proof of Lemma 2. The lemma is proved under the Gaussian design. For the bounded design, by definition $\hat{M}_j$ is essentially the same as $\check{M}_j$. Note that

$$\begin{aligned}
\max_{1\le j\le p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[|v_{ij}^0\epsilon_i| 1\{|v_{ij}^0\epsilon_i| \ge \tau_n\}] &\le Cn^{1/2} \max_{i,j} \mathbb{E}[|v_{ij}^0\epsilon_i| 1\{|v_{ij}^0\epsilon_i| \ge \tau_n\}] \\
&\le Cn^{1/2}(p+n)^{-1} \max_{i,j} \mathbb{E}[|v_{ij}^0\epsilon_i| e^{|v_{ij}^0\epsilon_i|}] \\
&\le Cn^{1/2}(p+n)^{-1},
\end{aligned}$$

where the last inequality follows from

$$\mathbb{E}[|v_{ij}^0\epsilon_i| e^{|v_{ij}^0\epsilon_i|}] \le C_1 \sqrt{\mathbb{E}(v_{ij}^0)^2} \sqrt{\mathbb{E}\exp(2|v_{ij}^0\epsilon_i|)} \le C_2$$

by sub-gaussianity of v_{ij}^0. Hence, if $\max_{i,j}|v_{ij}^0\epsilon_i| \le \tau_n$, then

$$\hat{Z}_{ij} = v_{ij}^0\epsilon_i - \mathbb{E}[v_{ij}^0\epsilon_i 1\{|v_{ij}^0\epsilon_i| \le \tau_n\}]$$

and thereby

$$\begin{aligned}
\max_j |\hat{M}_j - \check{M}_j| &\le \max_{1\le j\le p} \bigg| \frac{1}{\sqrt{nF_{jj}}} \sum_{i=1}^{n} \mathbb{E}[v_{ij}^0\epsilon_i 1\{|v_{ij}^0\epsilon_i| \le \tau_n\}] \bigg| \\
&= \max_{1\le j\le p} \bigg| \frac{1}{\sqrt{nF_{jj}}} \sum_{i=1}^{n} \mathbb{E}[v_{ij}^0\epsilon_i 1\{|v_{ij}^0\epsilon_i| \ge \tau_n\}] \bigg| \\
&\le \max_{1\le j\le p} \frac{1}{\sqrt{nF_{jj}}} \sum_{i=1}^{n} \mathbb{E}[|v_{ij}^0\epsilon_i| 1\{|v_{ij}^0\epsilon_i| \ge \tau_n\}] \\
&\le Cn^{1/2}(p+n)^{-1} \\
&= O(1/\log p).
\end{aligned}$$

Then we have

$$P_\theta\bigg(\max_j |\hat{M}_j - \check{M}_j| \ge C(\log p)^{-1}\bigg) \le P\bigg(\max_{i,j} |v_{ij}\epsilon_i| \ge \tau_n\bigg) = O(p^{-c}). \tag{2.13}$$

Now by the fact that

$$|\hat{M}_n - \check{M}_n| \le 2 \max_j |\check{M}_i| \max_j |\check{M}_j - \hat{M}_j| + \max_j |\check{M}_j - \hat{M}_j|^2,$$

it suffices to apply (2.13) and (2.6) in the proof of Lemma 1. $\square$

Proof of Lemma 6. By definition, we have

$$\begin{aligned}
\chi^2(g,f) &= \int \frac{g^2}{f} - 1 \\
&= \frac{1}{\binom{p}{k}^2} \int \frac{\left(\sum_{\beta\in\mathcal{H}_1} \prod_{i=1}^n p(X_i, y_i; \beta)\right)^2}{\prod_{i=1}^n p(X_i, y_i)} - 1 \\
&= \frac{1}{\binom{p}{k}^2} \sum_{\beta\in\mathcal{H}_1} \sum_{\beta'\in\mathcal{H}_1} \prod_{i=1}^n \int \frac{p(X_i, y_i; \beta) p(X_i, y_i; \beta')}{p(X_i, y_i)} - 1. \qquad (2.14)
\end{aligned}$$

Note that

$$\begin{aligned}
&\int \frac{p(X_i, y_i; \beta) p(X_i, y_i; \beta')}{p(X_i, y_i)} \qquad (2.15) \\
&= \frac{1}{(2\pi)^{p/2}} \int\int \frac{2\exp(-\frac{1}{2} X_i^\top X_i + y_i X_i^\top(\beta+\beta'))}{[1+\exp(X_i^\top \beta)][1+\exp(X_i^\top\beta')]} dy_i dX_i \\
&= \frac{1}{(2\pi)^{p/2}} \int \frac{2\exp(-\frac{1}{2} X_i^\top X_i + X_i^\top(\beta+\beta'))}{[1+\exp(X_i^\top \beta)][1+\exp(X_i^\top\beta')]} dX_i \\
&\quad + \frac{1}{(2\pi)^{p/2}} \int \frac{2\exp(-\frac{1}{2} X_i^\top X_i)}{[1+\exp(X_i^\top \beta)][1+\exp(X_i^\top\beta')]} dX_i \\
&= \mathbb{E} h(X; \beta, \beta') \qquad (2.16)
\end{aligned}$$

where in the last equality, the expectation is with respect to a standard multivariate normal random vector $X \sim N(0, \mathbf{I}_p)$ and

$$\begin{aligned}
h(X;\beta,\beta') &=: \frac{2(1+e^{X^\top(\beta+\beta')})}{(1+e^{X^\top\beta})(1+e^{X^\top\beta'})} = 1 + \frac{e^{X^\top\beta}-1}{e^{X^\top\beta}+1} \frac{e^{X^\top\beta'}-1}{e^{X^\top\beta'}+1} \\
&= 1 + \tanh\left(\frac{X^\top\beta}{2}\right) \tanh\left(\frac{X^\top\beta'}{2}\right)
\end{aligned}$$

Lemma 13. *If* $(X,Y) \sim N(0, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *for some* $\sigma^2 \le 1$*, then it follows*

$$\mathbb{E} \tanh\left(\frac{X}{2}\right) \tanh\left(\frac{Y}{2}\right) \le 6\sigma^2\rho.$$

Now since $X_i^\top \beta \sim N(0, k\rho^2)$, where we can choose ρ such that $k\rho^2 \le 1$. By Lemma 13, let $j = |\text{supp}(\beta) \cap \text{supp}(\beta')| = |I \cap I'|$ be the number of intersected components between β and β', we have

$$\chi^2(g, f) \le \frac{1}{\binom{p}{k}^2} \sum_{\beta \in \mathcal{H}_1} \sum_{\beta' \in \mathcal{H}_1} \left(1 + 6\beta^\top \beta'\right)^n - 1 = \frac{1}{\binom{p}{k}^2} \sum_{\beta \in \mathcal{H}_1} \sum_{\beta' \in \mathcal{H}_1} \left(1 + 6j\rho^2\right)^n - 1$$

Note that for β, β' uniformly picked from $\mathcal{H}_1$, j follows a hypergeometric distribution

$$P(J = j) = \frac{\binom{k}{j}\binom{p-k}{k-j}}{\binom{p}{k}}, \quad j = 0, 1, ..., k.$$

Then

$$\chi^2(g, f) \le \mathbb{E}(1 + 6\rho^2 J)^n - 1 = \mathbb{E}\exp(n \log(1 + 6\rho^2 J)) - 1 \le \mathbb{E}e^{6n\rho^2 J} - 1.$$

As shown on page 173 of (Aldous, 1985), J has the same distribution as the random variable $\mathbb{E}(Z|\mathcal{B}_n)$ where Z is a binomial random variable of parameters $(k, k/p)$ and $\mathcal{B}_n$ some suitable σ-algebra. Thus by Jensen's inequality we have

$$\mathbb{E}e^{6nJ\rho^2} \le \left(1 - \frac{k}{p} + \frac{k}{p} e^{6n\rho^2}\right)^k.$$

Let

$$\rho^2 = \frac{1}{6n} \log\left(1 + \frac{p}{h(\eta)k^2}\right),$$

where $h(\eta) = [\log(\eta^2 + 1)]^{-1}$ and $\eta = 1 - \alpha - \delta$, we have

$$\mathbb{E}e^{6n\rho^2 J} \le e^{1/h(\eta)},$$

so that

$$\chi^2(g, f) \le \eta^2 = (1 - \alpha - \delta)^2.$$

$\square$

Proof of Lemma 9. Note that

$$|M_j'| \le \frac{|\langle v_j, \epsilon \rangle|}{\|v_j\|_n} + \frac{|\langle v_j, Re \rangle|}{\|v_j\|_n} + \frac{|\langle v_j, \mathbf{h}_{-j} \rangle_n|}{\|v_j\|_n}.$$

We bound the above three terms one by one. Firstly, by concentration of sub-exponential random variables η_{ij}^2 (see (2.26) in the proof of Lemma 10 for details) and (2.18), we have

$$P_\theta\bigg(\bigg|\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2 - \mathbb{E}\eta_{ij}^2\bigg| \geq (\log p)^{-1}\bigg) = O(p^{-c}) \tag{2.17}$$

Then we have

$$\frac{|\langle v_j, \epsilon\rangle|}{\|v_j\|_n} \leq \frac{n^{-1/2}\sum_{i=1}^n \hat{\eta}_{ij}\epsilon_i/\dot{f}(\hat{u}_i)}{\sqrt{\sum_{i=1}^n \hat{\eta}_{ij}^2/n}} \leq \frac{C}{\sqrt{n}}\sum_{i=1}^n \hat{\eta}_{ij}\epsilon_i/\dot{f}(\hat{u}_i) \equiv \frac{C}{\sqrt{n}}\sum_{i=1}^n \xi_i.$$

Conditional on X and $\hat{\beta}$, we have $\mathbb{E}[\xi_i|X,\hat{\beta}] = 0$ and $\mathbb{E}[\xi_i^2|X,\hat{\beta}] \leq \hat{\eta}_{ij}^2/\dot{f}^2(\hat{u}_i) = \alpha_{ij}(n)$. By concentration inequality for independent sub-gaussian random variables $\xi_i|X,\hat{\beta}$, we have for any $t \geq 0$

$$P_\theta\bigg(\frac{1}{n}\sum_{i=1}^n \xi_i \geq t\bigg|X,\hat{\beta}\bigg) \leq \exp\bigg(-\frac{t^2n^2}{2\sum_{i=1}^n \alpha_{ij}(n)}\bigg).$$

It then follows that

$$P_\theta\bigg(\frac{1}{n}\sum_{i=1}^n \xi_i \geq t\bigg) = \int P_\theta\bigg(\frac{1}{n}\sum_{i=1}^n \xi_i \geq t\bigg|X,\hat{\beta}\bigg)dP_{X,\hat{\beta}} \leq \mathbb{E}\exp\bigg(-\frac{t^2n^2}{2\sum_{i=1}^n \alpha_{ij}(n)}\bigg).$$

Let $t = C\sqrt{\log p/n}$, we have

$$P_\theta\bigg(\frac{1}{n}\sum_{i=1}^n \xi_i \geq C\sqrt{\frac{\log p}{n}}\bigg) \leq \mathbb{E}\exp\bigg(-\frac{c\log p}{2\sum_{i=1}^n \alpha_{ij}(n)/n}\bigg). \tag{2.18}$$

Now since with probability at least $1 - O(n^{-c})$, $\alpha_{ij}(n) \leq \hat{\eta}_{ij}^2 L(n)^{-2}$, or

$$P_\theta\bigg(\frac{1}{n}\sum_{i=1}^n \alpha_{ij}(n) \geq L(n)^{-2}\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2\bigg) = O(n^{-c}),$$

by (2.17), we have

$$P\bigg(\frac{1}{n}\sum_{i=1}^n \alpha_{ij}(n) \geq CL(n)^{-2}\bigg) = O(n^{-c})$$

for some $C, c > 0$. Now notice that

$$\begin{aligned}
\mathbb{E}\exp\bigg(-\frac{c\log p}{2\sum_{i=1}^n v_{ij}^2/n}\bigg) &\le \mathbb{E}\bigg[\exp\bigg(-\frac{c\log p}{2\sum_{i=1}^n \alpha_{ij}/n}\bigg)1\bigg\{\frac{1}{n}\sum_{i=1}^n \alpha_{ij} \le CL(n)^{-2}\bigg\}\bigg] \\
&\quad + \mathbb{E}\bigg[\exp\bigg(-\frac{c\log p}{2\sum_{i=1}^n \alpha_{ij}/n}\bigg)1\bigg\{\frac{1}{n}\sum_{i=1}^n \alpha_{ij} > CL(n)^{-2}\bigg\}\bigg] \\
&\le p^{-1/(2CL^{-2}(n))} + O(n^{-c'}) \\
&= O(n^{-c}),
\end{aligned}$$

where we used the fact that

$$p^{-1/(2CL^{-2}(n))} \asymp p^{-\exp(-c_1\|\beta\|_2\sqrt{\log n})} \lesssim n^{-c}$$

for sufficiently small $c > 0$, as long as $\|\beta\|_2 = O\big(\frac{\log\log p}{\sqrt{\log n}}\big)$ and $\log p \gtrsim \log^{1+\delta} n$. As a result, by (2.18), we have

$$P_\theta\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \ge C\sqrt{\log p}\bigg) = O(n^{-c}). \tag{2.19}$$

To bound $|\langle v_j, Re\rangle|/\|v_j\|_n$, by mean value theorem,

$$\frac{|\langle v_j, Re\rangle|}{\|v_j\|_n} \le \frac{n^{-1/2}\big|\sum_{i=1}^n v_{ij}(\dot f(\hat u_i) - \dot f(u_i^*))(\hat u_i - u_i)\big|}{\sqrt{\sum_{i=1}^n \hat\eta_{ij}^2/n}}$$

Note that $\max_i |v_{ij}| = O(\sqrt{\log p}L^{-1}(n))$ with probability at least $1 - O(n^{-c})$, thereby

$$\begin{aligned}
&\bigg|\sum_{i=1}^n v_{ij}(\dot f(\hat u_i) - \dot f(u_i^*))(\hat u_i - u_i)\bigg| \le \sum_{i=1}^n (\hat u_i - u_i)^2 \cdot \max_{i,j}|v_{ij}| \\
&= \|X(\hat\beta - \beta)\|_2^2 \cdot O(L^{-1}(n)\sqrt{\log p}) = O(k\log^{3/2} pL^{-1}(n))
\end{aligned}$$

with probability at least $1 - O(n^{-c})$. Since $\|\beta\|_2 \le C\big(\frac{\log\log p}{\sqrt{\log n}}\big)$, for some $C \le \sqrt{2/\lambda_{\max}(\mathbf{\Sigma})}$, we have

$$\frac{|\langle v_j, Re\rangle|}{\|v_j\|_n} = O\bigg(\frac{k\log^{3/2} p}{L(n)\sqrt{n}}\bigg) = o(\sqrt{\log p}) \tag{2.20}$$

with probability at least $1 - O(n^{-c})$. Finally, to bound $\max_j |\langle v_j, \mathbf{h}_{-j}\rangle|/\|v_j\|_n$, by (2.11) we have

$$\frac{|\langle v_j, \mathbf{h}_{-j}\rangle|}{\|v_j\|_n} = O\bigg(\frac{k\log p}{\sqrt{n}}\bigg) = o(1) \tag{2.21}$$

with probability at least $1 - O(n^{-c})$. Combining (2.19) (2.20) and (2.21), we have proven Lemma 9. □

Proof of Lemma 11. Event A_0 and A_2 follows from Corollary 1, 2 and 5 of Negahban et al. (2010). For the event A_1, by Theorem 1.6 of Zhou (2009), the condition of Lemma 11 implies the restricted eigenvalue condition, which, by Lemma 2.1 and Figure 1 of van de Geer and Bühlmann (2009), implies event A_1. For event A_3, note that under A_2 we have

$$\max_{i,j}|\hat{\eta}_{ij}-\eta_{ij}| = \max_{i,j}|X_{i,-j}(\hat{\gamma}_j-\gamma_j)| \le \max_{i,j}\|X_{i,-j}\|_\infty \max_j \|\hat{\gamma}_j-\gamma_j\|_1 \le Ck\frac{\log p}{\sqrt{n}}$$

where the last inequality follows from that fact that

$$P_\theta\left(\max_{1\le i\le p} X_i \ge \sqrt{C\log p}\right) \le \frac{1}{p^c} \tag{2.22}$$

for some sufficiently large constant $C, c > 0$, which is a consequence of the Gaussian tail probability bound $1-\Phi(x) \le \frac{1}{x}\phi(x)$ by taking $x = \sqrt{C\log p}$ for some sufficiently large $C > 0$.

For event A_4, since $\|X\beta\|_\infty \le c_2$ for some constant $c_2 > 0$, there exists some constant $0 < \kappa < 1$ such that $\kappa < |f(u_i)| < 1-\kappa$ and thereby $\dot{f}(u_i) \ge \kappa(1-\kappa)$ for all i. A_4 then follows from the following lemma, event A_0 and (2.22).

Lemma 14. *Let $f(x) = \frac{e^x}{1+e^x}$, then uniformly over $a, b \in \mathbb{R}$, it holds that*

$$\left|\frac{1}{\dot{f}(a)} - \frac{1}{\dot{f}(b)}\right| \le \frac{\max\{\dot{f}(a), \dot{f}(b)\}}{\dot{f}(a)\dot{f}(b)}|a-b| \le \frac{1}{\dot{f}(a)\dot{f}(b)}|a-b|. \tag{2.23}$$

For event A_5, by the fact that $v_{ij} = \hat{\eta}_{ij}/\dot{f}(\hat{u}_i)$, it follows that

$$\begin{aligned}
\left|\frac{\|v_j\|_n}{\sqrt{n}} - F_{jj}^{1/2}\right| &= \left|\left[\frac{1}{n}\sum_{i=1}^n v_{ij}^2 \dot{f}(\hat{u}_i)\right]^{1/2} - F_{jj}^{1/2}\right| \le \left|\frac{1}{n}\sum_{i=1}^n v_{ij}^2\dot{f}(\hat{u}_i) - F_{jj}\right|^{1/2} \\
&\le \left|\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2/\dot{f}(\hat{u}_i) - \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2/\dot{f}(u_i)\right|^{1/2} + \left|\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2/\dot{f}(u_i) - F_{jj}\right|^{1/2} \\
&\le \left|\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2\left[\frac{1}{\dot{f}(\hat{u}_i)} - \frac{1}{\dot{f}(u_i)}\right]\right|^{1/2} + \left|\frac{1}{n}\sum_{i=1}^n(\hat{\eta}_{ij}^2 - \eta_{ij}^2)/\dot{f}(u_i)\right|^{1/2} \\
&\quad + \left|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}(u_i) - F_{jj}\right|^{1/2} \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

To bound I_2, note that $\|X\beta\|_\infty \le c$ implies $\max_i \dot{f}(u_i) \ge r$ for some constant $r > 0$, and that

$\hat{\eta}_j - \eta_j = X_{-j}(\hat{\gamma}_j - \gamma_j)$, we have

$$
\begin{aligned}
I_2^2 &\le \frac{1}{rn}\sum_{i=1}^n |\hat{\eta}_{ij}^2 - \eta_{ij}^2| \le \frac{1}{rn}\sum_{i=1}^n [|\hat{\eta}_{ij} - \eta_{ij}|^2 + 2|\hat{\eta}_{ij} - \eta_{ij}| \cdot |\eta_{ij}|] \\
&\le \frac{1}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2^2 + \frac{2C\sqrt{\log p}}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_1 \\
&\le \frac{1}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2^2 + \frac{2C\sqrt{\log p}}{r\sqrt{n}}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2. \qquad (2.24)
\end{aligned}
$$

Therefore, by event A_1, as long as $k < n$,

$$
I_2^2 \le C_1 k\frac{\log p}{n} + C_2\sqrt{k}\frac{\log p}{\sqrt{n}} = O\bigg(\sqrt{k}\frac{\log p}{\sqrt{n}}\bigg) \qquad (2.25)
$$

with probability at least $1 - O(p^{-c})$ for some $c > 0$. By A_4 and (2.25), we have, with probability at least $1 - O(p^{-c})$ for some $c > 0$,

$$
I_1^2 \le \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2 \cdot Ck\frac{\log p}{\sqrt{n}} \le \bigg[\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2 + o(1)\bigg] \cdot Ck\frac{\log p}{\sqrt{n}} \le C'k\frac{\log p}{\sqrt{n}},
$$

where the last inequality follows from the concentration inequality

$$
P_\theta\bigg(\bigg|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2 - \mathbb{E}\eta_{ij}^2\bigg| \ge \sqrt{\frac{\log p}{n}}\bigg) = O(p^{-c}). \qquad (2.26)
$$

To show this, we need to introduce the following norms for random variables. The sub-gaussian norm of a random variable U is defined as $\|U\|_{\psi_2} = \sup_{q\ge 1}\frac{1}{\sqrt{q}}(\mathbb{E}|U|^q)^{1/q}$, and the sub-exponential norm of a random variable is defined as $\|U\|_{\psi_1} = \sup_{q\ge 1} q^{-1}(\mathbb{E}|U|^q)^{1/q}$. By definition η_{ij} are sub-gaussian with $\|\eta_{ij}\|_{\psi_2} < C < \infty$ and therefore

$$
\|\eta_{ij}^2\|_{\psi_1} = \sup_{q\ge 1} q^{-1}(\mathbb{E}|\eta_{ij}|^{2q})^{1/q} = \sup_{q\ge 1}[q^{-1/2}(\mathbb{E}|\eta_{ij}|^{2q})^{1/2q}]^2 = \|\eta_{ij}\|_{\psi_2}^2 < C^2 < \infty.
$$

So η_{ij}^2 with $i = 1, ..., n$ are i.i.d. sub-exponential random variables. Then (2.26) follows from standard concentration inequality for sub-exponential random variables (see, for example, Proposition 5.16 in Vershynin (2010)). Similarly, we can show $\eta_{ij}^2/\dot{f}(u_i)$ are sub-exponential and therefore

$$
I_3^2 = \bigg|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}(u_i) - F_{jj}\bigg| = O\bigg(\sqrt{\frac{\log p}{n}}\bigg)
$$

with probability at least $1 - O(p^{-c})$ for some $c > 0$. Thus, $I_1 + I_2 + I_3 = O\big(\frac{\sqrt{k\log p}}{n^{1/4}}\big)$. □

Proof of Lemma 12. Events A_0 A_1 and A_2 follow the same argument as in Lemma 11. For event A_3', by A_1, A_2 and boundedness of X, we have

$$\max_{i,j} |\hat{\eta}_{ij} - \eta_{ij}| = \max_{i,j} |X_{i,-j}(\hat{\gamma}_j - \gamma_j)| \leq \max_{i,j} \|X_{i,-j}\|_\infty \max_j \|\hat{\gamma}_j - \gamma_j\|_1 \leq Ck\sqrt{\frac{\log p}{n}}$$

For event A_4', by (A4), there exists some constant $r > 0$ such that $\dot{f}(u_i) \geq r$ for all i with probability at least $1 - O(p^{-c})$. A_4' then follows from Lemma 14. For event A_5', as the proof of A_5 in Lemma 11, we have

$$\begin{aligned}\left|\sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2 \dot{f}(\hat{u}_i)} - F_{jj}^{1/2}\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2 \left[\frac{1}{\dot{f}(\hat{u}_i)} - \frac{1}{\dot{f}(u_i)}\right]\right|^{1/2} + \left|\frac{1}{n}\sum_{i=1}^n (\hat{\eta}_{ij}^2 - \eta_{ij}^2)/\dot{f}(u_i)\right|^{1/2} \\ &\quad + \left|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}(u_i) - F_{jj}\right|^{1/2} \\ &= I_1 + I_2 + I_3.\end{aligned}$$

To bound I_2, note that $P(\max_i \dot{f}(u_i) \leq r) = O(p^{-c})$, and that $\hat{\eta}_j - \eta_j = X_{-j}(\hat{\gamma}_j - \gamma_j)$, by A_1,

$$\begin{aligned}I_2^2 &\leq \frac{1}{rn}\sum_{i=1}^n |\hat{\eta}_{ij}^2 - \eta_{ij}^2| \leq \frac{1}{rn}\sum_{i=1}^n [|\hat{\eta}_{ij} - \eta_{ij}|^2 + 2|\hat{\eta}_{ij} - \eta_{ij}| \cdot |\eta_{ij}|] \\ &\leq \frac{1}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2^2 + \frac{2C}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_1 \\ &\leq \frac{1}{rn}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2^2 + \frac{2C}{r\sqrt{n}}\|X_{-j}(\hat{\gamma} - \gamma_j)\|_2 \\ &\leq C_1 k\frac{\log p}{n} + C_2\sqrt{k\frac{\log p}{n}} = O\left(\sqrt{k\frac{\log p}{n}}\right) \end{aligned} \tag{2.27}$$

with probability at least $1 - O(p^{-c})$ for some $c > 0$. For I_1, by A_4' and boundedness of X, we have, with probability at least $1 - O(p^{-c})$ for some $c > 0$,

$$I_1^2 \leq \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{ij}^2 \cdot Ck\sqrt{\frac{\log p}{n}} = O\left(k\sqrt{\frac{\log p}{n}}\right).$$

Finally, by concentration inequality for sub-exponential random variables $\eta_{ij}^2/\dot{f}(u_i)$ for $i = 1, ..., n$, we have

$$I_3^2 = \left|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2/\dot{f}(u_i) - F_{jj}\right| = O\left(\sqrt{\frac{\log p}{n}}\right)$$

with probability at least $1 - O(p^{-c})$ for some $c > 0$. Thus, $I_1 + I_2 + I_3 = O\big(\frac{\sqrt{k}\log^{1/4} p}{n^{1/4}}\big)$. $\square$

Proof of Lemma 13. By normalization, we only need to consider (X, Y) with $\text{Var}(X) = \text{Var}(Y) = 1$ and $\mathbb{E}XY = \rho$ and prove

$$\mathbb{E}\tanh\left(\frac{\sigma X}{2}\right)\tanh\left(\frac{\sigma Y}{2}\right) \le 10\sigma^2\rho. \tag{2.28}$$

Note that the inner product

$$\langle X, Y\rangle = \mathbb{E}XY$$

defines a Hilbert space on $L^2(\Omega, \mathcal{F}, \mu)$. Then the above inequality is equivalent to

$$\left\langle \tanh\left(\frac{\sigma X}{2}\right), \tanh\left(\frac{\sigma Y}{2}\right)\right\rangle \le \sigma^2\langle X, Y\rangle.$$

Consider the Hermite polynomials $H_n(x), x \in \mathbb{R}, n = 0, 1, ...$ which are defined as

$$H_n = \frac{(-1)^n}{n!}e^{x^2/2}\frac{d^n}{dx^n}(e^{-x^2/2}),$$

so that in particular $H_0(x) = 1, H_1(x) = x, H_2(x) = (x^2-1)/2$, and in general $H_n(x)$ is a polynomial of order n. The Hermite polynomials satisfy the following basic identities

$$\begin{aligned} H_n'(x) &= H_{n-1}(x) \\ (n+1)H_{n+1}(x) &= xH_n(x) - H_{n-1}(x), \\ H_n(-x) &= (-1)^n H_n(x), \end{aligned} \tag{2.29}$$

for all $n \ge 1$. For X, Y that are $N(0, 1)$ random variables that are jointly Gaussian, it can be shown (see, for example, Section 2.10 of Kolokoltsov (2011)) that

$$\langle H_n(X), H_m(Y)\rangle = \mathbb{E}(H_n(X)H_m(Y)) = \begin{cases} 0 & \text{if } m \neq n, \\ \frac{1}{n!}(\mathbb{E}XY)^n & \text{if } m = n. \end{cases} \tag{2.30}$$

Now we would like to expand the function $\tanh(\sigma x/2)$ in terms of orthogonal Hermite polynomials as

$$\tanh(\sigma x/2) = \sum_{n=0}^{\infty} C_n H_n(x).$$

To calculate the coefficients C_n, simply note that

$$C_n = \frac{\big\langle \tanh(\sigma X/2), H_n(X)\big\rangle}{\big\langle H_n(X), H_n(X)\big\rangle} = \frac{(-1)^n}{(2\pi)^{1/2}}\int \tanh\left(\frac{\sigma x}{2}\right)\frac{d^n}{dx^n}(e^{-x^2/2})dx.$$

Denote $\phi(x) = e^{-x^2/2}$, we have

$$C_n = \frac{(-1)^n}{\sqrt{2\pi}} \int \tanh\left(\frac{\sigma x}{2}\right) \phi^{(n)}(x) dx$$

Note that $\phi(x)$ is an even function and $\tanh(x)$ is an odd function, so the integrand $\phi^{(n)}(x)\tanh(\sigma x/2)$ is an odd function for all odd $n > 0$. Therefore $C_{2k} = 0$ for any $k \geq 0$. Now we calculate for $k \geq 0$, through integration by parts,

$$\begin{aligned} C_{2k+1} &= \frac{(-1)^{2k+1}}{\sqrt{2\pi}} \int \tanh\left(\frac{\sigma x}{2}\right) \phi^{(2k+1)}(x) dx = \frac{(-1)^{2k+1}}{\sqrt{2\pi}} \int \tanh^{(1)}\left(\frac{\sigma x}{2}\right) \phi^{(2k)}(x) dx \\ &= ... = \frac{(-1)^{2k+1}}{\sqrt{2\pi}} \int \tanh^{(2k+1)}\left(\frac{\sigma x}{2}\right) \phi(x) dx. \end{aligned}$$

By the fact that, for any $x \geq 0$,

$$\tanh^{(n)}(x/2) \leq \sinh^{(n)}(x),$$

we have

$$\begin{aligned} C_{2k+1} &\leq \frac{(-1)^{2k+1}}{\sqrt{2\pi}} \int \sinh^{(2k+1)}(\sigma x) \phi(x) dx \\ &= \frac{(-1)^{2k+1}}{\sqrt{2\pi}} \int \sinh(\sigma x) \phi^{(2k+1)}(x) dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^\infty \sinh(\sigma x) H_{2k+1}(x)(2k+1)! \phi(x) dx \\ &= \sigma^{2k+1} e^{\sigma^2/2}, \end{aligned}$$

where the last equation follows from Equation 7.387.1 of Gradshteyn and Ryzhik (2014). As a result,

$$\begin{aligned} &\left\langle \tanh\left(\frac{\sigma X}{2}\right), \tanh\left(\frac{\sigma Y}{2}\right) \right\rangle = \left\langle \sum_{n=0}^\infty C_n H_n(X), \sum_{n=0}^\infty C_n H_n(Y) \right\rangle = \sum_{n=0}^\infty C_n^2 \langle H_n(X), H_n(Y) \rangle \\ &= \sum_{k=0}^\infty \frac{C_{2k+1}^2 \rho^{2k+1}}{(2k+1)!} \leq e^{\sigma^2} \sum_{k=0}^\infty \frac{(\sigma^2 \rho)^{2k+1}}{(2k+1)!} = e^{\sigma^2} \sinh(\sigma^2 \rho). \end{aligned}$$

Now since $\sinh(x) \leq 2x$ for $0 \leq x \leq 1$. To see this, note that

$$\frac{d}{dx}(\sinh(x) - 2x) = \cosh(x) - 2 \leq 0$$

when $0 \leq x \leq 1$. So $\sinh(x) - 2x$ takes its maximum at $x = 0$, which is 0. Thus, given the fact that $\sigma^2 \leq 1$, we have

$$\sinh(\sigma^2 \rho) \leq 6\sigma^2 \rho, \tag{2.31}$$

which completes the proof. $\square$

Proof of Lemma 14. Since $\ddot{f}(x) = \frac{e^x(1-e^{2x})}{(1+e^x)^4} < \frac{e^x}{(1+e^x)^2} = \dot{f}(x)$ for all $x \in \mathbb{R}$, by mean value theorem, for any $a, b \in \mathbb{R}$, we have for some c between a and b,

$$|\dot{f}(a) - \dot{f}(b)| = |a - b|\ddot{f}(c) < |a - b|\dot{f}(c) \leq |a - b| \max\{\dot{f}(a), \dot{f}(b)\}$$

by monotonicity of $\dot{f}(x)$. The rest of the proof follows from

$$\left| \frac{1}{\dot{f}(a)} - \frac{1}{\dot{f}(b)} \right| = \frac{|\dot{f}(a) - \dot{f}(b)|}{\dot{f}(a)\dot{f}(b)}.$$

$\square$

3 Supplementary Tables and Figures of Section 5.2

In Section 5.2 of our main paper, we carried our simulations that compare different methods that control FDR. The design covariates were generated from a truncated multivariate Gaussian distribution, whose covariance matrix is a blockwise diagonal matrix of 10 identical unit diagonal Toeplitz matrices as follows

$$\begin{bmatrix} 1 & \frac{p-2}{10(p-1)} & \frac{p-3}{10(p-1)} & \cdots & \frac{1}{10(p-1)} & 0 \\ \frac{p-2}{10(p-1)} & 1 & \frac{p-2}{10(p-1)} & \cdots & \frac{1}{10(p-1)} & \frac{2}{10(p-1)} \\ \vdots & & \ddots & & & \\ 0 & \frac{1}{10(p-1)} & \frac{2}{10(p-1)} & \cdots & \frac{p-2}{10(p-1)} & 1 \end{bmatrix}.$$

Due to the space limit, we only presented the boxplots for the pooled empirical FDRs across all the settings. As a complement to Figure 2 in the main paper, the case-by-case empirical FDRs are shown in Figure 1.

A Technical Results for Two-Sample Testing

In this section, we discuss the implications of our results on single logistic regression problems to the two-sample settings.

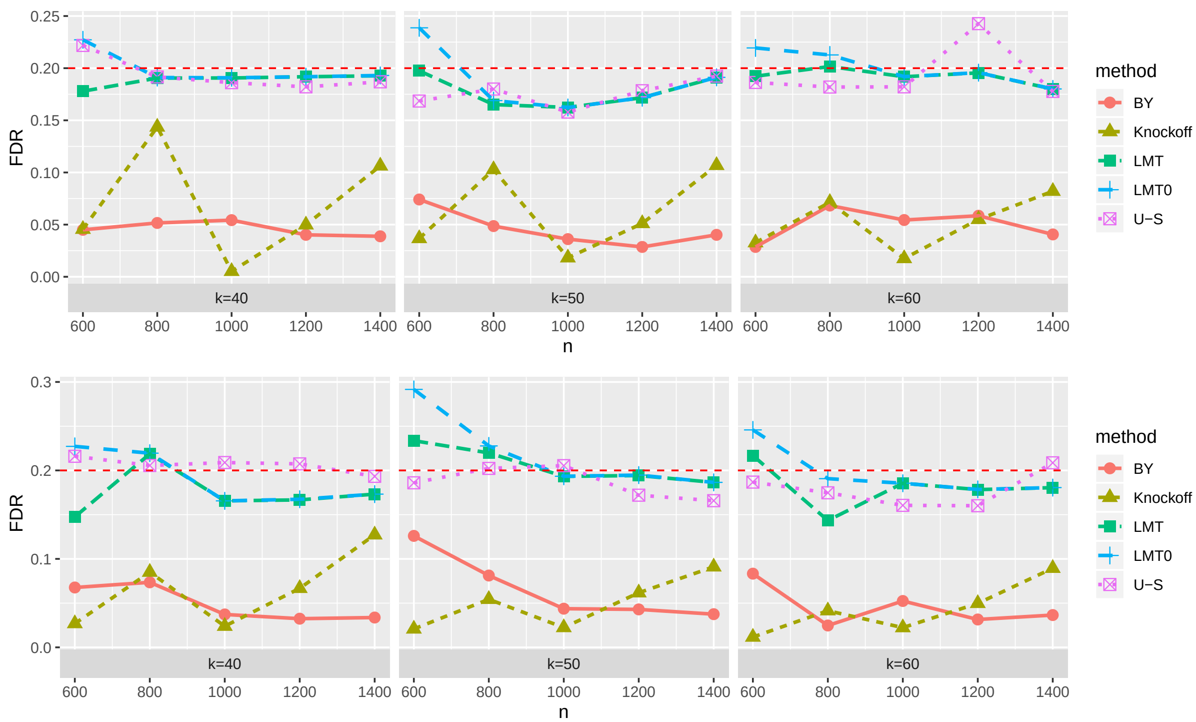

Figure 4: Empirical FDRs under nominal $\alpha = 0.2$ for $\rho = 3$ (top) and $\rho = 4$ (bottom).

A.1 Two-Sample Global Hypothesis Testing

For testing two-sample global null hypothesis

$$H_0 : \beta^{(1)} = \beta^{(2)} \quad \text{vs.} \quad H_1 : \beta^{(1)} \neq \beta^{(2)}.$$

Informed by the previous results, we construct the global two-sample testing procedure as follows. First we obtain $\breve{\beta}_j^{(\ell)}$ and $\tau_j^{(\ell)}$ for each group, and calculate the coordinate-wise standardized statistics

$$T_j = \frac{\breve{\beta}_j^{(1)}}{\sqrt{2}\tau_j^{(1)}} - \frac{\breve{\beta}_j^{(2)}}{\sqrt{2}\tau_j^{(2)}},$$

for $j = 1, ..., p$. Then we calculate the difference the global test statistics is defined as

$$T_n = \max_{1 \leq j \leq p} T_j^2. \tag{A.1}$$

The following corollary states the asymptotic null distribution for the global test statistics M_n under bounded design. In particular, we assume the parameters $(\beta^{(\ell)}, \mathbf{\Sigma}^{(\ell)})$ for $\ell = 1, 2$ come from the same parameter space $\Theta(k)$. We denote $\theta = (\beta^{(1)}, \mathbf{\Sigma}^{(1)}, \beta^{(2)}, \mathbf{\Sigma}^{(2)})$.

Theorem 6. *Let T_n be the test statistics defined in (A.1), $D^{(\ell)}$ be the diagonal of $[\mathbf{\Sigma}^{(\ell)}]^{-1}$ and $(\xi_{ij}^{(\ell)}) = [D^{(\ell)}]^{-1/2}[\mathbf{\Sigma}^{(\ell)}]^{-1}[D^{(\ell)}]^{-1/2}$. Suppose $\max_{1\le i<j\le p}|\xi_{ij}^{(\ell)}| \le c_0 < 1$ for some constant $0 < c_0 < 1$, $\log p = O(n^r)$ for some $0 < r < 1/5$. and*

1. *under the Gaussian design, we assume (A1) (A3) and $k = o\big(\sqrt{n}/\log^3 p\big)$; or*
2. *under the bounded design, we assume (A2) (A3) and $k = o\big(\sqrt{n}/\log^{5/2} p\big)$.*

Then under H_0, for any $x \in \mathbb{R}$, Then under H_0, for any $x \in \mathbb{R}$,

$$P_\theta\big(T_n - 2\log p + \log\log p \le x\big) \to \exp\left(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\right), \quad \textit{as } (n,p)\to\infty.$$

Based on the limiting null distribution, the asymptotically α level tests can be defined as follows:

$$\Phi_\alpha(T_n) = I\{T_n \ge 2\log p - \log\log p + q_\alpha\},$$

where $q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}$. The null hypothesis H_0 is rejected if and only if $\Phi_\alpha(T_n) = 1$.

A.2 Two-Sample Multiple Hypotheses Testing

Consider simultaneously testing the two-sample hypothesis

$$H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)} \quad \text{vs.} \quad H_{1,j} : \beta_j^{(1)} \ne \beta_j^{(2)}, \quad j = 1, ..., p.$$

As a consequence of the previous analysis, we propose the following two-sample multiple testing procedure controlling FDR/FDP or FDV.

Two-Sample FDR/FDP Control Procedure. Define test statistics

$$T_j = (\check\beta_j^{(1)}/\tau_j^{(1)} - \check\beta_j^{(2)}/\tau_j^{(2)})/\sqrt{2},$$

for $j = 1, ..., p$. Let $0 < \alpha < 1$ and define

$$\hat t = \inf\left\{0 \le t \le b_p : \frac{pG(t)}{\max\big\{\sum_{j=1}^p I\{|T_j| \ge t\}, 1\big\}} \le \alpha\right\}. \tag{A.2}$$

We reject $H_{0,j}$ whenever $|M_j| \ge \hat t$.

Two-Sample FDV Control Procedure. For a given tolerable number of falsely discovered variables r, let

$$\hat t_{FDV} = G^{-1}(r/p). \tag{A.3}$$

We reject $H_{0,j}$ whenever $|T_j| \geq \hat{t}_{FDV}$.

The following theorems provide the asymptotic behavior of our proposed testing procedures. For simplicity, we only consider the bounded design scenario.

Theorem 7. *Assume the conditions of Proposition 1 are satisfied for each group of the samples, we have*

$$\lim_{(n,p)\to\infty} \frac{FDR_\theta(\hat{t})}{\alpha p_0/p} \leq 1, \qquad \lim_{(n,p)\to\infty} P_\theta\bigg(\frac{FDP_\theta(\hat{t})}{\alpha p_0/p} \leq 1+\epsilon\bigg) = 1. \tag{A.4}$$

for any $\epsilon > 0$.

Theorem 8. *Suppose the conditions of Theorem 5 are satisfied for each group of the samples, then*

$$\lim_{(n,p)\to\infty} \frac{FDV_\theta(\hat{t}_{FDV})}{rp_0/p} \leq 1. \tag{A.5}$$

A.3 Proofs of the Theorems in Appendix A

In this section, to illustrate how the proofs of the one-sample tests extend to the two-sample tests, we prove Theorem 6 in our Appendix A. The proofs of Theorem 7 and Theorem 8 are similar and thus are omitted.

Proof of Theorem 6 Define $F_{jj}^{(\ell)} = \mathbb{E}[(\eta_{ij}^{(\ell)})^2/\dot{f}(u_i^{(\ell)})]$ and $n \asymp n_1 \asymp n_2$. Define statistics

$$\tilde{M}_j = \frac{\langle v_j^{(1)}, \epsilon^{(1)}\rangle}{\|v_j^{(1)}\|_n} - \frac{\langle v_j^{(2)}, \epsilon^{(2)}\rangle}{\|v_j^{(2)}\|_n}, \qquad \check{M}_j = \frac{\sum_{i=1}^{n_1} \eta_{ij}^{(1)}\epsilon_i^{(1)}/\dot{f}(u_i^{(1)})}{\sqrt{n_1 F_{jj}^{(1)}}} - \frac{\sum_{i=1}^{n_2} \eta_{ij}^{(2)}\epsilon_i^{(2)}/\dot{f}(u_i^{(2)})}{\sqrt{n_2 F_{jj}^{(2)}}},$$

for $j = 1, ..., p$, and thereby $\tilde{M}_n = \max_j \tilde{M}_j^2, \check{M}_n = \max_j \check{M}_j^2$.

Lemma 15. *Under the condition of Theorem 6, the following events*

$$B_1 = \Big\{|\tilde{M}_n - \check{M}_n| = o(1)\Big\}, \quad B_2 = \Big\{|\tilde{M}_n - M_n| = o\Big(\frac{1}{\log p}\Big)\Big\},$$

hold with probability at least $1 - O(p^{-c})$ for some constant $C, c > 0$.

The proof of the above lemma follows directly from the proof of Lemma 1.

It follows that under the event $B_1 \cap B_2$, let $y_p = 2\log p - \log\log p + x$ and $\epsilon_n = o(1)$, we have

$$P_\theta\big(\check{M}_n \leq y_p - \epsilon_n\big) \leq P_\theta\big(M_n \leq y_p\big) \leq P_\theta\big(\check{M}_n \leq y_p + \epsilon_n\big)$$

Therefore it suffices to prove that for any $t \in \mathbb{R}$, as $(n,p) \to \infty$,

$$P_\theta\big(\check{M}_n \leq y_p\big) \to \exp\bigg(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\bigg). \tag{A.6}$$

Now define

$$\hat{M}_j = \frac{\sum_{i=1}^{n_1} \hat{Z}_{ij}^{(1)}}{\sqrt{n_1 F_{jj}^{(1)}}} - \frac{\sum_{i=1}^{n_2} \hat{Z}_{ij}^{(2)}}{\sqrt{n_2 F_{jj}^{(2)}}} \qquad j = 1, ..., p.$$

where $\hat{Z}_{ij}^{(\ell)} = v_{ij}^{0,\ell}\epsilon_i^{(\ell)} 1\{|v_{ij}^{0,\ell}\epsilon_i^{(\ell)}| \le \tau_n\} - \mathbb{E}[v_{ij}^{0,\ell}\epsilon_i^{(\ell)} 1\{|v_{ij}^{0,\ell}\epsilon_i^{(\ell)}| \le \tau_n\}]$ for $\tau_n = \log p$, $v_{ij}^{0,\ell} = \eta_{ij}^{(\ell)}/\dot{f}(u_i^{(\ell)})$ and $\hat{M}_n = \max_j \hat{M}_j^2$. Equivalently, we can write

$$\hat{M}_j = \frac{1}{n_1} \sum_{i=1}^{n_1+n_2} w_{i_j} \qquad j = 1, ..., p.$$

where

$$w_{i_j} = \frac{\hat{Z}_{ij}^{(1)}}{\sqrt{F_{jj}^{(1)}}}, \qquad \text{for } i = 1, ..., n_1,$$

and

$$w_{i_j} = \sqrt{\frac{n_1}{n_2}} \frac{\hat{Z}_{ij}^{(2)}}{\sqrt{F_{jj}^{(2)}}}, \qquad \text{for } i = n_1 + 1, ..., n_1 + n_2.$$

By similar statement in Lemma 2, it suffices to prove that for any $t \in \mathbb{R}$, as $(n, p) \to \infty$,

$$P_\theta\big(\hat{M}_n \le y_p\big) \to \exp\left(- \frac{1}{\sqrt{\pi}} \exp(-x/2)\right). \tag{A.7}$$

By Lemma 3 in the main paper, for any integer $0 < q < p/2$,

$$\begin{aligned} \sum_{d=1}^{2q} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\left(\bigcap_{k=1}^{d} A_{j_k} \right) &\le P_\theta\left(\max_{1 \le j \le p} \hat{M}_j^2 \ge y_p \right) \\ &\le \sum_{d=1}^{2p-1} (-1)^{d-1} \sum_{1 \le j_1 < ... < j_d \le p} P_\theta\left(\bigcap_{k=1}^{d} A_{j_k} \right), \end{aligned} \tag{A.8}$$

where $A_{j_k} = \{\hat{M}_{j_k}^2 \ge y_p\}$. Now let $\mathbf{W}_i = (w_{i,j_1}, ..., w_{i,j_d})^T$ for $1 \le i \le n_1 + n_2$. Define $\|\mathbf{a}\|_{\min} = \min_{1 \le i \le d} |a_i|$ for any vector $\mathbf{a} \in \mathbb{R}^d$. Then we have

$$P_\theta\left(\bigcap_{k=1}^{d} A_{j_k} \right) = P_\theta\left(\left\| n_1^{-1/2} \sum_{i=1}^{n_1+n_2} \mathbf{W}_i \right\|_{\min} \ge y_p^{1/2} \right).$$

Then it follows from Theorem 1.1 in Zaitsev (1987) that

$$P_\theta\bigg(\bigg\|n_1^{-1/2}\sum_{i=1}^{n_1+n_2}\mathbf{W}_i\bigg\|_{\min}\geq y_p^{1/2}\bigg)\leq P_\theta\Big(\|\mathbf{N}_d\|_{\min}\geq y_p^{1/2}-\epsilon_n(\log p)^{-1/2}\Big)$$
$$+\,c_1 d^{5/2}\exp\bigg\{-\frac{n_1^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\bigg\}, \tag{A.9}$$

where $c_1>0$ and $c_2>0$ are constants, $\epsilon_n\to 0$ which will be specified later, and $\mathbf{N}_d=(N_{m_1},...,N_{m_d})$ is a normal random vector with $\mathbb{E}(\mathbf{N}_d)=0$ and $\mathrm{cov}(\mathbf{N}_d)=\mathrm{cov}(\mathbf{W}_1)+n_2/n_1\mathrm{cov}(\mathbf{W}_{n_1+1})$. Here d is a fixed integer that does not depend on n,p. Because $\log p=o(n^{1/5})$, we can let $\epsilon_n\to 0$ sufficiently slowly, say, $\epsilon_n=\sqrt{\frac{\log^5 p}{n}}$, so that for any large $c>0$,

$$c_1 d^{5/2}\exp\bigg\{-\frac{n^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\bigg\}=O(p^{-c}). \tag{A.10}$$

Combining (A.8), (A.9) and (A.10), we have

$$P_\theta\Big(\max_{1\leq j\leq p}\hat{M}_j^2\geq y_p\Big)\leq\sum_{d=1}^{2p-1}(-1)^{d-1}\sum_{1\leq j_1<\ldots<j_d\leq p}P_\theta\Big(\|\mathbf{N}_d\|_{\min}\geq y_p^{1/2}-\epsilon_n(\log p)^{-1/2}\Big)+o(1). \tag{A.11}$$

Similarly, one can derive

$$P_\theta\Big(\max_{1\leq j\leq p}\hat{M}_j^2\geq y_p\Big)\geq\sum_{d=1}^{2p}(-1)^{d-1}\sum_{1\leq j_1<\ldots<j_d\leq p}P_\theta\Big(\|\mathbf{N}_d\|_{\min}\geq y_p^{1/2}+\epsilon_n(\log p)^{-1/2}\Big)+o(1). \tag{A.12}$$

Using Lemma 4 in the main paper, it then follows from (A.11) and (A.12) that

$$\varlimsup_{n,p\to\infty}P_\theta\Big(\max_{1\leq j\leq p}\hat{M}_j^2\geq y_p\Big)\leq\sum_{d=1}^{2p}(-1)^{d-1}\frac{1}{d!}\Big(\frac{1}{\sqrt{\pi}}\exp(-t/2)\Big)^d,$$
$$\varliminf_{n,p\to\infty}P_\theta\Big(\max_{1\leq j\leq p}\hat{M}_j^2\geq y_p\Big)\geq\sum_{d=1}^{2p-1}(-1)^{d-1}\frac{1}{d!}\Big(\frac{1}{\sqrt{\pi}}\exp(-t/2)\Big)^d,$$

for any positive integer p. By letting $p\to\infty$, we obtain (A.7) and the proof is complete. □