LINEAR HYPOTHESIS TESTING FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS

By Chengchun Shi*, Rui Song*, Zhao Chen†, and Runze Li†

North Carolina State University and Pennsylvania State University

This paper is concerned with testing linear hypotheses in high-dimensional generalized linear models. To deal with linear hypotheses, we first propose constrained partial regularization method and study its statistical properties. We further introduce an algorithm for solving regularization problems with folded-concave penalty functions and linear constraints. To test linear hypotheses, we propose a partial penalized likelihood ratio test, a partial penalized score test and a partial penalized Wald test. We show that the limiting null distributions of these three test statistics are χ^2 distribution with the same degrees of freedom, and under local alternatives, they asymptotically follow non-central χ^2 distributions with the same degrees of freedom and noncentral parameter, provided the number of parameters involved in the test hypothesis grows to ∞ at a certain rate. Simulation studies are conducted to examine the finite sample performance of the proposed tests. Empirical analysis of a real data example is used to illustrate the proposed testing procedures.

1. Introduction. During the last three decades, there are many works devoted to developing variable selection techniques for high dimensional regression models. Fan and Lv (2010) presents a selective overview on this topic. There are some recent works for hypothesis testing on Lasso estimate (Tibshirani, 1996) in high-dimensional linear models. Lockhart et al. (2014) proposed the covariance test which produces a sequence of p-values as the tuning parameter, λ_n, decreases, and features become non-zero in the Lasso. This approach does not give confidence intervals or p-values for an individual variable's coefficient. Taylor et al. (2014) and Lee et al. (2016) extended the covariance testing framework to test hypotheses about individual features, after conditioning on a model selected by the Lasso. However, their framework permits inference only about features which have non-zero coefficients in a Lasso regression; this set of features likely varies across samples, making the interpretation difficult. Moreover, these work focused on high di-

*Supported by NSF grant DMS 1555244, NCI grant P01 CA142538.

†Supported by NSF grant DMS 1512422, NIH grants P50 DA039838 and P50 DA036107, and T32 LM012415, and NNSFC grants 11690014 and 11690015.

Keywords and phrases: High-dimensional testing, Linear hypothesis, Likelihood ratio statistics, Score test, Wald test

mensional linear regression models, and it remains unknown whether their results can be extended to a more general setting.

This paper will focus on generalized linear models (GLM, McCullagh and Nelder, 1989). Let Y be the response, and $\boldsymbol{X}$ be its associate fixed-design covariate vector. The GLM assumes that the distribution of Y belongs to the exponential family. The exponential family with canonical link has the following probability density function

$$\exp\left(\frac{Y\boldsymbol{\beta}_0^T\boldsymbol{X} - b(\boldsymbol{\beta}_0^T\boldsymbol{X})}{\phi_0}\right)c(Y), \tag{1.1}$$

where $\boldsymbol{\beta}_0$ is a p-dimensional vector of regression coefficients, and ϕ_0 is some positive nuisance parameter. In this paper, we assume that $b(\cdot)$ is thrice continuously differentiable with $b''(\cdot) > 0$.

We study testing linear hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{t}$ in GLM, where $\boldsymbol{\beta}_{0,\mathcal{M}}$ is a subvector of $\boldsymbol{\beta}_0$, the true regression coefficients. The number of covariates p can be much larger than the sample size n, while the number of parameters in $\boldsymbol{\beta}_{0,\mathcal{M}}$ is assumed to be much smaller than n. Such type of hypotheses is of particular interests when the goal is to explore the group structure of $\boldsymbol{\beta}_0$. Moreover, it also includes a very important class of hypotheses $\boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$ by setting $\boldsymbol{C}$ to be the identity matrix and $\boldsymbol{t} = \mathbf{0}$. In the literature, Fan and Peng (2004) proposed penalized likelihood ratio test for $H_{0a} : \boldsymbol{C}\boldsymbol{\beta}_{0,S} = \mathbf{0}$ in GLM, where $\boldsymbol{\beta}_{0,S}$ is the vector consisting of all nonzero elements of $\boldsymbol{\beta}_0$ when $p = o(n^{1/5})$ where n stands for the sample size. Wang and Cui (2013) extended Fan and Peng (2004)'s proposal and considered a penalized likelihood ratio statistic for testing $H_{0b} : \boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$, requiring $p = o(n^{1/5})$. Ning and Liu (2017) proposed a decorrelated score test for $H_{0c} : \boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$ under the setting of high dimensional penalized M-estimators with nonconvex penalties. Recently, Fang, Ning and Liu (2017) extended the proposal of Ning and Liu (2017) and developed a class of decorrelated Wald, score and partial likelihood ratio tests for Cox's model with high dimensional survival data. Zhang and Cheng (2017) proposed a maximal type statistic based on the desparsified Lasso estimator (van de Geer et al., 2014) and a bootstrap-assisted testing procedure for $H_{0d} : \boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$, allowing the cardinality of $\mathcal{M}$ to be an arbitrary subset of $[1, \ldots, p]$. In this paper, we aim to develop theory of Wald test, score test and likelihood ratio test for $H_0 : \boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{t}$ in GLM under ultrahigh dimensional setting (i.e., p grows exponentially with n).

It is well known that the Wald, score and likelihood ratio tests are equivalent in the fixed p case. However, it can be challenging to generalize these statistics to the setting with ultrahigh dimensionality. To better understand

this point, we take the Wald statistic for illustration. Consider the null hypothesis $H_0 : \boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$. Analogous to the classical Wald statistic, in the high dimensional setting, one might consider the statistic $\hat{\boldsymbol{\beta}}_{\mathcal{M}}^T\{\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\}^{-1}\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ for some penalized regression estimator $\hat{\boldsymbol{\beta}}$ and its variance estimator $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})$. The choice of the estimators is essential here: some penalized regression estimator such as the Lasso, or the Dantzig estimator (Candes and Tao, 2007) cannot be used due to their large biases when $p \gg n$. The non-concave penalized estimator does not have this bias issue, but the minimal signal conditions imposed in Fan and Peng (2004) and Fan and Lv (2011) implies that the associated Wald statistic does not have any power for local alternatives of the type $H_a : \boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{h}_n$ for some sequence $\boldsymbol{h}_n$ such that $\|\boldsymbol{h}_n\|_2 \ll \lambda_n$ where $\|\cdot\|_2$ is the Euclidean norm. Moreover, to implement the score and the likelihood ratio statistics, we need to estimate the regression parameter under the null, which involves penalized likelihood under linear constraints. This is a very challenging task and has rarely been studied: (a) the associated estimation and variable selection property is not standard from a theoretical perspective, and (b) there is a lack of constrained optimization algorithms that can produce sparse estimators from a computational perspective.

We briefly summarize our contributions as follows. First, we consider a more general form of hypothesis. In contrast, existing literature mainly focuses on testing $\boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$. Besides, we also allow the number of linear constraints to diverge with n. Our tests are therefore applicable to a wider range of real applications for testing a growing set of linear hypotheses. Second, we propose a partial penalized Wald, a partial penalized score and a partial penalized likelihood-ratio statistic based on the class of folded-concave penalty functions, and show their equivalence in the high dimensional setting. We derive the asymptotic distributions of our test statistics under the null hypothesis and the local alternatives. Third, we systematically study the partial penalized estimator with linear constraints. We derive its rate of convergence and limiting distribution. These results are significant in their own rights. The unconstrained and constrained estimators share similar forms, but the constrained estimator is more efficient due to the additional information contained in the constraints under the null hypothesis. Fourth, we introduce an algorithm for solving regularization problems with folded-concave penalty functions and equality constraints, based on the alternating direction method of multipliers (ADMM, cf. Boyd et al., 2011).

The rest of the paper is organized as follows. We study the statistical properties of the constrained partial penalized estimator with folded concave penalty functions in Section 2. We formally define our partial penalized Wald, score and likelihood-ratio statistics, establish their limiting distribu-

tions, and show their equivalence in Section 3. Detailed implementations of our testing procedures are given in Section 3.3, where we introduce our algorithm for solving the constrained partial penalized regression problems. Simulation studies are presented in Section 4. The proof of Theorem 3.1 is presented in Section 5. Other proofs and addition numerical results are presented in the supplementary material (Shi et al., 2018).

2. Constrained partial penalized regression.

2.1. *Model setup.* Suppose that $\{\boldsymbol{X}_i, Y_i\}$, $i = 1, \cdots, n$ is a sample from model (1.1). Denote by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ the n-dimensional response vector and $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$ is the $n \times p$ design matrix. We assume the covariates $\boldsymbol{X}_i$ are fixed design. Let $\boldsymbol{X}^j$ denote the jth column of $\boldsymbol{X}$. To simplify the presentation, for any $r \times q$ matrix $\boldsymbol{\Phi}$ and any set $J \subset [1, 2, \ldots, q]$, we denote by $\boldsymbol{\Phi}_J$ the submatrix of $\boldsymbol{\Phi}$ formed by columns in J. Similarly, for any q-dimensional vector $\boldsymbol{\phi}$, $\boldsymbol{\phi}_J$ stands for the subvector of $\boldsymbol{\phi}$ formed by elements in J. We further denote $\boldsymbol{\Phi}_{J_1,J_2}$ as the submatrix of $\boldsymbol{\Phi}$ formed by rows in J_1 and columns in J_2 for any $J_1 \subseteq [1, \ldots, r]$ and $J_2 \subseteq [1, \ldots, q]$. Let $|J|$ be the number of elements in J. Define $J^c = [1, \ldots, q] - J$ to be the complement of J.

In this paper, we assume $\log p = O(n^a)$ for some $0 < a < 1$ and focus on the following testing problem:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{t}, \tag{2.1}$$

for a given $\mathcal{M} \subseteq [1, \ldots, p]$, an $r \times |\mathcal{M}|$ matrix $\boldsymbol{C}$ and an r-dimensional vector $\boldsymbol{t}$. We assume that the matrix $\boldsymbol{C}$ is of full row rank. This implies there are no redundant or contradictory constraints in (2.1). Let $m = |\mathcal{M}|$, we have $r \leq m$.

Define the partial penalized likelihood function

$$Q_n(\boldsymbol{\beta}, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i\boldsymbol{\beta}^T\boldsymbol{X}_i - b(\boldsymbol{\beta}^T\boldsymbol{X}_i)\} - \sum_{j \notin \mathcal{M}} p_\lambda(|\beta_j|).$$

for some penalty function $p_\lambda(\cdot)$ with a tuning parameter λ. Further define

$$\hat{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}, \lambda_{n,0}) \text{ subject to } \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}, \tag{2.2}$$

$$\hat{\boldsymbol{\beta}}_a = \arg\max_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}, \lambda_{n,a}). \tag{2.3}$$

Note that in (2.2) and (2.3), we do not add penalties on parameters involved in the constraints. This enables to avoid imposing minimal signal condition

on elements of $\boldsymbol{\beta}_{0,\mathcal{M}}$. Thus, the corresponding likelihood ratio test, Wald test and score test have power at local alternatives.

We present a lemma to characterize the constrained local maximizer $\hat{\boldsymbol{\beta}}_0$ in the supplementary material (see Lemma S.1). In Section 3, we show that these partial penalized estimators help us to obtain valid statistical inference about the null hypothesis.

2.2. *Partial penalized regression with linear constraint.* In this section, we study the statistical properties of $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_a$ by restricting p_λ to the class of folded concave penalty functions. Popular penalty functions such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) belong to this class. Let $\rho(t_0,\lambda) = p_\lambda(t_0)/\lambda$ for $\lambda > 0$. We assume that $\rho(t_0,\lambda)$ is increasing and concave in $t_0 \in [0,\infty)$, and has a continuous derivative $\rho'(t_0,\lambda)$ with $\rho'(0+,\lambda) > 0$. In addition, assume $\rho'(t_0,\lambda)$ is increasing in $\lambda \in (0,\infty)$ and $\rho'(0+,\lambda)$ is independent of λ. For any vector $\boldsymbol{v} = (v_1,\ldots,v_q)^T$, define

$$\bar{\boldsymbol{\rho}}(\boldsymbol{v},\lambda) = \{\mathrm{sgn}(v_1)\rho'(|v_1|,\lambda),\cdots,\mathrm{sgn}(v_q)\rho'(|v_q|,\lambda)\}^T,$$
$$\boldsymbol{\mu}(\boldsymbol{v}) = \{b'(v_1),\ldots,b'(v_q)\}, \quad \boldsymbol{\Sigma}(\boldsymbol{v}) = \mathrm{diag}\{b''(v_1),\ldots,b''(v_q)\},$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function. We further define the local concavity of the penalty function ρ at $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_0 = q$ as

$$\kappa(\rho,\boldsymbol{v},\lambda) = \lim_{\epsilon\to 0^+}\max_{1\le j\le q}\sup_{t_1<t_2\in(|v_j|-\epsilon,|v_j|+\epsilon)} -\frac{\rho'(t_2,\lambda)-\rho'(t_1,\lambda)}{t_2-t_1}.$$

We assume that the true regression coefficient $\boldsymbol{\beta}_0$ is sparse and satisfies $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} - \boldsymbol{t} = \boldsymbol{h}_n$ for some sequence of vectors $\boldsymbol{h}_n \to \boldsymbol{0}$. When $\boldsymbol{h}_n = \boldsymbol{0}$, the null holds. Otherwise, the alternative holds. Let $S = \{j \in \mathcal{M}^c : \beta_{0,j} \neq 0\}$ and $s = |S|$. Let d_n be the half minimum signal of $\boldsymbol{\beta}_{0,S}$, i.e, $d_n = \min_{j\in S}|\beta_j|/2$. Define $\mathcal{N}_0 = \{\boldsymbol{\beta}\in\mathbb{R}^p : \|\boldsymbol{\beta}_{S\cup\mathcal{M}} - \boldsymbol{\beta}_{0,S\cup\mathcal{M}}\|_2 \le \sqrt{(s+m)\log(n)/n}, \boldsymbol{\beta}_{(S\cup\mathcal{M})^c} = 0\}$. We impose the following conditions.

(A1) Assume that

$$\max_{1\le j\le p}\|\boldsymbol{X}^j\|_\infty = O\left(\sqrt{n/\log(p)}\right), \qquad \max_{1\le j\le p}\|\boldsymbol{X}^j\|_2 = O(\sqrt{n}),$$
$$\inf_{\boldsymbol{\beta}\in\mathcal{N}_0}\lambda_{\min}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{S\cup\mathcal{M}}\right) \ge cn,$$
$$\lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O(n),$$
$$\|\boldsymbol{X}_{(S\cup\mathcal{M})^c}^T\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{\beta}_0)\boldsymbol{X}_{S\cup\mathcal{M}}\|_{2,\infty} = O(n),$$
$$\max_{1\le j\le p}\sup_{\boldsymbol{\beta}\in\mathcal{N}_0}\lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\mathrm{diag}\{|\boldsymbol{X}^j|\circ|b'''(\boldsymbol{X}\boldsymbol{\beta})|\}\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O(n),$$

for some constant $c > 0$, where for any vector $\boldsymbol{v} = (v_1, \ldots, v_q)^T$, $\mathrm{diag}(\boldsymbol{v})$ denotes a diagonal matrix with the j-th diagonal elements being v_j, $|\boldsymbol{v}| = (|v_1|, \ldots, |v_q|)^T$, and $\|\boldsymbol{B}\|_{2,\infty} = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} \|\boldsymbol{B}\boldsymbol{v}\|$ for any matrix $\boldsymbol{B}$ with q rows.

(A2) Assume that $d_n \gg \lambda_{n,j} \gg \max\{\sqrt{(s+m)/n}, \sqrt{(\log p)/n}\}$, $p'_{\lambda_{n,j}}(d_n) = o((s+m)^{-1/2} n^{-1/2})$, $\lambda_{n,j}\kappa_{0,j} = o(1)$ where $\kappa_{0,j} = \max_{\beta \in \mathcal{N}_0} \kappa(\rho, \boldsymbol{\beta}, \lambda_{n,j})$, for $j = 0, a$.

(A3) Assume that there exist some constants M and v_0 such that

$$\max_{1 \le i \le n} \mathrm{E}\left\{ \exp\left(\frac{|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|}{M} \right) - 1 - \frac{|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|}{M} \right\} M^2 \le \frac{v_0}{2}.$$

(A4) Assume that $\|\boldsymbol{h}_n\|_2 = O\left(\sqrt{\min(s+m-r, r)/n}\right)$, and $\lambda_{\max}\left((\boldsymbol{C}\boldsymbol{C}^T)^{-1}\right) = O(1)$.

In Section S4.1 of the supplementary material, we show that Condition (A1) holds with probability tending to 1 if the covariate vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are uniformly bounded or realizations from a sub-Gaussian distribution. The first condition in (A2) is a minimum signal assumption imposed on nonzero elements in $\mathcal{M}^c$ only. This is due to partial penalization, which enables us to evaluate the uncertainty of the estimation for small signals. Such conditions are not assumed in van de Geer et al. (2014) and Ning and Liu (2017) for testing $H_0 : \boldsymbol{\beta}_{0,\mathcal{M}} = 0$. However, we note that these authors impose some additional assumptions on the design matrix. For example, the validity of the decorrelated score statistic depends on the sparsity of $\mathbf{w}^*$. For testing univariate parameters, this requires the degree of a particular node in the graph to be relatively small when the covariate follows a Gaussian graphical model (see Remark 6 in Ning and Liu, 2017). In Section S4.3 of the supplementary material, we show Condition (A3) holds for linear, logistic, and Poisson regression models.

THEOREM 2.1. *Suppose that Conditions (A1)-(A4) hold, and $s + m = o(\sqrt{n})$, then the following holds: (i) With probability tending to 1, $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_a$ defined in (2.2) and (2.3) must satisfy $\hat{\boldsymbol{\beta}}_{0,(S\cup\mathcal{M})^c} = \hat{\boldsymbol{\beta}}_{a,(S\cup\mathcal{M})^c} = 0$. (ii) $\|\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{a,S\cup\mathcal{M}}\|_2 = O_p(\sqrt{(s+m)/n})$ and $\|\hat{\boldsymbol{\beta}}_{0,S\cup\mathcal{M}} - \boldsymbol{\beta}_{0,S\cup\mathcal{M}}\|_2 = O_p(\sqrt{(s+m-r)/n})$. If further $s + m = o(n^{1/3})$, then we have*

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} = \frac{1}{\sqrt{n}} \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + o_p(1),$$

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix} = \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}$$
$$- \sqrt{n}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\\ \boldsymbol{0}\end{pmatrix}+o_p(1),$$

where $\boldsymbol{I}$ *is the identity matrix,* $\boldsymbol{K}_n$ *is the* $(m+s)\times(m+s)$ *matrix*

$$\boldsymbol{K}_n=\frac{1}{n}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{S}}\\ \boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{S}}\end{pmatrix},$$

and $\boldsymbol{P}_n$ *is the* $(m\times s)\times(m\times s)$ *projection matrix*

$$\boldsymbol{P}_n=\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\left\{(\boldsymbol{C}\,\boldsymbol{O}_{r\times s})\,\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\right\}^{-1}(\boldsymbol{C}\,\boldsymbol{O}_{r\times s})\,\boldsymbol{K}_n^{-1/2},$$

where $\boldsymbol{O}_{r\times s}$ *is an* $r\times s$ *zero matrix.*

REMARK 2.1. Since $d_n \gg \sqrt{(s+m)/n}$, Theorem 2.1(ii) implies that each element in $\hat{\boldsymbol{\beta}}_{0,S}$ and $\hat{\boldsymbol{\beta}}_{a,S}$ is nonzero. This together with result (i) shows the sign consistency of $\hat{\boldsymbol{\beta}}_{0,\mathcal{M}^c}$ and $\hat{\boldsymbol{\beta}}_{a,\mathcal{M}^c}$.

REMARK 2.2. Theorem 2.1 implies that the constrained estimator $\hat{\boldsymbol{\beta}}_0$ converges at a rate of $O_p(\sqrt{s+m-r}/\sqrt{n})$. In contrast, the unconstrained estimator converges at a rate of $O_p(\sqrt{s+m}/\sqrt{n})$. This suggests that when $\boldsymbol{h}_n$ is relatively small, the constrained estimator $\hat{\boldsymbol{\beta}}_0$ converges faster than the unconstrained $\hat{\boldsymbol{\beta}}_a$ defined in (2.3), when $s+m-r \ll s+m$. This result is expected with the following intuition: the more information about $\boldsymbol{\beta}_0$ we have, the more accurate the estimator will be.

REMARK 2.3. Under certain regularity conditions, Theorem 2.1 implies that

$$\sqrt{n}\{(\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}})^T,(\hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S})^T\}\to N(-\boldsymbol{\xi}_0,\boldsymbol{V}_0),$$

where $\boldsymbol{\xi}_0$ and $\boldsymbol{V}_0$ are limits of $\sqrt{n}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}(\boldsymbol{h}_n^T,\boldsymbol{0}^T)^T$ and $\boldsymbol{K}_n^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}$, respectively. Similarly, we can show

$$\sqrt{n}\{(\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}})^T,(\hat{\boldsymbol{\beta}}_{a,S}-\boldsymbol{\beta}_{0,S})^T\}\to N(\boldsymbol{0},\boldsymbol{V}_a),$$

where $\boldsymbol{V}_a=\lim_n \boldsymbol{K}_n^{-1}$. Note that $\boldsymbol{a}^T\boldsymbol{V}_0\boldsymbol{a}\le\boldsymbol{a}^T\boldsymbol{V}_a\boldsymbol{a}$ for any $\boldsymbol{a}\in\mathbb{R}^{s+m}$. Under the null, we have $\boldsymbol{\xi}_0=\boldsymbol{0}$, which suggests that $\hat{\boldsymbol{\beta}}_0$ is more efficient than $\hat{\boldsymbol{\beta}}_a$ in terms of a smaller asymptotic variance. Under the alternative, $\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}$ is asymptotically biased. This can be interpreted as a bias-variance trade-off between $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_a$.

3. Partial penalized Wald, score and likelihood ratio statistics.

3.1. *Test statistics.* We begin by introducing our partial penalized likelihood ratio statistic,

$$T_L = 2n\{L_n(\hat{\boldsymbol{\beta}}_a) - L_n(\hat{\boldsymbol{\beta}}_0)\}/\hat{\phi}, \tag{3.1}$$

where $L_n(\boldsymbol{\beta}) = \sum_i \{Y_i \boldsymbol{\beta}^T \boldsymbol{X}_i - b(\boldsymbol{\beta}^T \boldsymbol{X}_i)\}/n$, $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_a$ are defined in (2.2) and (2.3) respectively, and $\hat{\phi}$ is some consistent estimator for ϕ_0 in (1.1). For Gaussian linear models, ϕ_0 corresponds to the error variance. For logistic or Poisson regression models, $\phi_0 = 1$.

The partial penalized Wald statistic is based on $\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})$. Define $\boldsymbol{\Omega}_n = \boldsymbol{K}_n^{-1}$, and denote $\boldsymbol{\Omega}_{mm}$ as the first m rows and columns of $\boldsymbol{\Omega}_n$. It follows from Theorem 2.1 that its asymptotic variance is equal to $\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$. Let $\widehat{S}_a = \{j \in \mathcal{M}^c : \hat{\beta}_{a,j} \neq 0\}$. Then, with probability tending to 1, we have $\widehat{S}_a = S$. Define

$$\widehat{\boldsymbol{\Omega}}_a = n \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\widehat{S}_a} \\ \boldsymbol{X}_{\widehat{S}_a}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\widehat{S}_a}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\widehat{S}_a} \end{pmatrix}^{-1},$$

and $\widehat{\boldsymbol{\Omega}}_{a,mm}$ as its submatrix formed by its first m rows and columns. The partial penalized Wald statistic is defined by

$$T_W = (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})^T \left(\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})/\hat{\phi}. \tag{3.2}$$

Analogous to the classical score statistic, we define our partial penalized score statistic as

$$T_S = \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}} \\ \boldsymbol{X}_{\widehat{S}_0} \end{pmatrix} \widehat{\boldsymbol{\Omega}}_0 \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}} \\ \boldsymbol{X}_{\widehat{S}_0} \end{pmatrix}^T \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}/\hat{\phi}, \tag{3.3}$$

where $\widehat{S}_0 = \{j \in \mathcal{M}^c : \hat{\beta}_{0,j} \neq 0\}$, and

$$\widehat{\boldsymbol{\Omega}}_0 = n \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\widehat{S}_0} \\ \boldsymbol{X}_{\widehat{S}_0}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\widehat{S}_0}^T \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\widehat{S}_0} \end{pmatrix}^{-1}.$$

3.2. *Limiting distributions of the test statistics.* For a given significance level α, we reject the null hypothesis when $T > \chi^2_\alpha(r)$ for $T = T_L, T_W$ or T_S where $\chi^2_\alpha(r)$ is the upper α-quantile of a central χ^2 distribution with r degrees of freedom and r is the number of constraints. Assume r is fixed. When $\hat{\phi}$ is consistent to ϕ_0, it follows from Theorem 2.1 that T_L, T_W and T_S

converge asymptotically to a (non-central) χ^2 distribution with r degrees of freedom. However, when r diverges with n, there is no such theoretical guarantee. This is because the concept of weak convergence is not well defined in such settings. To resolve this issue, we observe that when the following holds,

$$\sup_x |\Pr(T \leq x) - \Pr(\chi^2(r, \gamma_n) \leq x)| \to 0,$$

where $\chi^2(r, \gamma_n)$ is a chi square random variable with r degrees of freedom and noncentrality parameter γ_n which is allowed to vary with n, our testing procedure is still valid using χ^2 approximation.

THEOREM 3.1. *Assume Conditions (A1)-(A4) hold,* $s + m = o(n^{1/3})$, *and* $|\hat{\phi} - \phi_0| = o_p(1)$. *Further assume the following holds:*

$$\frac{r^{1/4}}{n^{3/2}} \sum_{i=1}^{n} \left\{ (\boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}})^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}} \right\}^{3/2} \to 0. \tag{3.4}$$

Then, we have

$$\sup_x \left| Pr(T \leq x) - Pr(\chi^2(r, \gamma_n) \leq x) \right| \to 0, \tag{3.5}$$

for $T = T_W, T_S$ *or* T_L, *where* $\gamma_n = n\boldsymbol{h}_n^T \left(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T \right)^{-1} \boldsymbol{h}_n/\phi_0$.

REMARK 3.1. By (3.5), it is immediate to see that

$$\sup_x |\Pr(T_1 \leq x) - \Pr(T_2 \leq x)| \to 0,$$

for any $T_1, T_2 \in \{T_W, T_S, T_L\}$. This establish the equivalence between the partial penalized Wald, score and likelihood-ratio statistics. Condition (3.4) is the key to guarantee χ^2 approximation in (3.5). When $r = O(1)$, this condition is equivalent to

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \left\{ (\boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}})^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}} \right\}^{3/2} \to 0,$$

which corresponds to the Lyaponuv condition that ensures the asymptotic normality of $\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}$ and $\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S}$. When r diverges, (3.4) guarantees that the following Lyaponuv type bound goes to 0,

$$\sup_{\mathcal{C}} |\Pr\left((\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1/2} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})/\sqrt{\phi}_0 \in \mathcal{C} \right) - \Pr(\boldsymbol{Z} \in \mathcal{C})| \to 0,$$

where $\boldsymbol{Z}$ represents an r-dimensional multivariate normal with identity covariance matrix, and the supremum is taken over all convex subsets $\mathcal{C}$ in $\mathbb{R}^m$. The scaling factor $r^{1/4}$ accounts for the dependence of the above Lyaponuv type estimate on the dimension and it remains unknown whether the factor $r^{1/4}$ can be improved (see related discussions in Bentkus, 2004).

REMARK 3.2. Theorem 3.1 implies that our testing procedures are consistent. When the null holds, we have $\boldsymbol{h}_n = \mathbf{0}$ and hence $\gamma_n = 0$. This together with equation (3.5) suggests that our tests have correct size under the null. Under the alternative, we have $\boldsymbol{h}_n \neq \mathbf{0}$ and hence $\gamma_n \neq 0$. Since $\chi^2(r, 0)$ is stochastically smaller than $\chi^2(r, \gamma_n)$, (3.5) implies that our tests have non-negligible powers under H_a. We summarize these results in the following corollary.

COROLLARY 3.1. *Assume Conditions (A1)-(A3) and (3.4) hold,* $s + m = o(n^{1/3})$, $\lambda_{\max}((\boldsymbol{C}\boldsymbol{C}^T)^{-1}) = O(1)$, *and* $|\hat{\phi} - \phi_0| = o_p(1)$. *Then, under the null hypothesis, for any* $0 < \alpha < 1$, *we have*

$$\lim_n Pr(T > \chi^2_\alpha(r)) = \alpha,$$

for $T = T_W, T_L$ *and* T_S, *where* $\chi^2_\alpha(r)$ *is the critical value of* χ^2*-distribution with* r *degrees of freedom at level* α. *Under the alternative* $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} - \boldsymbol{t} = \boldsymbol{h}_n$ *for some* $\boldsymbol{h}_n$ *satisfying* $\boldsymbol{h}_n = O(\sqrt{\min(s + m - r, r)/n})$, *we have for any* $0 < \alpha < 1$, *and* $T = T_W, T_S$ *and* T_L,

$$\lim_n |Pr(T > \chi^2_\alpha(r)) - Pr(\chi^2(r, \gamma_n) > \chi^2_\alpha(r))| = 0,$$

where $\gamma_n = n\boldsymbol{h}_n^T \left(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\right)^{-1} \boldsymbol{h}_n/\phi_0$.

REMARK 3.3. Corollary 3.1 shows that the asymptotic power functions of the proposed test statistics are

$$\Pr(\chi^2(r, \gamma_n) > \chi^2_\alpha(r)). \tag{3.6}$$

It follows from Theorem 2 in Ghosh (1973) that the asymptotic power function decreases as r increases for a given γ_n. This is the same as that for traditional likelihood ratio test, score test and Wald's test. However, $\boldsymbol{h}_n$ is an r-dimensional vector in our setting. Thus, one may easily construct an example in which γ_n grows as r increases. As a result, the asymptotic power function may not be monotone increasing function of r.

In Section S3 of Shi et al. (2018), we study in depth that how the penalty on individual coefficient affects the power, and find that the tests are most advantageous if each unpenalized variable is either an important variable (i.e., in $\mathcal{S}$) or a variable in $\mathcal{M}$.

REMARK 3.4. Notice that the null hypothesis reduces to $\boldsymbol{\beta}_{0,\mathcal{M}} = \mathbf{0}$ if we set $\boldsymbol{C}$ to be the identity matrix and $\boldsymbol{t} = \mathbf{0}$. The Wald test based on the desparsified Lasso estimator (van de Geer et al., 2014) and the decorrelated score test (Ning and Liu, 2017) can also be applied to testing such hypothesis. Based on (3.6), we show that these two tests achieve less power than the proposed partial penalized tests in Section S1 of Shi et al. (2018). This is due to the increased variances of the de-sparsified Lasso estimator and the decorrelated score statistic after the debiasing procedure.

3.3. *Some implementation issues.*

3.3.1. *Constrained partial penalized regression.* To construct our test statistics, we need to compute the partial penalized estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_a$. Our algorithm is based upon the alternating direction method of multipliers (ADMM), which is a variant of standard augmented Lagrangian method. Below, we present our algorithm for estimating $\hat{\boldsymbol{\beta}}_0$. The unconstrained estimator $\hat{\boldsymbol{\beta}}_a$ can be similarly computed. For a fixed regularization parameter λ, define

$$\hat{\boldsymbol{\beta}}_0^{\lambda} = \arg\min_{\boldsymbol{\beta}} \left(-L_n(\boldsymbol{\beta}) + \sum_{j \in \mathcal{M}^c} p_\lambda(|\beta_j|) \right), \text{ subject to } \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}.$$

The above optimization problem is equivalent to

$$(\hat{\boldsymbol{\beta}}_0^{\lambda}, \hat{\boldsymbol{\theta}}_0^{\lambda}) = \arg \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ \boldsymbol{\theta} \in \mathbb{R}^{p-m}}} \left(-L_n(\boldsymbol{\beta}) + \sum_{j=1}^{p-m} p_\lambda(|\theta_j|) \right), \quad \text{subject to } \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t},\ \boldsymbol{\beta}_{\mathcal{M}^c} = \boldsymbol{\theta}. \tag{3.7}$$

The augmented Lagrangian for (3.7) is

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{v}) &= -L_n(\boldsymbol{\beta}) + \sum_{j=1}^{p-m} p_\lambda(|\theta_j|) + \boldsymbol{v}^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta} \end{pmatrix} \\ &+ \frac{\rho}{2} \|\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t}\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta}\|_2^2, \end{aligned}$$

for a given $\rho > 0$. Applying dual ascent method yields the following algorithm:

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{v}^k)^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta}^k \end{pmatrix} + \frac{\rho}{2} \left\| \begin{matrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta}^k \end{matrix} \right\|_2^2 - L_n(\boldsymbol{\beta}) \right\},$$

$$\boldsymbol{\theta}^{k+1} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{j=1}^{p-m} p_\lambda(|\theta_j|) + \frac{\rho}{2} \|\boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta}\|_2^2 + (\boldsymbol{v}^k)^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}}^{k+1} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta} \end{pmatrix} \right\},$$

$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \rho \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}}^{k+1} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta}^{k+1} \end{pmatrix},$$

for the $(k+1)$th iteration.

Since L_n is twice differentiable, $\boldsymbol{\beta}^{k+1}$ can be obtained by the Newton-Raphson algorithm. $\boldsymbol{\theta}^{k+1}$ may have a closed form for some popular penalties such as Lasso, SCAD or MCP penalty. In our implementation, we use the SCAD penalty,

$$p_\lambda(|\beta_j|) = \lambda \int_0^{|\beta_j|} \left\{ I(t \le \lambda) + \frac{(a\lambda - t)_+}{a-1} I(t > \lambda) \right\} dt,$$

and set $a = 3.7$, $\rho = 1$.

To obtain $\hat{\boldsymbol{\beta}}_0$, we compute $\hat{\boldsymbol{\beta}}_0^\lambda$ for a series of log-spaced values in $[-\lambda_{\min}, \lambda_{\max}]$ for some $\lambda_{\min} < \lambda_{\max}$. Then we choose $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0^{\hat{\lambda}}$ by minimizing the following information criterion:

$$\hat{\lambda} = \arg\min_{\lambda} \left(-nL_n(\lambda) + c_n \|\hat{\boldsymbol{\beta}}^\lambda\|_0 \right),$$

where $c_n = \max\{\log n, \log(\log(n)) \log(p)\}$. Using similar arguments in Schwarz (1978) and Fan and Tang (2013), we can show such information criterion is consistent in both fixed p and ultrahigh dimension setting.

3.3.2. *Estimation of the nuisance parameter.* It can be shown that $\phi_0 = 1$ for logistic or Poisson regression models. In linear regression models, we have $\phi_0 = \mathrm{E}(Y_i - \boldsymbol{\beta}_0^T \boldsymbol{X}_i)^2$. In our implementation, we estimate ϕ_0 by

$$\hat{\phi} = \frac{1}{n - |\widehat{S}_a| - m} \sum_{i=1}^n (Y_i - \hat{\boldsymbol{\beta}}_a^T \boldsymbol{X}_i)^2,$$

where $\hat{\boldsymbol{\beta}}_a$ is defined in (2.3).

In Section S2 of the supplementary material (Shi et al., 2018), we show $\hat{\phi} = \phi_0 + O_p(n^{-1/2})$, under the conditions in Theorem 2.1, which implies selection consistency. Alternatively, one can estimate ϕ_0 using refitted cross-validation (Fan, Guo and Hao, 2012) or scaled lasso (Sun and Zhang, 2013).

TABLE 1
Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%), under the setting where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$.

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$			$H_0^{(1)}$			
0	4.33(0.83)	4.33(0.83)	4.67(0.86)	5.67(0.94)	5.67(0.94)	5.67(0.94)
0.1	13.17(1.38)	13.50(1.40)	13.50(1.40)	11.67(1.31)	11.67(1.31)	11.67(1.31)
0.2	39.83(2.00)	40.17(2.00)	40.00(2.00)	39.67(2.00)	39.67(2.00)	39.67(2.00)
0.4	92.33(1.09)	93.17(1.03)	93.17(1.03)	92.67(1.06)	92.67(1.06)	92.67(1.06)
$h^{(2)}$			$H_0^{(2)}$			
0	5.17(0.90)	5.17(0.90)	5.67(0.94)	5.33(0.92)	5.33(0.92)	5.33(0.92)
0.1	11.00(1.28)	11.00(1.28)	11.33(1.29)	12.50(1.35)	12.50(1.35)	12.50(1.35)
0.2	30.67(1.88)	30.67(1.88)	31.00(1.89)	33.67(1.93)	33.67(1.93)	33.67(1.93)
0.4	85.17(1.45)	85.00(1.46)	85.00(1.46)	87.83(1.33)	87.83(1.33)	87.83(1.33)
$h^{(2)}$			$H_0^{(3)}$			
0	6.50 (1.01)	6.33(0.99)	6.50(1.01)	5.67(0.94)	5.67(0.94)	5.67(0.94)
0.1	11.83 (1.32)	11.67(1.31)	11.67(1.31)	11.00(1.28)	11.00(1.28)	11.00(1.28)
0.2	31.67 (1.90)	31.50(1.90)	31.67(1.90)	33.17(1.92)	33.17(1.92)	33.17(1.92)
0.4	84.33 (1.48)	84.17(1.49)	84.50(1.48)	86.00(1.42)	86.17(1.41)	86.17(1.41)

4. Numerical Examples. In this section, we examine the finite sample performance of the proposed tests. Simulation results for linear regression and logistic regression are presented in the main text. In the supplementary material (Shi et al., 2018), we present simulation results for Poisson log-linear model and illustrate the proposed methodology by a real data example.

4.1. *Linear regression.* Simulated data with sample size $n = 100$ were generated from

$$Y = 2X_1 - (2 + h^{(1)})X_2 + h^{(2)}X_3 + \varepsilon$$

where $\varepsilon \sim N(0, 1)$ and $\boldsymbol{X} \sim N(\mathbf{0}_p, \mathbf{\Sigma})$, and $h^{(1)}$ and $h^{(2)}$ are some constants. The true value $\boldsymbol{\beta}_0 = (2, -2-h^{(1)}, h^{(2)}, \mathbf{0}_{p-3}^T)^T$ where $\mathbf{0}_q$ denotes a zero vector of length q.

4.1.1. *Testing linear hypothesis.* We focus on testing the following three pairs of hypotheses:

$$H_0^{(1)} : \beta_{0,1} + \beta_{0,2} = 0, \quad v.s \quad H_a^{(1)} : \beta_{0,1} + \beta_{0,2} \neq 0.$$
$$H_0^{(2)} : \beta_{0,2} + \beta_{0,3} = -2, \quad v.s \quad H_a^{(2)} : \beta_{0,2} + \beta_{0,3} \neq -2.$$
$$H_0^{(3)} : \beta_{0,3} + \beta_{0,4} = 0, \quad v.s \quad H_a^{(3)} : \beta_{0,3} + \beta_{0,4} \neq 0.$$

These hypotheses test linear structures between two regression coefficients. When testing $H_0^{(1)}$, we set $h^{(2)} = 0$, and hence $H_0^{(1)}$ holds if and only if $h^{(1)} = 0$. Similarly when testing $H_0^{(2)}$ and $H_0^{(3)}$, we set $h^{(1)} = 0$, and hence the hull hypotheses hold if and only if $h^{(2)} = 0$.

We consider two different dimensions, $p = 50$ and $p = 200$, and two different covariance matrices $\mathbf{\Sigma}$, corresponding to $\mathbf{\Sigma} = \boldsymbol{I}$ and $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. This yields a total of 4 settings. For each hypothesis and each setting, we further consider four scenarios, by setting $h^{(j)} = 0, 0.1, 0.2, 0.4$. Therefore, the null holds under the first scenario and the alternative holds under the rest three. Table 1 summarizes the rejection probabilities for $H_0^{(1)}$, $H_0^{(2)}$ and $H_0^{(3)}$ under the settings where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}$. Rejection probabilities of the proposed tests under the settings where $\mathbf{\Sigma} = \boldsymbol{I}$ are given in Table S1 in the supplementary material. The rejection probabilities are evaluated via 600 simulation replications.

Based on the results, it can be seen that under these null hypotheses, Type I error rates of the three tests are well controlled and close to the nominal level for all four settings. Under the alternative hypotheses, the powers of these three test statistics increase as $h^{(1)}$ or $h^{(2)}$ increases, showing the consistency of our testing procedure. Moreover, the empirical rejection rates between these three test statistics are very close across all different scenarios and settings. For example, the rejection rates are exactly the same for testing $H_0^{(1)}$ and $H_0^{(2)}$ when $p = 200$ in Table 1, although we observed that the values of these three statistics in our simulation are slightly different. This is consistent with our theoretical findings that these statistics are asymptotically equivalent even in high dimensional settings. Figures S1, S2 and S3 in the supplementary material depicts the kernel density estimates of three test statistics under $H_0^{(1)}$ and $H_0^{(2)}$ with different combinations of p and the covariance matrices respectively. It can be seen that these three test statistics converge to their limiting distributions under the null hypotheses.

4.1.2. *Testing univariate parameter.* Consider testing the following two pairs of hypotheses:

$$H_0^{(4)} : \beta_{0,2} = -2, \quad v.s \quad H_a^{(1)} : \beta_{0,2} \neq -2.$$
$$H_0^{(5)} : \beta_{0,3} = 0, \quad v.s \quad H_a^{(2)} : \beta_{0,3} \neq 0.$$

We set $h^{(2)} = 0$ when testing $H_0^{(4)}$, and set $h^{(1)} = 0$ when testing $H_0^{(5)}$. Therefore, $H_0^{(4)}$ is equivalent to $h^{(1)} = 0$ and $H_0^{(5)}$ is equivalent to $h^{(2)} = 0$. We use the same 4 settings described in Section 4.1.1. For each setting,

TABLE 2

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics, the Wald test statistic based on the de-sparsified Lasso estimator and the decorrelated score statistic under the settings where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$*, with standard errors in parenthesis (%).*

	T_L	T_W	T_S	T_W^D	T_S^D
$h^{(1)}$			$H_0^{(4)}$ and $p = 50$		
0	5.17(0.90)	5.33(0.92)	5.50(0.93)	12.67(1.36)	7.00(1.04)
0.1	15.67(1.48)	16.00(1.50)	16.00(1.50)	6.00(0.97)	14.67(1.44)
0.2	41.00(2.01)	41.33(2.01)	41.50(2.01)	14.83(1.45)	38.83(1.99)
0.4	92.50(1.08)	93.00(1.04)	93.00(1.04)	67.67(1.91)	88.67(1.29)
			$H_0^{(4)}$ and $p = 200$		
0	4.83(0.88)	4.83(0.88)	4.83(0.88)	21.83(1.69)	5.50(0.93)
0.1	11.00(1.28)	11.00(1.28)	11.00(1.28)	5.83(0.96)	10.83(1.27)
0.2	40.50(2.00)	40.50(2.00)	40.50(2.00)	6.17(0.98)	37.83(1.98)
0.4	91.50(1.14)	91.50(1.14)	91.50(1.14)	49.33(2.04)	88.00(1.33)
$h^{(2)}$			$H_0^{(5)}$ and $p = 50$		
0	6.33(0.99)	6.00(0.97)	6.50(1.00)	5.33(0.92)	3.00(0.70)
0.1	13.67(1.40)	13.50(1.40)	14.00(1.42)	5.33(0.92)	9.17(1.18)
0.2	40.17(2.00)	40.33(2.00)	40.50(2.00)	15.67(1.48)	28.50(1.84)
0.4	90.83(1.18)	91.33(1.15)	91.67(1.13)	69.17(1.89)	83.33(1.52)
			$H_0^{(5)}$ and $p = 200$		
0	5.67(0.94)	5.67(0.94)	5.67(0.94)	6.50(1.01)	2.67(0.66)
0.1	13.67(1.40)	13.67(1.40)	13.67(1.40)	3.67(0.77)	8.17(1.12)
0.2	39.17(1.99)	39.17(1.99)	39.17(1.99)	9.67(1.21)	24.67(1.76)
0.4	91.50(1.14)	91.50(1.14)	91.50(1.14)	51.33(2.04)	80.50(1.62)

we set $h^{(1)} = 0.1, 0.2, 0.4$ under $H_a^{(4)}$ and $h^{(2)} = 0.1, 0.2, 0.4$ under $H_a^{(5)}$. Comparison is made among the following test statistics:

(i) The proposed likelihood ratio (T_L), Wald (T_W) and score (T_S) statistic.
(ii) The Wald test statistic based on the de-sparisfied Lasso estimator (T_W^D).
(iii) The decorrelated score statistic. (T_S^D).

The test statistic T_W^D is computed via the R package `hdi` (Dezeure et al., 2015). We calculate T_S^D according to Section 4.1 in Ning and Liu (2017). More specifically, the initial estimator $\hat{\boldsymbol{\beta}}$ is computed by a penalized linear regression with SCAD penalty function, and $\hat{\boldsymbol{\omega}}$ is computed by a penalized linear regression with l_1 penalty function (see Equation (4.4) in Ning and Liu, 2017). These penalized regressions are implemented via the R package `ncvreg` (Breheny and Huang, 2011). The tuning parameters are selected via 10-folded cross-validation. The rejection probabilities of these test statistics under the settings where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}$ are reported in Table 2. In the supplementary material, we report the rejection probabilities of these test

TABLE 3
Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%), under the settings where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$.

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$			$H_0^{(6)}$			
0	4.83(0.88)	4.50(0.85)	4.67(0.86)	4.83(0.88)	4.83(0.88)	4.83(0.88)
0.2	28.17(1.84)	28.17(1.84)	28.50(1.84)	28.50(1.84)	28.50(1.84)	28.50(1.84)
0.4	80.33(1.62)	80.17(1.63)	80.33(1.62)	79.83(1.64)	79.83(1.64)	79.83(1.64)
0.8	99.83(0.17)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)
$h^{(1)}$			$H_0^{(7)}$			
0	4.50(0.85)	4.50(0.85)	4.50(0.85)	5.00(0.89)	5.00(0.89)	5.00(0.89)
0.2	18.17(1.57)	18.33(1.58)	18.33(1.58)	18.33(1.58)	18.33(1.58)	18.33(1.58)
0.4	53.83(2.04)	54.17(2.03)	54.00(2.03)	57.33(2.02)	57.33(2.02)	57.33(2.02)
0.8	98.50(0.50)	99.00(0.41)	99.00(0.41)	98.50(0.50)	98.50(0.50)	98.50(0.50)
$h^{(1)}$			$H_0^{(8)}$			
0	5.17 (0.90)	5.00(0.89)	5.17(0.90)	5.67(0.94)	5.67(0.94)	5.67(0.94)
0.2	14.33 (1.43)	14.33(1.43)	14.33(1.43)	13.67(1.40)	13.67(1.40)	13.67(1.40)
0.4	42.00 (2.01)	42.17(2.02)	42.17(2.02)	41.67(2.01)	41.67(2.01)	41.67(2.01)
0.8	92.83 (1.05)	92.83(1.05)	92.83(1.05)	93.00(1.04)	93.00(1.04)	93.00(1.04)

statistics under the settings where $\mathbf{\Sigma} = \boldsymbol{I}$ in Table S2. Results are averaged over 600 simulation replications.

From Table 2, it can be seen that T_W^D failed to test $H_0^{(4)}$ under the settings where $\mathbf{\Sigma} = \{0.5^{|i-j|}\}$. Under the null hypotheses, the Type I error rates of T_W^D are greater than 12%. Under the alternative hypotheses, the proposed test statistics and the decorrelated score test are more powerful than T_W^D in almost all cases. Besides, we note that T_L, T_W, T_S and T_S^D perform comparable under the settings where $\mathbf{\Sigma} = \boldsymbol{I}$. When $\mathbf{\Sigma} = \{0.5^{|i-j|}\}$ however, the proposed test statistics achieve greater power than T_S^D. This is in line with our theoretical findings (see Section S1 of the supplementary material for details).

4.1.3. *Effects on* m. In Section 4.1.1, we consider linear hypotheses involving two parameters only. As suggested by one of the referee, we further examine our test statistics under settings where more regression parameters are involved in the hypotheses. More specifically, we consider the following

three pairs of hypotheses:

$$H_0^{(6)}: \sum_{j=1}^{4} \beta_{0,j} = 0, \quad v.s \quad H_a^{(6)}: \sum_{j=1}^{4} \beta_{0,j} \neq 0.$$
$$H_0^{(7)}: \sum_{j=1}^{8} \beta_{0,j} = 0, \quad v.s \quad H_a^{(7)}: \sum_{j=1}^{8} \beta_{0,j} \neq 0.$$
$$H_0^{(8)}: \sum_{j=1}^{12} \beta_{0,j} = 0, \quad v.s \quad H_a^{(8)}: \sum_{j=1}^{12} \beta_{0,j} \neq 0.$$

The numbers of parameters involved in $H_0^{(6)}$, $H_0^{(7)}$ and $H_0^{(8)}$ are equal to 4, 8 and 12, respectively. We consider the same 4 settings described in Section 4.1.1. For each setting, we set $h^{(1)} = 0, 0.2, 0.4, 0.8$ and $h^{(2)} = 0$. Hence, the null hypotheses hold when $h^{(1)} = 0$ and the alternatives hold when $h^{(1)} > 0$. We report the rejection probabilities over 600 replications in Table 3, under the settings where $\boldsymbol{\Sigma} = \{0.5^{|i-j|}\}$. Rejection probabilities under the settings where $\boldsymbol{\Sigma} = \boldsymbol{I}$ are reported in Table S3 in the supplementary material.

The Type I error rates of the three test statistics are close to the nominal level under the null hypotheses. The powers of the test statistics increase as $h^{(1)}$ increases, under the alternative hypotheses. Moreover, we note that the powers decrease as m increases. This is in line with Corollary 3.1 which states that the asymptotic power function of our test statistics is a function of r and γ_n. Recall that $\gamma_n = n\boldsymbol{h}_n^T \left(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\right)^{-1} \boldsymbol{h}_n/\phi_0$. Consider the following sequence of null hypotheses indexed by $m \geq 2$: $\boldsymbol{C}_m\boldsymbol{\beta}_0 = 0$ where $\boldsymbol{C}_m = (1, \cdots, 1, \boldsymbol{0}_{p-m})$. Let $\gamma_{n,m} = n\boldsymbol{h}_n^T \left(\boldsymbol{C}_m\boldsymbol{\Omega}_{mm}\boldsymbol{C}_m^T\right)^{-1} \boldsymbol{h}_n/\phi_0$. Under the given settings, we have $\boldsymbol{\Omega}_{mm} = (\omega_{ij})$ is a banded matrix with $\omega_{ij} = 0$ for $|i-j| \geq 2$, $\omega_{ij} = -1/(1-\rho^2)$ for $|i-j| = 1$, $\omega_{11} = \omega_{mm} = 1/\{\rho(1-\rho^2)\}$, and $\omega_{jj} = (1+\rho^2)/\{\rho(1-\rho^2)\}$ for $j \neq 1$ and m, where ρ is the auto-correlation between X_1 and X_2. It is immediate to see $\gamma_{n,m}$ decreases as m increases.

4.2. *Logistic regression.* In this example, we generate data with sample size $n = 300$ from the logistic regression model

$$\text{logit}\{\Pr(Y = 1|\boldsymbol{X})\} = 2X_1 - (2 + h^{(1)})X_2 + h^{(2)}X_3,$$

where $\text{logit}(p) = \log\{p/(1-p)\}$, the logit link function, and $\boldsymbol{X} \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma})$.

4.2.1. *Testing linear hypothesis.* We consider the same linear hypotheses as those in Section 4.1.1:

$$H_0^{(1)}: \beta_{0,1}+\beta_{0,2}=0, \quad v.s \quad H_a^{(1)}: \beta_{0,1}+\beta_{0,2}\neq 0.$$
$$H_0^{(2)}: \beta_{0,2}+\beta_{0,3}=-2, \quad v.s \quad H_a^{(2)}: \beta_{0,2}+\beta_{0,3}\neq -2.$$
$$H_0^{(3)}: \beta_{0,3}+\beta_{0,4}=0, \quad v.s \quad H_a^{(3)}: \beta_{0,3}+\beta_{0,4}\neq 0.$$

Similarly, we set $h^{(2)}=0$ when testing $H_0^{(1)}$, and set $h^{(1)}=0$ when testing $H_0^{(2)}$. Therefore, $H_0^{(1)}$ is equivalent to $h^{(1)}=0$ and $H_0^{(2)}$ is equivalent to $h^{(2)}=0$. We use the same 4 settings described in Section 4.1.1. For each of the four settings, we set $h^{(j)}=0.2, 0.4, 0.8$ under $H_a^{(j)}$.The rejection probabilities for $H_0^{(1)}$ and $H_0^{(2)}$ over 600 replications are given in Table S4 in the supplementary material. We also plot the kernel density estimates of three test statistics under $H_0^{(1)}$ and $H_0^{(2)}$ in Figures S4, S5 and S6 in the supplementary material. The findings are very similar to those in the previous examples.

4.2.2. *Testing univariate parameter.* To compare the proposed partial penalized Wald (T_W), score (T_S) and likelihood ratio (T_L) test statistics with the Wald test based on the de-sparsified Lasso estimator (T_W^D) and the decorrelated score test (T_S^D), we consider testing the following hypotheses:

$$H_0^{(4)}: \beta_{0,2}=-2, \quad v.s \quad H_a^{(4)}: \beta_{0,2}\neq -2.$$
$$H_0^{(5)}: \beta_{0,3}=0, \quad v.s \quad H_a^{(5)}: \beta_{0,3}\neq 0.$$

Similar to Section 4.1.2, we set $h^{(2)}=0$ when testing $H_0^{(4)}$, and set $h^{(1)}=0$ when testing $H_0^{(5)}$. We set $h^{(1)}=0$ under $H_0^{(4)}$, $h^{(1)}=0.2, 0.4, 0.8$ under $H_a^{(4)}$ and set $h^{(2)}=0$ under $H_0^{(5)}$, $h^{(2)}=0.2, 0.4, 0.8$ under $H_a^{(5)}$. We consider the same 4 settings described in Section 4.1.1. The test statistic T_W^D is computed via the R package `hdi` and T_S^D is obtained according to Section 4.2 of Ning and Liu (2017). We compute the initial estimator $\hat{\boldsymbol{\beta}}$ in T_S^D by fitting a penalized logistic regression with SCAD penalty function, and calculate $\hat{\boldsymbol{\omega}}$ by fitting a penalized linear regression with l_1 penalty function. These penalized regressions are implemented via the R package `ncvreg`. We report the rejection probabilities of T_W,T_S,T_L,T_W^D and T_S^D in Table S5 in the supplementary article, based on 600 simulation replications.

Based on the results, it can be seen that the Type I error rates of T_W^D and T_S^D are significantly larger than the nominal level in almost all cases for testing $H_0^{(4)}$. On the other hand, the Type I error rates of the proposed test

statistics are close to the nominal level under $H_0^{(4)}$. Besides, under $H_a^{(5)}$, the powers of the proposed test statistics are greater than or equal to T_W^D and T_S^D in all cases.

4.2.3. *Effects on m.* As in Section 4.1.3, we further examine the proposed test statistics by allowing more regression coefficients to appear in the linear hypotheses. Similarly, we consider the following three pairs of hypotheses:

$$H_0^{(6)}: \sum_{j=1}^{4} \beta_{0,j} = 0, \quad v.s \quad H_a^{(6)}: \sum_{j=1}^{4} \beta_{0,j} \neq 0.$$
$$H_0^{(7)}: \sum_{j=1}^{8} \beta_{0,j} = 0, \quad v.s \quad H_a^{(7)}: \sum_{j=1}^{8} \beta_{0,j} \neq 0.$$
$$H_0^{(8)}: \sum_{j=1}^{12} \beta_{0,j} = 0, \quad v.s \quad H_a^{(8)}: \sum_{j=1}^{12} \beta_{0,j} \neq 0.$$

We set $h^{(2)} = 0$, and set $h^{(1)} = 0$ under the null hypotheses, $h^{(1)} = 0.4, 0.8, 1.6$ under the alternative hypotheses. The same 4 settings described in Section 4.1.1 are used. The rejection probabilities of the proposed test statistics are reported in Table S6 in the supplementary article. Results are averaged over 600 replications. Findings are very similar to those in Section 4.1.3.

5. Technical proofs. This section consists of the proof of Theorem 3.1. To establish Theorem 3.1, we need the following lemma. The proof of this lemma is given in Section 5.1. For any symmetric and positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{q\times q}$, it follows from the spectral theorem that $\boldsymbol{A} = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U}$ for some orthogonal matrix $\boldsymbol{U}$ and diagonal matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. Since the diagonal elements in $\boldsymbol{\Lambda}$ are positive, we use $\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{\Lambda}^{-1/2}$ to denote the diagonal matrices $\mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_q^{1/2})$ and $\mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_q^{-1/2})$, respectively. In addition, we define $\boldsymbol{A}^{1/2} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}$ and $\boldsymbol{A}^{-1/2} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}$.

LEMMA 5.1. *Under the conditions in Theorem 3.1, we have*

$$\lambda_{\max}(\boldsymbol{K}_n) = O(1), \tag{5.1}$$
$$\lambda_{\max}(\boldsymbol{K}_n^{1/2}) = O(1), \tag{5.2}$$
$$\lambda_{\max}(\boldsymbol{K}_n^{-1/2}) = O(1), \tag{5.3}$$
$$\lambda_{\max}\left((\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}\right) = O(1), \tag{5.4}$$
$$\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\|_2 = O(1), \tag{5.5}$$

$$\|\boldsymbol{\Psi}^{1/2}(\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T)^{-1}\boldsymbol{\Psi}^{1/2}-\boldsymbol{I}\|_2=O_p\left(\frac{s+m}{\sqrt{n}}\right), \tag{5.6}$$

$$\|\boldsymbol{I}-\boldsymbol{K}_n^{1/2}\widehat{\boldsymbol{K}}_{n,0}^{-1}\boldsymbol{K}_n^{1/2}\|_2=O_p\left(\frac{s+m}{\sqrt{n}}\right), \tag{5.7}$$

where $\boldsymbol{\Psi}=\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$ *and*

$$\widehat{\boldsymbol{K}}_{n,0}=\frac{1}{n}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_S\\ \boldsymbol{X}_S^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_S^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\boldsymbol{X}_S\end{pmatrix}.$$

We break the proof into four steps. In the first three steps, we show T_W/r, T_S/r and T_L/r are equivalent to T_0/r, respectively, where

$$T_0=\frac{1}{\phi_0}(\boldsymbol{\omega}_n+\sqrt{n}\boldsymbol{h}_n)^T(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}(\boldsymbol{\omega}_n+\sqrt{n}\boldsymbol{h}_n),$$

and

$$\boldsymbol{\omega}_n=\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{s\times r}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}.$$

In the final step, we show the χ^2 approximation (3.5) holds for T_W, T_S and T_L.

Step 1: We first show that T_W/r is equivalent to T_0/r. It follows from Theorem 2.1 that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{a,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix}=\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}+\boldsymbol{R}_a,$$

for some vector $\boldsymbol{R}_a$ that satisfies

$$\|\boldsymbol{R}_a\|_2=o_p(1). \tag{5.8}$$

Therefore, we have

$$\sqrt{n}\boldsymbol{C}(\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}})=\boldsymbol{\omega}_n+\boldsymbol{C}\boldsymbol{R}_{a,J_0}, \tag{5.9}$$

where $J_0=[1,\ldots,m]$. Since $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}}=\boldsymbol{t}+\boldsymbol{h}_n$, it follows from (5.9) that

$$\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\boldsymbol{t})=\boldsymbol{\omega}_n+\boldsymbol{C}\boldsymbol{R}_{a,J_0}+\sqrt{n}\boldsymbol{h}_n,$$

and hence

$$\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\boldsymbol{t})=\boldsymbol{\Psi}^{-1/2}(\boldsymbol{\omega}_n+\boldsymbol{C}\boldsymbol{R}_{a,J_0}+\sqrt{n}\boldsymbol{h}_n). \tag{5.10}$$

By (5.8) and (5.5) in Lemma 5.1, we have

$$\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\boldsymbol{R}_{a,J_0}\|_2 \leq \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\|\|\boldsymbol{R}_{a,J_0}\|_2 \leq \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\|\|\boldsymbol{R}_{a}\|_2 = o_p(1).$$

This together with (5.10) gives

$$\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t}) = \boldsymbol{\Psi}^{-1/2}(\boldsymbol{\omega}_n + \sqrt{n}\boldsymbol{h}_n) + o_p(1). \tag{5.11}$$

Note that

$$\mathrm{E}\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{\omega}_n\|_2^2 = \mathrm{tr}\left(\boldsymbol{\Psi}^{-1/2}\mathrm{E}\boldsymbol{\omega}_n\boldsymbol{\omega}_n^T\boldsymbol{\Psi}^{-1/2}\right) = \phi_0\mathrm{tr}\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1/2}\right) = r\phi_0.$$

By Markov's inequality, we have

$$\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{\omega}_n\|_2 = O_p(\sqrt{r}). \tag{5.12}$$

Besides, it follows from (5.4) in Lemma 5.1 and Condition (A4) that

$$\|\sqrt{n}\boldsymbol{\Psi}^{-1/2}\boldsymbol{h}_n\|_2 = O(\sqrt{r}). \tag{5.13}$$

This together with (5.11) and (5.12) implies that

$$\|\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})\|_2 = O_p(\sqrt{r}). \tag{5.14}$$

Combining this together with (5.6) in Lemma 5.1 gives

$$\begin{aligned}&\|\{\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})\}^T\{\boldsymbol{\Psi}^{1/2}(\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T)^{-1}\boldsymbol{\Psi}^{1/2} - \boldsymbol{I}\}\{\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})\}\|_2^2\\ &\leq \|\{\sqrt{n}\boldsymbol{\Psi}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})\}\|_2^2\|\{\boldsymbol{\Psi}^{1/2}(\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T)^{-1}\boldsymbol{\Psi}^{1/2} - \boldsymbol{I}\}\|_2 = O_p\left(\frac{r(s+m)}{\sqrt{n}}\right).\end{aligned}$$

The last term is $o_p(r)$ under the condition $s+m = o(n^{1/3})$. By the definition of T_W, we have shown that

$$\hat{\phi}|T_W - T_{W,0}| = o_p(r), \tag{5.15}$$

where

$$T_{W,0} = \frac{n(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t})}{\hat{\phi}}.$$

Under the conditions in Theorem 3.1, we have $\hat{\phi} = \phi_0 + o_p(1)$. Since $\phi_0 > 0$, we have

$$1/\hat{\phi} = O_p(1), \tag{5.16}$$

which together with (5.15) entails that $T_W = T_{W,0} + o_p(r)$.

It follows from (5.10)-(5.13) and the condition $s + m = o(n^{1/3})$ that

$$\begin{aligned}(5.17)\quad \hat{\phi} T_{W,0} &= \left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n + o_p(1) \right\|_2^2 \\ &= \left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2 + o_p(1) + o_p\left(\boldsymbol{\Psi}^{-1/2} (\boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{h}_n) \right) \\ &= \left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2 + o_p(1) + o_p(r) \\ &= \left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2 + o_p(r) = \hat{\phi} T_{W,1} + o_p(r),\end{aligned}$$

where

$$T_{W,1} = \frac{\left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2}{\hat{\phi}}.$$

By (5.16), we obtain $T_{W,0} = T_{W,1} + o_p(r)$ and hence $T_W = T_{W,1} + o_p(r)$. In the following, we show $T_{W,1} = T_0 + o_p(r)$.

Observe that

$$(5.18)\qquad |T_{W,1} - T_0| = \frac{|\phi_0 - \hat{\phi}|}{\hat{\phi}\phi_0} \left\| \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2.$$

It follows from (5.12), (5.13), (5.16) and the condition $|\hat{\phi} - \phi_0| = o_p(1)$ that right-hand side (RHS) of (5.18) is of the order $o_p(r)$. This proves $T_{W,1} = T_0 + o_p(r)$.

Step 2: We show that T_S/r is equivalent to T_0/r. Based on the proof of Theorem 2.1 in Section S5.1 of the supplementary article, we have

$$\begin{aligned}(5.19)\quad &\frac{1}{\sqrt{n}} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\} = \frac{1}{\sqrt{n}} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} \\ &- \frac{1}{\sqrt{n}} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} + o_p(1),\end{aligned}$$

and

$$\begin{aligned}\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} &= \frac{1}{\sqrt{n}} \boldsymbol{K}_n^{-1/2} (\boldsymbol{I} - \boldsymbol{P}_n) \boldsymbol{K}_n^{-1/2} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} \\ (5.20)\qquad &- \sqrt{n} \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix} \boldsymbol{\Psi}^{-1} \boldsymbol{h}_n + o_p(1).\end{aligned}$$

Combining (5.1) with (5.20) gives

$$\begin{aligned}\sqrt{n}\boldsymbol{K}_n\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix} &= \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\ &- \sqrt{n}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n+o_p(1),\end{aligned}$$

which together with (5.19) implies that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}=\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}^T\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}+o_p(1)\\ -\;&\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}+\sqrt{n}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n\\ =\;&\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}+\sqrt{n}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n+o_p(1).\end{aligned}$$

By (5.3), we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\} &= \frac{1}{\sqrt{n}}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\ &+\sqrt{n}\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n+o_p(1).\end{aligned}\tag{5.21}$$

It follows from (5.5) and (5.13) that

$$\left\|\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n\right\|_2\le\|\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2}\|_2\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{h}_n\|_2=O_p(\sqrt{r/n}).$$

This together with (5.3) yields

$$\sqrt{n}\left\|\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n\right\|_2=O_p(\sqrt{r}).\tag{5.22}$$

Notice that

$$\mathrm{E}\left\|\frac{1}{\sqrt{n}}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\right\|_2^2=\mathrm{tr}(\boldsymbol{P}_n)=\mathrm{rank}(\boldsymbol{P}_n)=r.$$

It follows from Markov's equality that

$$\left\|\frac{1}{\sqrt{n}}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\right\|_2=O_p(\sqrt{r}).$$

Combining this with (5.21) and (5.22) yields

$$\left\|\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\|_2 = O_p(\sqrt{r}). \tag{5.23}$$

This together with (5.7) and the condition $s+m=o(n^{1/3})$ gives that

$$\left|\frac{1}{n}\left\{\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\}^T(\boldsymbol{K}_n^{-1}-\widehat{\boldsymbol{K}}_{n,0}^{-1})\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right| \le$$

$$\frac{1}{n}\left\|\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\|_2^2\|\boldsymbol{I}-\boldsymbol{K}_n^{1/2}\widehat{\boldsymbol{K}}_{n,0}^{-1}\boldsymbol{K}_n^{1/2}\|_2 = O_p\left(\frac{r(s+m)}{\sqrt{n}}\right)=o_p(r).$$

When $\hat{S}_0=S$, we have

$$\hat{\phi}T_S=\frac{1}{n}\left\{\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\}^T\widehat{\boldsymbol{K}}_{n,0}^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}.$$

Since $\Pr(\hat{S}_0=S)\to 1$, we obtain $\hat{\phi}|T_S-T_{S,0}|=o_p(r)$, where

$$T_{S,0}=\frac{1}{n\hat{\phi}}\left\{\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}.$$

This together with (5.16) implies that $|T_S-T_{S,0}|=o_p(r)$. Using similar arguments in (5.17) and (5.18), we can show that $T_{S,0}/r$ is equivalent to $T_{S,1}/r$, where $T_{S,1}$ is defined as

$$\frac{1}{\phi_0}\left\|\frac{1}{\sqrt{n}}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}+\sqrt{n}\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n\right\|_2^2.$$

Recall that

$$\boldsymbol{P}_n=\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1/2},$$

we have

$$\begin{aligned}T_{S,1} &= \frac{1}{\phi_0}\left\|\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{\omega}_n+\sqrt{n}\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n\right\|_2^2\\ &= \frac{1}{\phi_0}\left\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{\omega}_n+\sqrt{n}\boldsymbol{\Psi}^{-1/2}\boldsymbol{h}_n\right\|_2^2=T_0.\end{aligned}$$

This proves the equivalence between T_S/r and T_0/r.

Step 3: By Theorem 2.1, we have

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{a,S}-\hat{\boldsymbol{\beta}}_{0,S}\end{pmatrix} = \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + \sqrt{n}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\\ \boldsymbol{0}\end{pmatrix}+o_p(1).$$

Notice that

$$\begin{aligned}&\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\\ \boldsymbol{0}\end{pmatrix}\\ =\ &\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\begin{pmatrix}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\\ \boldsymbol{0}\end{pmatrix}=\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n.\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{a,S}-\hat{\boldsymbol{\beta}}_{0,S}\end{pmatrix} &= \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\ &+ \sqrt{n}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n+o_p(1).\end{aligned}\tag{5.24}$$

Similar to (5.23), we can show that

$$n\|\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S}-\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}\|_2^2=O_p(r).\tag{5.25}$$

Under the event $\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}=\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S}=0$, using third-order Taylor expansion, we obtain that

$$\begin{aligned}&L_n(\hat{\boldsymbol{\beta}}_0)-L_n(\hat{\boldsymbol{\beta}}_a)=\frac{1}{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\hat{\boldsymbol{\beta}}_{a,S}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\}\\ -\ &\frac{1}{2n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\hat{\boldsymbol{\beta}}_{a,S}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}^T\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\hat{\boldsymbol{\beta}}_{a,S}\end{pmatrix}\\ +\ &\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\hat{\boldsymbol{\beta}}_{a,S}\end{pmatrix}^T\boldsymbol{R},\end{aligned}$$

where $n\|\boldsymbol{R}\|_\infty$ is upper bounded by

$$\begin{aligned}&\max_{j\in\mathcal{M}\cup S}\left|(\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S})^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\mathrm{diag}\{|\boldsymbol{X}^j|\circ|b'''(\boldsymbol{X}\boldsymbol{\beta}^*)|\}\boldsymbol{X}_{\mathcal{M}\cup S}(\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S})\right|\\ &\qquad\le\|\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}-\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S}\|_2^2\max_{j\in\mathcal{M}\cup S}\lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}\cup S}^T\mathrm{diag}\{|\boldsymbol{X}^j|\circ|b'''(\boldsymbol{X}\boldsymbol{\beta}^*)|\}\boldsymbol{X}_{\mathcal{M}\cup S}\right),\end{aligned}$$

for some $\boldsymbol{\beta}^*$ lying on the line segment between $\hat{\boldsymbol{\beta}}_a$ and $\hat{\boldsymbol{\beta}}_0$. By Theorem 2.1, we have $\boldsymbol{\beta}^*_{(\mathcal{M}\cup\mathcal{S})^c} = 0$ and $\|\boldsymbol{\beta}^*_{\mathcal{M}\cup\mathcal{S}} - \boldsymbol{\beta}_{0,\mathcal{M}\cup\mathcal{S}}\|_2 \le \sqrt{(s+m)\log n/n}$ with probability tending to 1. By Condition (A1), we obtain

$$\|\boldsymbol{R}\|_\infty = O_p\left(\frac{r}{n}\right).$$

This together with (5.25) yields that

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \hat{\boldsymbol{\beta}}_{0,S} \end{pmatrix}^T \boldsymbol{R} \right\|_2 \le \|\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup\mathcal{S}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup\mathcal{S}}\|_1 \|\boldsymbol{R}\|_\infty = o_p\left(\frac{r}{n}\frac{\sqrt{r}}{\sqrt{n}}\sqrt{s+m}\right).$$

The last term is $o_p(\sqrt{r}/n)$ since $r \le s+m$ and $s+m = o(n^{1/3})$.

Similarly, we can show

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \hat{\boldsymbol{\beta}}_{0,S} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \hat{\boldsymbol{\beta}}_{0,S} \end{pmatrix} -$$
$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \hat{\boldsymbol{\beta}}_{0,S} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \hat{\boldsymbol{\beta}}_{0,S} \end{pmatrix} \Bigg\|_2 = o_p(\sqrt{r}).$$

As a result, we have

$$\begin{aligned} n\{L_n(\hat{\boldsymbol{\beta}}_0) - L_n(\hat{\boldsymbol{\beta}}_a)\} &= \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\} \\ &\quad - \frac{n}{2} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix}^T \boldsymbol{K}_n \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix} + o_p(\sqrt{r}). \end{aligned} \tag{5.26}$$

Recall that $\hat{\boldsymbol{\beta}}_a$ is the maximizer of $nL_n(\boldsymbol{\beta}) - n\sum_{j\notin\mathcal{M}} p_{\lambda_{n,a}}(|\beta_j|)$. By Theorem 2.1, we have with probability tending to 1 that $\min_{j\in S}|\hat{\beta}_{a,j}| \ge d_n$. Under the condition $p'_{\lambda_{n,a}}(d_n) = o((s+m)^{-1/2}n^{-1/2})$, we have

$$\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\} = n \begin{pmatrix} \boldsymbol{0} \\ \bar{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{a,S}, \lambda_{n,a}) \end{pmatrix} = o_p(n^{1/2}).$$

This together with (5.25) yields

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\} = o_p(\sqrt{r}).$$

By (5.26), we obtain that

$$n\{L_n(\hat{\boldsymbol{\beta}}_0) - L_n(\hat{\boldsymbol{\beta}}_a)\} = -\frac{n}{2} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix}^T \boldsymbol{K}_n \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix} + o_p(\sqrt{r}).$$

In view of (5.24), using similar arguments in (5.17), we can show that

$$\left| n \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix}^T \boldsymbol{K}_n \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \hat{\boldsymbol{\beta}}_{a,S} \end{pmatrix} \right.$$
$$\left. - \frac{1}{n} \left\| \boldsymbol{P}_n \boldsymbol{K}_n^{-1/2} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + n\boldsymbol{K}_n^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix} \boldsymbol{\Psi}^{-1} \boldsymbol{h}_n \right\|_2^2 \right| = o_p(\sqrt{r}).$$

As a result, we have

$$\frac{1}{n} \left\| \boldsymbol{P}_n \boldsymbol{K}_n^{-1/2} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + n\boldsymbol{K}_n^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix} \boldsymbol{\Psi}^{-1} \boldsymbol{h}_n \right\|_2^2$$
$$-2n\{L_n(\hat{\boldsymbol{\beta}}_a) - L_n(\hat{\boldsymbol{\beta}}_0)\} = o_p(\sqrt{r}).$$

By (5.16), this shows

$$\left| T_L - \frac{\phi_0}{\hat{\phi}} T_0 \right| = o_p(r).$$

Under the condition $|\hat{\phi} - \phi_0| = o_p(1)$, we can show $|T_0(1 - \phi_0/\hat{\phi})| = o_p(r)$. As a result, we have $T_L = T_0 + o_p(r)$.

Step 4: We first show the χ^2 approximation (3.5) holds for $T = T_0$. Recall that

$$T_0 = \frac{1}{\phi_0} \left\| \frac{1}{\sqrt{n}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n + \sqrt{n} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2^2.$$

By the definition of $\boldsymbol{\omega}_n$, we have

$$\frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_n = \frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}$$
$$= \sum_{i=1}^n \frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \{Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)\} \begin{pmatrix} X_{i,\mathcal{M}} \\ X_{i,S} \end{pmatrix} = \sum_{i=1}^n \boldsymbol{\xi}_i.$$

With some calculation, we can show that

$$\sum_i^n \mathrm{cov}(\boldsymbol{\xi}_i) = \boldsymbol{I}_r. \tag{5.27}$$

It follows from Condition (A3) that

$$\max_{i=1,\ldots,n} \mathrm{E} \left(\frac{|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|^3}{6M^3} M^2 \right)$$
$$\leq \max_{i=1,\ldots,n} \mathrm{E} \left\{ \exp\left(\frac{|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|}{M} \right) - 1 - \frac{|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|}{M} \right\} M^2 \leq \frac{v_0}{2}.$$

This implies $\max_{i=1,\ldots,n} \mathrm{E}|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|^3 = O(1)$.

Hence, with some calculations, we have

$$
\begin{aligned}
& r^{1/4} \sum_i^n \mathrm{E}\|\boldsymbol{\xi}_i\|_2^3 \\
= \; & \frac{r^{1/4}}{(n\phi_0)^{3/2}} \sum_i^n \mathrm{E} \left\| \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup S} \{Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)\} \right\|_2^3 \\
= \; & \frac{r^{1/4}}{(n\phi_0)^{3/2}} \sum_i^n \left\| \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup S} \right\|_2^3 \mathrm{E}|Y_i - \mu(\boldsymbol{\beta}_0^T \boldsymbol{X}_i)|^3 \\
= \; & O(1) \frac{r^{1/4}}{(n\phi_0)^{3/2}} \sum_i^n \left\| \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup S} \right\|_2^3 \\
\leq \; & O(1) \frac{r^{1/4}}{(n\phi_0)^{3/2}} \sum_i^n \left\| \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1/2} \right\|_2^3 \left\| \boldsymbol{K}_n^{-1/2} \boldsymbol{X}_{i,\mathcal{M}\cup S} \right\|_2^3 \\
\leq \; & O(1) \frac{r^{1/4}}{(n\phi_0)^{3/2}} \sum_{i=1}^n \left\{ (\boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}})^T \boldsymbol{K}_n^{-1} \boldsymbol{X}_{i,\mathcal{M}\cup\mathcal{S}} \right\}^{3/2} = o(1),
\end{aligned}
$$

where $O(1)$ denotes some positive constant, the first inequality follows from Cauchy-Schwarz inequality, the last inequality follows from the fact that

$$
\left\| \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1/2} \right\|_2^2 = \lambda_{\max} \left\{ \boldsymbol{\Psi}^{-1/2} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix} \boldsymbol{\Psi}^{-1/2} \right\} = 1,
$$

and the last equality is due to Condition (3.4).

This together with (5.27) and an application of Lemma S.3 in the supplementary material gives that

$$
\sup_{\mathcal{C}} \left| \Pr\left(\frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_0 \in \mathcal{C} \right) - \Pr(\boldsymbol{Z} \in \mathcal{C}) \right| \to 0, \tag{5.28}
$$

where $\boldsymbol{Z} \in \mathbb{R}^r$ stands for a mean zero Gaussian random vector with identity covariance matrix, and the supremum is taken over all convex sets $\mathcal{C} \in \mathbb{R}^r$.

Consider the following class of sets:

$$
\mathcal{C}_x = \left\{ \boldsymbol{z} \in \mathbb{R}^r : \left\| \boldsymbol{z} - \sqrt{\frac{n}{\phi_0}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n \right\|_2 \leq x \right\},
$$

indexed by $x \in \mathbb{R}$. It follows from (5.28) that

$$\sup_x \left| \Pr\left(\frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{\omega}_0 \in \mathcal{C}_x \right) - \Pr(\boldsymbol{Z} \in \mathcal{C}_x) \right| \to 0.$$

Note that $\frac{1}{\sqrt{n\phi_0}} \boldsymbol{\Psi}^{-1} \boldsymbol{\omega}_0 \in \mathcal{C}_x$ is equivalent to $T_0 \le x$, and $\Pr(\boldsymbol{Z} \in \mathcal{C}_x) = \Pr(\chi^2(r, \gamma_n) \le x)$ where $\gamma_n = n \boldsymbol{h}_n^T \boldsymbol{\Psi}^{-1/2} \boldsymbol{h}_n / \phi_0$. This implies

$$\sup_x |\Pr(T_0 \le x) - \Pr(\chi^2(r, \gamma_n) \le x)| \to 0. \tag{5.29}$$

Consider any statistic $T^* = T_0 + o_p(r)$. For any x and $\varepsilon > 0$, it follows from (5.29) that

$$\begin{aligned} &\Pr(\chi^2(r, \gamma_n) \le x - r\varepsilon) + o(1) \le \Pr(T_0 \le x - r\varepsilon) + o(1) \\ \le\ & \Pr(T^* \le x) \le \Pr(T_0 \le x + r\varepsilon) + o(1) \le \Pr(\chi^2(r, \gamma_n) \le x + r\varepsilon) + o(1). \end{aligned} \tag{5.30}$$

Besides, by Lemma S.4, we have

$$\lim_{\varepsilon \to 0} \limsup_n |\Pr(\chi^2(r, \gamma_n) \le x + r\varepsilon) - \Pr(\chi^2(r, \gamma_n) \le x - r\varepsilon)| \to 0. \tag{5.31}$$

Combining (5.30) with (5.31), we obtain that

$$\sup_x |\Pr(T^* \le x) - \Pr(\chi^2(r, \gamma_n) \le x)| \to 0. \tag{5.32}$$

In the first three steps, we have shown $T_0 = T_S + o_p(1) = T_W + o_p(1) = T_L + o_p(1)$. This together with (5.32) implies that the χ^2 approximation holds for our partial penalized Wald, score and likelihood ratio statistics. The proof is hence completed.

5.1. *Proof of Lemma 5.1.* Assertion (5.1) is directly implies by Condition (A1). This means the square root of the maximum eigenvalue of $\boldsymbol{K}_n$ is $O(1)$. By definition, this proves (5.2). Under Condition (A1), we have $\lambda_{\max}(\boldsymbol{K}_n^{-1}) = O(1)$. Using the same arguments, we have $\lambda_{\max}(\boldsymbol{K}_n^{-1/2}) = O(1)$. Hence, (5.3) is proven. We now show (5.4) holds. It follows from the condition $\lambda_{\max}\left((\boldsymbol{C}\boldsymbol{C}^T)^{-1}\right) = O(1)$ in Condition (A4) that $\liminf_n \lambda_{\min}(\boldsymbol{C}\boldsymbol{C}^T)^{-1} > 0$, and hence

$$a_0 \stackrel{\Delta}{=} \liminf_n \inf_{\boldsymbol{a} \in \mathbb{R}^r : \|\boldsymbol{a}\|_2 = 1} \|\boldsymbol{C}^T \boldsymbol{a}\|_2^2 = \liminf_n \inf_{\boldsymbol{a} \in \mathbb{R}^r : \|\boldsymbol{a}\|_2 = 1} \boldsymbol{a}^T \boldsymbol{C}\boldsymbol{C}^T \boldsymbol{a} > 0.$$

This implies that for sufficiently large n, we have

$$\|\boldsymbol{C}^T \boldsymbol{a}\|_2 > \sqrt{a_0/2} \|\boldsymbol{a}\|_2, \quad \forall \boldsymbol{a} \neq 0. \tag{5.33}$$

By (5.1), we have $\liminf_n \lambda_{\min}(\boldsymbol{\Omega}_n) > 0$, or equivalently,

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1} \liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_n\boldsymbol{a} > 0.$$

Hence, we have

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1,\boldsymbol{a}^c_{J_0}=0} \liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_n\boldsymbol{a} > 0,$$

where $J_0 = [1,\ldots,m]$. Note that this implies

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1} \liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_{mm}\boldsymbol{a} > 0.$$

Therefore, we obtain

$$\liminf_n \lambda_{\min}(\boldsymbol{\Omega}_{mm}) > 0. \tag{5.34}$$

Combining this together with (5.33) yields

$$\inf_{\boldsymbol{a}\in\mathbb{R}^r:\|\boldsymbol{a}\|_2=1} \liminf_n \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\boldsymbol{a} \geq \inf_{\boldsymbol{a}\in\mathbb{R}^m:\|\boldsymbol{a}\|_2=\sqrt{a_0/2}} \liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_{mm}\boldsymbol{a} > 0.$$

By definition, this suggests

$$\liminf_n \lambda_{\min}(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T) > 0,$$

or equivalently,

$$\lambda_{\max}\left((\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}\right) = O(1).$$

This gives (5.4).

Using Cauchy-Schwarz inequality, we have

$$\|(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1/2}\boldsymbol{C}\|_2 \leq \underbrace{\|(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1/2}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\|_2}_{I_1}\underbrace{\|\boldsymbol{\Omega}_{mm}^{-1/2}\|_2}_{I_2}.$$

Observe that

$$I_1^2 = \lambda_{\max}\left((\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1/2}\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1/2}\right) = 1. \tag{5.35}$$

Besides, by (5.34), we have

$$I_2^2 = \lambda_{\max}\left((\boldsymbol{\Omega}_{mm})^{-1}\right) = O(1),$$

which together with (5.35) implies that $I_1 I_2 = O(1)$. This proves (5.5).

We now show (5.6) holds. Assume for now, we have

$$\|\boldsymbol{K}_n - \widehat{\boldsymbol{K}}_{n,a}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right), \tag{5.36}$$

where

$$\widehat{\boldsymbol{K}}_{n,a} = \frac{1}{n}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_S \\ \boldsymbol{X}_S^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_S^T\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a)\boldsymbol{X}_S \end{pmatrix}.$$

Note that

$$\begin{aligned} \liminf_n \lambda_{\min}(\widehat{\boldsymbol{K}}_{n,a}) &\geq \liminf_n \inf_{\substack{\boldsymbol{a}\in\mathbb{R}^{m+s} \\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^T\boldsymbol{K}_n\boldsymbol{a} - \limsup_n \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{m+s} \\ \|\boldsymbol{a}\|_2=1}} |\boldsymbol{a}^T(\widehat{\boldsymbol{K}}_{n,a} - \boldsymbol{K}_n)\boldsymbol{a}| \\ &\geq \liminf_n \lambda_{\min}(\boldsymbol{K}_{n,a}) - \limsup_n \|\boldsymbol{K}_n - \widehat{\boldsymbol{K}}_{n,a}\|_2. \end{aligned}$$

Under Condition (A1), we have $\liminf_n \lambda_{\min}(\boldsymbol{K}_n) > 0$. Under the condition $\max(s,m) = o(n^{1/2})$, this together with (5.36) implies

$$\liminf_n \lambda_{\min}(\widehat{\boldsymbol{K}}_{n,a}) > 0, \tag{5.37}$$

with probability tending to 1. Hence, we have

$$\begin{aligned} &\|\boldsymbol{K}_n^{-1} - \widehat{\boldsymbol{K}}_{n,a}^{-1}\|_2 = \|\boldsymbol{K}_n^{-1}(\boldsymbol{K}_n - \widehat{\boldsymbol{K}}_{n,a})\widehat{\boldsymbol{K}}_{n,a}^{-1}\|_2 \\ \leq\ &\lambda_{\max}(\boldsymbol{K}_n^{-1})\|\boldsymbol{K}_n - \widehat{\boldsymbol{K}}_{n,a}\|_2\lambda_{\max}(\widehat{\boldsymbol{K}}_n^{-1}) = O_p\left(\frac{s+m}{\sqrt{n}}\right). \end{aligned} \tag{5.38}$$

By Lemma S.2, this gives

$$\sup_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T(\boldsymbol{K}_n^{-1} - \widehat{\boldsymbol{K}}_{n,a}^{-1})\boldsymbol{a}| = O_p\left(\frac{s+m}{\sqrt{n}}\right),$$

and hence,

$$\sup_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1,\boldsymbol{a}_{J_0}^c=0} |\boldsymbol{a}^T(\boldsymbol{K}_n^{-1} - \widehat{\boldsymbol{K}}_{n,a}^{-1})\boldsymbol{a}| = O_p\left(\frac{s+m}{\sqrt{n}}\right),$$

where $J_0 = [1,\ldots,m]$. Using Lemma S.2 again, we obtain

$$\|(\boldsymbol{K}_n^{-1})_{J_0,J_0} - (\widehat{\boldsymbol{K}}_{n,a}^{-1})_{J_0,J_0}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right). \tag{5.39}$$

By definition, we have $\boldsymbol{\Omega}_{mm} = (\boldsymbol{K}_n^{-1})_{J_0,J_0}$. According to Theorem 2.1, we have that with probability tending to 1, $\widehat{S}_a = S$ where $\widehat{S}_a = \{j \in \mathcal{M}^c :$

$\hat{\beta}_{a,j} \neq 0\}$. When $\widehat{S}_a = S$, we have $\widehat{\boldsymbol{K}}_{n,a}^{-1} = \widehat{\boldsymbol{\Omega}}_a$ and $(\widehat{\boldsymbol{K}}_{n,a}^{-1})_{J_0,J_0} = \widehat{\boldsymbol{\Omega}}_{a,mm}$. Therefore, by (5.39), we have

$$\|\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right).$$

Using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} & \|\boldsymbol{\Omega}_{mm}^{-1/2}(\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm})\boldsymbol{\Omega}_{mm}^{-1/2}\|_2 \\ \leq\ & \|\boldsymbol{\Omega}_{mm}^{-1/2}\|_2^2 \|\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right), \end{aligned} \tag{5.40}$$

by (5.34). Let $\boldsymbol{\Psi} = \boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$, we obtain

$$\begin{aligned} & \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}(\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm})\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2}\|_2 \\ \leq\ & \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{\Omega}_{mm}^{-1/2}(\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm})\boldsymbol{\Omega}_{mm}^{-1/2}\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2}\|_2 \\ \leq\ & \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\|_2^2 \|\boldsymbol{\Omega}_{mm}^{-1/2}(\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm})\boldsymbol{\Omega}_{mm}^{-1/2}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right), \end{aligned} \tag{5.41}$$

by (5.40) and that

$$\|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\|_2^2 = \lambda_{\max}\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1/2}\right) = O(1).$$

Similar to (5.37), by (5.41), we can show that

$$\liminf_n \lambda_{\min}\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2}\right) > 0. \tag{5.42}$$

Combining (5.41) together with (5.42), we obtain

$$\begin{aligned} & \|(\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2})^{-1} - \boldsymbol{I}_m\|_2 \\ \leq\ & \|(\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2})^{-1}\|_2 \|\boldsymbol{\Psi}^{-1/2}\boldsymbol{C}(\boldsymbol{\Omega}_{mm} - \widehat{\boldsymbol{\Omega}}_{a,mm})\boldsymbol{C}^T\boldsymbol{\Psi}^{-1/2}\|_2 \\ =\ & O_p\left(\frac{s+m}{\sqrt{n}}\right). \end{aligned}$$

This proves (5.6).

Similar to (5.38), we can show

$$\|\boldsymbol{K}_n^{-1} - \widehat{\boldsymbol{K}}_{n,0}^{-1}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right).$$

By (5.2), we obtain

$$\|\boldsymbol{I} - \boldsymbol{K}_n^{1/2}\widehat{\boldsymbol{K}}_{n,0}^{-1}\boldsymbol{K}_n^{1/2}\|_2 \leq \|\boldsymbol{K}_n^{1/2}\|_2 \|\boldsymbol{K}_n^{-1} - \widehat{\boldsymbol{K}}_{n,0}^{-1}\|_2 \|\boldsymbol{K}_n^{1/2}\|_2 = O_p\left(\frac{s+m}{\sqrt{n}}\right).$$

This proves (5.7).

It remains to show (5.36). Since $\boldsymbol{K}_n$ and $\widehat{\boldsymbol{K}}_{n,a}$ are symmetric, by Lemma S.5, it suffices to show

$$\|\boldsymbol{K}_n - \widehat{\boldsymbol{K}}_{n,a}\|_\infty = O_p\left(\frac{s+m}{\sqrt{n}}\right).$$

By definition, this requires to show

$$\max_{j\in\mathcal{S}\cup\mathcal{M}} \|(\boldsymbol{X}^j)^T\{\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) - \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\|_1 = O_p\left(\sqrt{n}(s+m)\right),$$

For any vector $\boldsymbol{a}\in\mathbb{R}^q$, we have $\|\boldsymbol{a}\|_1 \le \sqrt{q}\|\boldsymbol{a}\|_2$. Hence, it suffices to show

(5.43)
$$\max_{j\in\mathcal{S}\cup\mathcal{M}} \|(\boldsymbol{X}^j)^T\{\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) - \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\|_2 = O_p\left(\sqrt{n(s+m)}\right).$$

Using Taylor's theorem, we have

(5.44)
$$\begin{aligned}
&(\boldsymbol{X}^j)^T\{\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) - \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}} \\
\le\ & \int_0^1 (\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_0)^T\boldsymbol{X}\mathrm{diag}\left\{\boldsymbol{X}^j \circ b'''(\boldsymbol{X}\{t\hat{\boldsymbol{\beta}}_a + (1-t)\boldsymbol{\beta}_0\})\right\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}dt.
\end{aligned}$$

By Theorem 2.1, we have $\Pr\left(\hat{\boldsymbol{\beta}}_a \in \mathcal{N}_0\right) \to 1$. Hence, we have

$$\Pr\left(\bigcup_{t\in[0,1]}\left\{t\hat{\boldsymbol{\beta}}_a + (1-t)\boldsymbol{\beta}_0 \in \mathcal{N}_0\right\}\right) \to 1.$$

By Condition (A1),

$$\sup_{t\in[0,1]} \lambda_{\max}\left\{\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\mathrm{diag}\left(\boldsymbol{X}^j \circ b'''(\boldsymbol{X}\{t\hat{\boldsymbol{\beta}}_a + (1-t)\boldsymbol{\beta}_0\})\right)\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\right\} = O_p(n).$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\|(\boldsymbol{X}^j)^T\{\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) - \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\|_2 \\
\le\ & \sup_{t\in[0,1]}\left\|(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_0)^T\boldsymbol{X}\mathrm{diag}\left\{\boldsymbol{X}^j \circ b'''(\boldsymbol{X}\{t\hat{\boldsymbol{\beta}}_a + (1-t)\boldsymbol{\beta}_0\})\right\}\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\right\|_2 \\
\le\ & \|\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_0\|_2 \sup_{t\in[0,1]} \lambda_{\max}\left\{\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\mathrm{diag}\left(\boldsymbol{X}^j \circ b'''(\boldsymbol{X}\{t\hat{\boldsymbol{\beta}}_a + (1-t)\boldsymbol{\beta}_0\})\right)\boldsymbol{X}_{\mathcal{M}\cup\mathcal{S}}\right\} \\
=\ & O_p(\sqrt{n(s+m)}).
\end{aligned}$$

This proves (5.43).

Acknowledgements. The authors wish to thank the Associate Editor and anonymous referees for their constructive comments, which lead to significant improvement of this work.

SUPPLEMENTARY MATERIAL

Supplement to "Partial penalization for high dimensional testing with linear constraints":
(doi: COMPLETED BY THE TYPESETTER; .pdf). This supplemental material includes power comparions with existing test statistics, additional numerical studies on Poisson regression and a real data application, discussions of Condition (A1)-(A4), some technical lemmas and the proof of Theorem 2.1.

References.

Bentkus, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.* **49** 400–410.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3** 1–122.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253.

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

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552.

Fang, E. X., Ning, Y. and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. Roy. Statist. Soc. Ser. B* **79** 1415–1437.

Ghosh, B. K. (1973). Some monotonicity theorems for χ^2, F and t distributions with applications. *J. Roy. Statist. Soc. Ser. B* **35** 480–492.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468.

McCullagh and Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

Shi, C., Song, R., Chen, Z. and Li, R. (2018). Supplement to "Partial penalization for high dimensional testing with linear constraints".

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418.

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

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

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

Wang, S. and Cui, H. (2013). Partial Penalized Likelihood Ratio Test under Sparse Case. *arXiv preprint arXiv:1312.3723*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768.

Chengchun Shi and Rui Song
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203
USA
E-mail: cshi4@ncsu.edu
E-mail: rsong@ncsu.edu

Zhao Chen and Runze Li
Department of Statistics,
and The Methodology Center
the Pennsylvania State University,
University Park, PA 16802-2111
USA
E-mail: zuc4@psu.edu
E-mail: rzli@psu.edu