Journal of Machine Learning Research 17 (2016) 1-34 Submitted 2/16; Published 11/16

Guarding against Spurious Discoveries in High Dimensions

Jianqing Fan JQFAN@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton Univeristy
Princeton, NJ 08544, USA

Wen-Xin Zhou WENXINZ@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton Univeristy
Princeton, NJ 08544, USA

Editor: Hui Zou

Abstract

Many data mining and statistical machine learning algorithms have been developed to select a subset of covariates to associate with a response variable. Spurious discoveries can easily arise in high-dimensional data analysis due to enormous possibilities of such selections. How can we know statistically our discoveries better than those by chance? In this paper, we define a measure of goodness of spurious fit, which shows how good a response variable can be fitted by an optimally selected subset of covariates under the null model, and propose a simple and effective LAMM algorithm to compute it. It coincides with the maximum spurious correlation for linear models and can be regarded as a generalized maximum spurious correlation. We derive the asymptotic distribution of such goodness of spurious fit for generalized linear models and L_1 regression. Such an asymptotic distribution depends on the sample size, ambient dimension, the number of variables used in the fit, and the covariance information. It can be consistently estimated by multiplier bootstrapping and used as a benchmark to guard against spurious discoveries. It can also be applied to model selection, which considers only candidate models with goodness of fits better than those by spurious fits. The theory and method are convincingly illustrated by simulated examples and an application to the binary outcomes from German Neuroblastoma Trials.

Keywords: Bootstrap, Gaussian approximation, generalized linear models, L_1 regression, model selection, sparsity, spurious correlation, spurious fit

1. Introduction

Technological developments in science and engineering lead to collections of massive amounts of high-dimensional data. Scientific advances have become more and more data-driven, and researchers have been making efforts to understand the contemporary large-scale and complex data. Among these efforts, variable selection plays a pivotal role in high-dimensional statistical modeling, where the goal is to extract a small set of explanatory variables that are associated with given responses such as biological, clinical, and societal outcomes. Toward this end, in the past two decades, statisticians have developed many data learning methods and algorithms, and have applied them to solve problems arising from diverse fields of sciences, engineering and humanities, ranging from genomics, neurosciences and health

sciences to economics, finance and machine learning. For an overview, see Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015).

Linear regression is often used to investigate the relationship between a response variable Y and explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. In the high-dimensional linear model $Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$, the coefficient $\boldsymbol{\beta}^*$ is assumed to be sparse with support $S_0 = \mathrm{supp}(\boldsymbol{\beta}^*)$. Variable selection techniques such as the forward stepwise regression, the Lasso (Tibshirani, 1996) and folded concave penalized least squares (Fan and Li, 2001; Zou and Li, 2008) are frequently used. However, it has been recently noted in Fan, Guo and Hao (2012) that high dimensionality introduces large spurious correlations between response and unrelated covariates, which may lead to wrong statistical inference and false scientific discoveries. As an illustration, Fan, Shao and Zhou (2015) considered a real data example using the gene expression data from the international 'HapMap' project (Thorisson et al., 2005). There, the sample correlation between the observed and post-Lasso fitted responses is as large as 0.92. While conventionally it is a common belief that a correlation of 0.92 between the response and a fit is noteworthy, in high-dimensional scenarios, this intuition may no longer be true. In fact, even if the response and all the covariates are scientifically independent in the sense that $\boldsymbol{\beta}^* = \mathbf{0}$, simply by chance, some covariates will appear to be highly correlated with the response. As a result, the findings obtained via any variable selection techniques are hardly impressive unless they are proven to be better than by chance. To simplify terminology, in this paper we say that the discovery (by a variable selection method) is spurious if it is no better than by chance.

To guard against spurious discoveries, one naturally asks how good a response can be fitted by optimally selected subsets of covariates, even when the response variable and the covariates are not causally related to each other, that is, when they are independent. Such a measure of the goodness of spurious fit (GOSF) is a random variable whose distribution can provide a benchmark to gauge whether the discoveries by statistical machine learning methods any better than a spurious fit (chance). Measuring such a goodness of spurious fit and estimating its theoretical distributions are the aims of this paper. This problem arises from not only high-dimensional linear models and generalized linear models, but also robust regression and other statistical model fitting. To formally measure the degree of spurious fit, Fan, Shao and Zhou (2015) derived the distributions of maximum spurious correlations, which provide a benchmark to assess the strength of the spurious associations (between response and independent covariates) and to judge whether discoveries by a certain variable selection technique are any better than by chance.

The response, however, is not always a quantitative value. Instead, it is often binary; for example, positive or negative, presence or absence and success or failure. In this regard, generalized linear models (GLIM) serve as a flexible parametric approach to modeling the relationship between explanatory and response variables (McCullagh and Nelder, 1989). Prototypical examples include linear, logistic and Poisson regression models which are frequently encountered in practice.

In GLIM, the relationship between the response and covariates is more complicated and cannot be effectively measured via Pearson correlation coefficient, which is essentially a measure of the linear correlation between two variables. We need to extend the concept of spurious correlation or the measure of goodness of spurious fit to more general models and study its null distribution. A natural measure of goodness of fit is the likelihood

ratio statistic, denoted by $\mathcal{LR}_n(s,p)$, where n is the sample size and s is size of optimally fitted model. It measures the goodness of spurious fit when $\boldsymbol{X}$ and Y are independent. This generalization is consistent with the spurious correlation studied in Fan, Shao and Zhou (2015), that is, applying $\mathcal{LR}_n(s,p)$ to linear regression yields the maximum spurious correlation. We plan to study the limiting null distribution of $2\mathcal{LR}_n(s,p)$ under various scenarios. This reference distribution then serves as a benchmark to determine whether the discoveries are spurious.

To gain further insights, let us illustrate the issue by using the gene expression profiles for $10,707$ genes from 251 patients in the German Neuroblastoma Trials NB90-NB2004 (Oberthuer et al., 2006). The response labeled as "3-year event-free survival" (3-year EFS) is a binary outcome indicating whether each patient survived 3 years after the diagnosis of neuroblastoma. Excluding five outlier arrays, there are 246 subjects (101 females and 145 males) with 3-year EFS information available. Among them, 56 are positives and 190 are negatives. We apply Lasso using the logistic regression model with tuning parameter selected via ten-fold cross validation (40 genes are selected). The fitted likelihood ratio $2\widehat{\mathcal{LR}} = 211.96$. To judge the credibility of the finding of these 40 genes, we should compare the value 211.96 with the distribution of the Goodness Of Spurious Fit (GOSF) $2\mathcal{LR}_n(s,p)$ when $\boldsymbol{X}$ and Y are indeed independent, where $n = 246$, $p = 10,707$ and $s = 40$. This requires some new methodology and technical work. Figure 1 shows the distribution of the GOSF estimated by our proposed method below and indicates how abnormal the value 211.96 is. It can be concluded that the goodness of fit to the binary outcome is not statistically significantly better than GOSF.

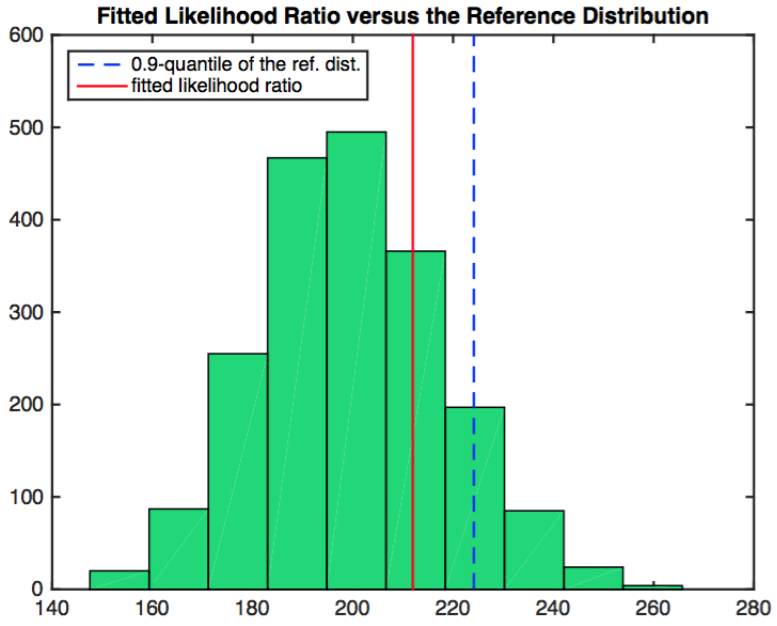

Figure 1: Lasso fitted likelihood ratio $2\widehat{\mathcal{LR}}$ in comparison to the distribution of GOSF $2\mathcal{LR}_n(s,p)$ with $n = 246$, $p = 10,707$ and $s = 40$.

The above result shows that the 10-fold cross-validation chooses a too large model with 40 variables. This prompts us to reduce the model sizes along the Lasso path such that their fits are better than GOSF. The results are reported in Table 2. The largest model along the LASSO path that fits better than GOSF has model size 17. We can use the cross-validation to select a model with model size no more than 17 or to select a best model among all models that fit better than GOSF. This is another important application of our method.

1.1 Structure of the paper

In Section 2, we introduce a general measure of spurious fit via generalized likelihood ratios, which extends the concept of spurious correlation in the linear model to more general models, including generalized linear models and robust linear regression. We also introduce a local adaptive majorization-minimization (LAMM) algorithm to compute the GOSF. Section 3 presents the main results on the limiting laws of goodness of spurious fit and their bootstrap approximations. For conducting inference, we use the proposed LAMM algorithm to compute the bootstrap statistic. In Section 4, we discuss an application of our theoretical findings to high-dimensional statistical inference and model selection. Section 5 presents numerical studies. Proofs of the main results, Theorems 2 and 6, are provided in Section 6; in each case, we break down the key steps in a series of lemmas with proofs deferred to the appendix.

1.2 Notations

We collect standard pieces of notation here for readers' convenience. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n/b_n \le C$ for all sufficiently large n; we write $a_n \asymp b_n$ if there exist constants $C_1, C_2 > 0$ such that, for all n large enough, $C_1 \le a_n/b_n \le C_2$; and we write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, respectively. For $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$.

For every positive integer ℓ, we write $[\ell] = \{1, 2, \ldots, \ell\}$, and for any set S, we use S^{c} to denote its complement and $|S|$ for its cardinality. For any real-valued random variable X, its sub-Gaussian norm is defined by $\|X\|_{\psi_2} = \sup_{\ell \ge 1} \ell^{-1/2} (\mathbb{E}|X|^\ell)^{1/\ell}$. We say that a random variable X is sub-Gaussian if $\|X\|_{\psi_2} < \infty$.

Let p, q be two positive integers. For every p-vector $\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}}$, we define its ℓ_q-norm to be $\|\boldsymbol{u}\|_q = \big(\sum_{i=1}^p |u_i|^q\big)^{1/q}$, and set $\|\boldsymbol{u}\|_0 = \sum_{i=1}^p I\{u_i \neq 0\}$. Let $\mathbb{S}^{p-1} = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^p$. Moreover, for each subset $S \subseteq [p]$ with $|S| = s \in [p]$, we denote by $\boldsymbol{u}_S$ the s-variate sub-vector of $\boldsymbol{u}$ containing only the coordinates indexed by S. We use $\|\mathbf{M}\|$ to denote the spectral norm of a matrix $\mathbf{M}$.

2. Goodness of spurious fit

Let $Y, Y_1, \ldots, Y_n$ be independent and identically distributed (i.i.d.) random variables with mean zero and variance $\sigma^2 > 0$, and $\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. p-dimensional random vectors. We write

$$\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}, \mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p} \text{ and } \boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}, \quad i = 1, \ldots, n.$$

For $s \in [p]$, the maximum s-multiple correlation between Y and $\boldsymbol{X}$ is given by

$$\widehat{R}_n(s,p) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^p:\|\boldsymbol{\alpha}\|_0\le s} \widehat{\mathrm{corr}}_n(Y, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}), \tag{1}$$

where $\widehat{\mathrm{corr}}_n(\cdot,\cdot)$ denotes the sample Pearson correlation coefficient. When Y and $\boldsymbol{X}$ are independent, we regard $\widehat{R}_n(s,p)$ as the maximum spurious (multiple) correlation. The limiting distribution of $\widehat{R}_n(s,p)$ is studied in Cai and Jiang (2012) and Fan, Guo and Hao (2012) when $s = 1$ and $X \sim N(\mathbf{0}, \mathbf{I}_p)$ (the standard normal distribution in $\mathbb{R}^p$), and later in Fan, Shao and Zhou (2015) under a general setting where $s \ge 1$ and $\boldsymbol{X}$ is sub-Gaussian with an arbitrary covariance matrix.

For binary data, the sample Pearson correlation is not effective for measuring the regression effect. We need a new metric. In classical regression analysis, the multiple correlation coefficient, also known as the R^2, is the proportion of variance explained by the regression model. For each submodel $S \subseteq [p]$, its R^2 statistic can be computed as

$$R^2_S = \max_{\boldsymbol{\theta}\in\mathbb{R}^s} \widehat{\mathrm{corr}}^2_n(Y, \boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{\theta}). \tag{2}$$

Then, the maximum s-multiple correlation $\widehat{R}_n(s,p)$ can be expressed as the maximum R^2 statistic:

$$\widehat{R}^2_n(s,p) = \max_{S\subseteq[p]:|S|=s} R^2_S. \tag{3}$$

The concept of R^2 can be extended to more general models. For binary response models, Maddala (1983) suggested the following generalization: $-\log(1-R^2) = \frac{2}{n}\{\ell(\widehat{\boldsymbol{\beta}}) - \ell(\mathbf{0})\}$, where $\ell(\widehat{\boldsymbol{\beta}}) = \log L(\widehat{\boldsymbol{\beta}})$ and $\ell(\mathbf{0}) = \log L(\mathbf{0})$ denote the log-likelihoods of the fitted and the null model, respectively. This motivates us to use the likelihood ratio as a generalization of the goodness of fit beyond the linear model.

Let $L_n(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathbb{R}^p$ be the negative logarithm of a quasi-likelihood process of the sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. For a given model size $s \in [p]$, the best subset fit is $\widehat{\boldsymbol{\beta}}(s) := \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} L_n(\boldsymbol{\beta})$. The goodness of such a fit, in comparison with the baseline fit $L_n(\mathbf{0})$, can be measured by

$$\mathcal{LR}_n(s,p) := L_n(\mathbf{0}) - L_n(\widehat{\boldsymbol{\beta}}(s)) = L_n(\mathbf{0}) - \min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} L_n(\boldsymbol{\beta}). \tag{4}$$

When $\boldsymbol{X}$ and Y are independent, it becomes the Goodness OF Spurious Fit (GOSF). According to (2) and (3), this definition is consistent with the maximum spurious correlation when it is applied to the linear model with Gaussian quasi-likelihood, where $L_n(\boldsymbol{\beta}; \beta_0, \sigma) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\|\boldsymbol{Y} - \beta_0 - \mathbb{X}\boldsymbol{\beta}\|_2^2/\sigma^2$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$.

Throughout, we refer to $L_n(\cdot)$ as the loss function which is assumed to be convex. This setup encompasses the generalized linear models (McCullagh and Nelder, 1989) with $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \{b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - Y_i\,\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}$ under the canonical link where $b(\cdot)$ is a model-dependent convex function (we take the dispersion parameter as one, as we don't consider the dispersion issue), robust regression with $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n |Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}|$, the hinge loss $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n (1 - Y_i\,\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})_+$ in the support vector machine (Vapnik, 1995) and exponential loss $L_n(\boldsymbol{\beta}) =$

$\sum_{i=1}^n \exp(-Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})$ in AdaBoost (Freund and Schapire, 1997) in classification with Y taking values ± 1.

The prime goal of this paper is to derive the limiting laws of GOSF $\mathcal{LR}_n(s,p)$ in the null setting where the response Y and the explanatory variables $\boldsymbol{X}$ are independent. Here, both s and p can depend on n, as we shall use double-array asymptotics. We will mainly focus on the GLIM and robust linear regression that are of particular interest in statistics.

2.1 Generalized linear models

Recall that $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ are i.i.d. copies of $(Y, \boldsymbol{X})$. Assume that the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^p$ belongs to the canonical exponential family with the probability density function taking the form (McCullagh and Nelder, 1989)

$$f(y; \boldsymbol{x}, \boldsymbol{\beta}^*) = \exp\big[\{y\, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - b(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\}/\phi + c(y,\phi)\big], \tag{5}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ is the unknown p-dimensional vector of regression coefficients, and $\phi > 0$ is the dispersion parameter. The log-likelihood function with respect to the given data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ is $\sum_{i=1}^n c(Y_i,\phi) + \phi^{-1}\sum_{i=1}^n \{Y_i\, \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}$. For simplicity, we take $\phi = 1$ with the exception that in the linear model with Gaussian noise, $\phi = \sigma^2$ is the variance. Two other showcases are

1. Logistic regression: $b(u) = \log(1+e^u)$, $u \in \mathbb{R}$ and $\phi = 1$.
2. Poisson regression: $b(u) = e^u$, $u \in \mathbb{R}$ and $\phi = 1$.

In GLIM, the loss function is $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \{b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - Y_i\, \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}$. By (4), the generalized measure of goodness of fit for GLIM is

$$\mathcal{LR}_n(s,p) = nb(0) - \min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0 \le s} L_n(\boldsymbol{\beta}). \tag{6}$$

In Section 3, we derive under mild regularity conditions the limiting distribution of GOSF $\mathcal{LR}_n(s,p)$ in the null model. This extends the classical Wilks theorem (Wilks, 1938). Here, we interpret $\mathcal{LR}_n(s,p)$ as the degree of spuriousness caused by the high-dimensionality.

2.2 L_1 regression

In this section, we revisit the high-dimensional linear model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \quad \text{or} \quad Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i,\ i = 1, \ldots, n, \tag{7}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is the response vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is the n-vector of measurement errors. Robustness considerations lead to least absolute deviation (LAD) regression and more generally quantile regression (Koenker, 2005). For simplicity, we consider the ℓ_1-loss $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n |Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}|$, $\boldsymbol{\beta} \in \mathbb{R}^p$. The generalized measure of goodness of fit (4) now becomes

$$\mathcal{LR}_n(s,p) = \|\boldsymbol{Y}\|_1 - \min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0 \le s} L_n(\boldsymbol{\beta}). \tag{8}$$

The limiting distribution of GOSF $\mathcal{LR}_n(s,p)$ is studied in Section 3.4.

In particular, if $\varepsilon_1, \ldots, \varepsilon_n$ in (7) are i.i.d. from the double exponential distribution with the density $f_\varepsilon(u) = \frac{1}{2}e^{-|u|}$, $u \in \mathbb{R}$, the ℓ_1-loss $L_n(\cdot)$ corresponds to the negative log-likelihood function. In general, we assume that the regression error ε_i has median zero, that is, $\mathbb{P}(\varepsilon_i \le 0) = \frac{1}{2}$. Hence, the conditional median of Y_i given $\boldsymbol{X}_i$ is $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*$ for $i \in [n]$, and $\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\mathbb{E}_{\boldsymbol{X}}\{L_n(\boldsymbol{\beta})\}$, where $\mathbb{E}_{\boldsymbol{X}}(\cdot) = \mathbb{E}(\cdot\,|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ denotes the conditional expectation given $\{\boldsymbol{X}_i\}_{i=1}^n$.

2.3 An LAMM algorithm

The computation of the best subset regression coefficient $\widehat{\boldsymbol{\beta}}(s)$ in (4) requires solving a combinatorial optimization problem with a cardinality constraint, and therefore is NP-hard. In the following, we suggest a fast and easily implementable method, which combines the forward selection (stepwise addition) algorithm and a local adaptive majorization-minimization (LAMM) algorithm (Lange, Hunter and Yang, 2000; Fan et al., 2015) to provide an approximate solution.

Our optimization problem is $\min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} f(\boldsymbol{\beta})$, where $f(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta})$. We say that a function $g(\boldsymbol{\beta}\,|\,\boldsymbol{\beta}^{(k)})$ majorizes $f(\boldsymbol{\beta})$ at the point $\boldsymbol{\beta}^{(k)}$ if $f(\boldsymbol{\beta}^{(k)}) = g(\boldsymbol{\beta}^{(k)}\,|\,\boldsymbol{\beta}^{(k)})$ and $f(\boldsymbol{\beta}) \le g(\boldsymbol{\beta}\,|\,\boldsymbol{\beta}^{(k)})$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. An majorization-minimization (MM) algorithm initializes at $\boldsymbol{\beta}^{(0)}$ and then iteratively computes $\boldsymbol{\beta}^{(k+1)} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} g(\boldsymbol{\beta}\,|\,\boldsymbol{\beta}^{(k)})$. The target value of such an algorithm is non-increasing since

$$f(\boldsymbol{\beta}^{(k+1)}) \overset{\text{majorization}}{\le} g(\boldsymbol{\beta}^{(k+1)}\,|\,\boldsymbol{\beta}^{(k)}) \overset{\text{minimization}}{\le} g(\boldsymbol{\beta}^{(k)}\,|\,\boldsymbol{\beta}^{(k)}) \overset{\text{initialization}}{=} f(\boldsymbol{\beta}^{(k)}). \tag{9}$$

We now majorize $f(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}^{(k)}$ by an isotropic quadratic function

$$g_\lambda(\boldsymbol{\beta}\,|\,\widehat{\boldsymbol{\beta}}^{(k)}) = f(\boldsymbol{\beta}) + \left\langle \nabla f(\widehat{\boldsymbol{\beta}}^{(k)}), \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(k)} \right\rangle + \frac{\lambda}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(k)}\|_2^2, \quad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{10}$$

This is a valid majorization as long as $\lambda \ge \max_{\boldsymbol{\beta}} \|\nabla^2 f(\boldsymbol{\beta})\|$ (this will be relaxed below). The isotropic form on the right-hand side of (10) allows a simple analytic solution given by

$$\widehat{\boldsymbol{\beta}}_\lambda^{(k+1)} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} g(\boldsymbol{\beta}\,|\,\boldsymbol{\beta}^{(k)}) = \big\{\widehat{\boldsymbol{\beta}}^{(k)} - \lambda^{-1}\nabla f(\widehat{\boldsymbol{\beta}}^{(k)})\big\}_{[1:s]}.$$

Here, we used the notation that for any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta}_{[1:s]} \in \mathbb{R}^p$ retains the s largest (in magnitude) entries of $\boldsymbol{\beta}$ and assigns the rest to zero.

Remark 1 To implement the MM algorithm, we need to compute the gradient of the objective function of interest. In the L_1 regression, the loss function $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n |Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}|$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is not differentiable everywhere. Recall that the subdifferential of the absolute function $h(x) = |x|$, $x \in \mathbb{R}$ is given by

$$\partial h(x) = \begin{cases} \{1\}, & \text{if } x > 0, \\ [-1, 1], & \text{if } x = 0, \\ \{-1\}, & \text{if } x < 0. \end{cases}$$

With slight abuse of notation, we suggest a randomized algorithm using the stochastic subgradient $\nabla L_n(\boldsymbol{\beta}) = \sum_{i=1}^n I(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} > 0) - I(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} > 0) + U_i I(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} = 0)$, where $U_1, \ldots, U_n$ are i.i.d. random variables uniformly distributed on $[-1, 1]$.

We propose to use the stepwise forward selection algorithm to compute an initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$. As the MM algorithm decreases the target value as shown in (9), the resulting target value is no larger than that produced by the stepwise forward selection algorithm.

To properly choose the isotropic parameter $\lambda > 0$ without computing the maximum eigenvalue, we use the local adaptive procedure as in Fan et al. (2015). Note that, in order to have a non-increasing target value, the majorization is not actually required. As long as $f(\boldsymbol{\beta}^{(k+1)}) \le g(\boldsymbol{\beta}^{(k+1)} \,|\, \boldsymbol{\beta}^{(k)})$, arguments in (9) hold. Starting from a prespecified value $\lambda = \lambda_0$, we successfully inflate λ by a factor $\rho > 1$. After the ℓth iteration, $\lambda = \lambda_\ell = \rho^{\ell-1}\lambda_0$. We take the first ℓ such that $f(\widehat{\boldsymbol{\beta}}_{\lambda_\ell}^{(k+1)}) \le g_{\lambda_\ell}(\widehat{\boldsymbol{\beta}}_{\lambda_\ell}^{(k+1)} \,|\, \widehat{\boldsymbol{\beta}}^{(k)})$ and set $\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}_{\lambda_\ell}^{(k+1)}$. Such an ℓ always exists as a large ℓ will major the function f. We then continue with the iteration in the MM part. A simple criteria for stopping the iteration is that $|f(\widehat{\boldsymbol{\beta}}^{(k+1)}) - f(\widehat{\boldsymbol{\beta}}^{(k)})| \le \epsilon$ for a sufficiently small ϵ, say 10^{-5}. We refer to Fan et al. (2015) for a detailed computational complexity analysis of the LAMM algorithm.

While the LAMM algorithm can be applied to compute $\widehat{\boldsymbol{\beta}}(s)$ in a general setting, in our application, the algorithm is mainly applied to compute GOSF under the null model (see Figure 1 and Section 3.5). From our simulation experiences, our algorithm delivers a good enough solution under the null model. It always provides an upper certificate $f(\widehat{\boldsymbol{\beta}}_0)$ to the problem $\min_{\|\boldsymbol{\beta}\|_0 \le s} f(\boldsymbol{\beta})$, where $\widehat{\boldsymbol{\beta}}_0$ is the output of the LAMM algorithm. As in Bertsimas, King and Mazumder (2016), if needed to verify the accuracy of our method, a lower certificate is $f(\widehat{\boldsymbol{\beta}}_1)$, where $\widehat{\boldsymbol{\beta}}_1$ is the solution to the convex problem $\min_{\|\boldsymbol{\beta}\|_1 \le B_s} f(\boldsymbol{\beta})$, and B_s is a sufficient large constant so that the L_0-solution satisfies $\|\widehat{\boldsymbol{\beta}}(s)\|_1 \le B_s$. For example, under the null model, it is well known that $\|\widehat{\boldsymbol{\beta}}(s)\|_1 = O_{\mathbb{P}}\{s\sqrt{(\log p)/n}\}$. Therefore, we can take $B_s = C_s s\sqrt{(\log p)/n}$ for a sufficiently large constant C_s. A data-driven heuristic approach is to take $B_s = 2\|\widehat{\boldsymbol{\beta}}_1(s)\|_1$ along the Lasso path such that $\|\widehat{\boldsymbol{\beta}}_1(s)\|_0 = s$.

Note that the minimum target value falls in the interval $[f(\widehat{\boldsymbol{\beta}}_1), f(\widehat{\boldsymbol{\beta}}_0)]$. If this interval is very tight, we have certified that $\widehat{\boldsymbol{\beta}}_0$ is an accurate solution.

3. Asymptotic distribution of goodness of spurious fit

3.1 Preliminaries

Define $p \times p$ covariance matrices

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}. \tag{11}$$

For $s \in [p]$, we say that $S \subseteq [p]$ is an s-subset if $|S| = s$. For every s-subset $S \subseteq [p]$, let $\boldsymbol{\Sigma}_{SS}$ and $\widehat{\boldsymbol{\Sigma}}_{SS}$ be the $s \times s$ sub-matrices of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ containing the entries indexed by $S \times S$, that is,

$$\boldsymbol{\Sigma}_{SS} = \mathbb{E}(\boldsymbol{X}_S\boldsymbol{X}_S^{\mathrm{T}}), \quad \widehat{\boldsymbol{\Sigma}}_{SS} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}. \tag{12}$$

Condition 3.1 The covariates are standardized to have unit second moment, that is, $\mathbb{E}(X_j^2) = 1$ for $j = 1, \ldots, p$. There exits a random vector $\boldsymbol{U} \in \mathbb{R}^p$ satisfying $\mathbb{E}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}) = \mathbf{I}_p$, such that $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}$ and $A_0 := \sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{U}\|_{\psi_2} < \infty$.

For $1 \le s \le p$, the s-sparse condition number of $\mathbf{\Sigma}$ is given by

$$\gamma_s = \gamma_s(\mathbf{\Sigma}) = \sqrt{\lambda_{\max}(s)/\lambda_{\min}(s)}, \tag{13}$$

where $\lambda_{\max}(s) = \max_{\boldsymbol{u}\in\mathbb{S}^{p-1}:\|\boldsymbol{u}\|_0\le s} \boldsymbol{u}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{u}$ and $\lambda_{\min}(s) = \min_{\boldsymbol{u}\in\mathbb{S}^{p-1}:\|\boldsymbol{u}\|_0\le s} \boldsymbol{u}^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{u}$ denote the s-sparse largest and smallest eigenvalues of $\mathbf{\Sigma}$, respectively.

Let $\boldsymbol{G} = (G_1, \ldots, G_p)^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Sigma})$ be a centered Gaussian random vector with covariance matrix $\mathbf{\Sigma}$. For any s-subset $S \subseteq [p]$, $\boldsymbol{G}_S \sim N(\mathbf{0}, \mathbf{\Sigma}_{SS})$. Define the random variable

$$R_0(s,p) = \max_{S\subseteq[p]:|S|=s} \|\mathbf{\Sigma}_{SS}^{-1/2}\boldsymbol{G}_S\|_2, \tag{14}$$

which is the maximum of the ℓ_2-norms of a sequence of dependent chi-squared random variables with s degrees of freedom. The distribution of $R_0(s,p)$ depends on the unknown $\mathbf{\Sigma}$ and can be estimated by the multiplier bootstrap in Section 3.5. It will be shown that this distribution is the asymptotic distribution of GOSF. In particular, for the isotropic case where $\mathbf{\Sigma} = \mathbf{I}_p$, $R_0(s,p) = G_{(1)}^2 + \cdots + G_{(s)}^2$, the sum of the largest s order statistics of p independent χ_1^2 random variables.

3.2 Generalized linear models

For i.i.d. observations $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from the distribution in (5), define individual residuals $\varepsilon_i = Y_i - \mathbb{E}_{\boldsymbol{X}}(Y_i) = Y_i - b'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$ with conditional variance $\mathrm{Var}_{\boldsymbol{X}}(\varepsilon_i) = \phi b''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$, where $\mathrm{Var}_{\boldsymbol{X}}(\cdot) = \mathbb{E}_{\boldsymbol{X}}\{\cdot - \mathbb{E}_{\boldsymbol{X}}(\cdot)\}^2$. In particular, under the null model, Y is independent of $\boldsymbol{X}$ with mean $\mu_Y := \mathbb{E}(Y) = b'(0)$ and variance $\sigma_Y^2 := \mathrm{Var}(Y) = \phi b''(0)$.

Condition 3.2 There exists $a_0 > 0$ such that $\mathbb{E}\exp\{u\sigma_Y^{-1}(Y-\mu_Y)\} \le \exp(a_0u^2/2)$ holds for all $u \in \mathbb{R}$. The function $b(\cdot)$ in (5) satisfies

$$\min_{u:|u|\le 1} b''(u) \ge a_1 \quad \text{and} \quad \max_{u:|u|\le 1} |b'''(u)| \le A_1 \tag{15}$$

for some constants $a_1, A_1 > 0$.

Condition 3.2 is satisfied by a wide class of GLIMs, including the logistic and Poisson regression models. The following theorem shows that, under certain moment and regularity conditions, the distribution of the generalized likelihood ratio statistic $2\mathcal{LR}_n(s,p)$ can be consistently approximated by that of $R_0^2(s,p)$ given in (14).

Theorem 2 *Let Conditions 3.1 and 3.2 be satisfied. Assume that $\phi = 1$ in (5), $p, n \ge 3$ and $1 \le s \le \min(p,n)$. Then, under the null model (7) with $\boldsymbol{\beta}^* = \mathbf{0}$,*

$$\sup_{t\ge 0} \big|\mathbb{P}\big\{2\mathcal{LR}_n(s,p) \le t\big\} - \mathbb{P}\{R_0^2(s,p) \le t\}\big| \le C\big[\{s\log(\gamma_s pn)\}^{7/8}n^{-1/8} + \gamma_s^{1/2}\{s\log(\gamma_s pn)\}^2 n^{-1/2}\big], \tag{16}$$

where $C > 0$ is a constant depending only on a_0, a_1, A_0, A_1 in Conditions 3.1 and 3.2.

Remark 3 We regard Theorem 2 as a nonasymptotic, high-dimensional version of the celebrated Wilks theorem. In the low-dimensional setting where $s = p$ is fixed, Theorem 2 reduces to the conventional Wilks theorem, which asserts that the generalized likelihood ratio statistic converges in distribution to χ_p^2. In addition, we also provide a Berry-Esseen bound in (16).

3.3 Linear least squares regression

As a specific case of GLIM, we consider the linear regression model (7) with the loss function $L_n(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$. The corresponding likelihood ratio statistic

$$\mathcal{LR}_n(s,p) = \frac{1}{2}\|\boldsymbol{Y}\|_2^2 - \min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} L_n(\boldsymbol{\beta}) \tag{17}$$

then coincides with that in (6) with $b(u) = \frac{1}{2}u^2$. We state the null limiting distribution of $\mathcal{LR}_n(s,p)$ in a general case, where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. copies of a sub-Gaussian random variable ε. Specifically, we assume that

Condition 3.3 ε is a centered, sub-Gaussian random variable with $\mathrm{Var}(\varepsilon) = \sigma^2 > 0$ and $K_0 := \|\varepsilon\|_{\psi_2} < \infty$. Moreover, write $v_\ell = \mathbb{E}(|\varepsilon|^\ell)$ for $\ell \ge 3$.

The following corollary is a particular case of the general result Theorem 2 with $b(u) = \frac{1}{2}u^2$, $u \in \mathbb{R}$ and $\phi = \sigma^2$. By examining the proof of Theorem 2 and noting that $b''' \equiv 0$, it can be easily shown that the second term on the right-side of (16) vanishes. Hence, the proof is omitted.

Corollary 4 *Let Conditions 3.1 and 3.3 hold. Assume that $p, n \ge 3$ and $1 \le s \le \min(p,n)$. Then, under the null model (7) with $\boldsymbol{\beta}^* = \boldsymbol{0}$,*

$$\sup_{t\ge 0}\left|\mathbb{P}\big\{2\mathcal{LR}_n(s,p) \le t\big\} - \mathbb{P}\big\{\sigma^2 R_0^2(s,p) \le t\big\}\right| \le C\{s\log(\gamma_s pn)\}^{7/8} n^{-1/8},$$

where $C > 0$ is a constant depending only on A_0 and K_0 in Conditions 3.1 and 3.3.

Remark 5 Under the null model, the variance σ^2 can be consistently estimated by $\widehat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$. Under the same conditions of Corollary 4, it can be proved that

$$\sup_{t\ge 0}\left|\mathbb{P}\big\{2\mathcal{LR}_n(s,p) \le t\big\} - \mathbb{P}\big\{\widehat{\sigma}_0^2 R_0^2(s,p) \le t\big\}\right| \lesssim \{s\log(\gamma_s pn)\}^{7/8} n^{-1/8},$$

which is in line with Theorem 3.1 in Fan, Shao and Zhou (2015). To see this, note that

$$\begin{aligned} 2\mathcal{LR}_n(s,p) &= \|\boldsymbol{Y}\|_2^2 - \min_{S\subseteq[p]:|S|=s}\min_{\boldsymbol{\theta}\in\mathbb{R}^s}\|\boldsymbol{Y} - \mathbb{X}_S\boldsymbol{\theta}\|_2^2 \\ &= \max_{S\subseteq[p]:|S|=s} \boldsymbol{Y}^{\mathrm{T}}\mathbb{X}_S(\mathbb{X}_S^{\mathrm{T}}\mathbb{X}_S)^{-1}\mathbb{X}_S^{\mathrm{T}}\boldsymbol{Y} = \max_{\boldsymbol{\alpha}\in\mathbb{R}^p:\|\boldsymbol{\alpha}\|_0\le s} (\boldsymbol{Y}^{\mathrm{T}}\mathbb{X}\boldsymbol{\alpha})^2/\|\mathbb{X}\boldsymbol{\alpha}\|_2^2. \end{aligned}$$

The estimator $\widehat{\sigma}_0^2$, used in computing the maximum spurious correlation, can be seriously biased beyond the null model and hence adversely affect the power. Thus, we suggest using either the refitted cross-validation procedure (Fan, Guo and Hao, 2012) or the scaled Lasso estimator (Sun and Zhang, 2012) to estimate σ^2.

3.4 Linear median regression

We now state an analogous result to Theorem 2 regarding the ℓ_1-loss considered in Section 2.2.

Condition 3.4 The noise $\varepsilon_1, \ldots, \varepsilon_n$ in (7) are i.i.d. copies of a random variable ε satisfying $\mathbb{E}|\varepsilon|^\kappa < \infty$ for some $1 < \kappa \leq 2$. There exist positive constants $a_2 < (\mathbb{E}|\varepsilon|)^{-1}$, A_2 and A_3 such that the distribution function $F_\varepsilon(\cdot)$ and the density function $f_\varepsilon(\cdot)$ of ε satisfy

$$2\max\{1 - F_\varepsilon(u), F_\varepsilon(-u)\} \leq (1 + a_2 u)^{-1} \quad \text{for all } u \geq 0, \tag{18}$$

$$\max_{u\in\mathbb{R}} f_\varepsilon(u) \leq A_2 \quad \text{and} \quad \max_{u:|u|\leq 1} \max\big\{|f_\varepsilon'(u+)|, |f_\varepsilon'(u-)|\big\} \leq A_3. \tag{19}$$

Theorem 6 *If $p, n \geq 3$ and $1 \leq s \leq \min(p, n)$, then under the null model (7) with $\boldsymbol{\beta}^* = \mathbf{0}$ and Conditions 3.1 and 3.4, we have*

$$\begin{aligned}\sup_{t\geq 0} \big|\mathbb{P}\big\{2\mathcal{LR}_n(s,p) \leq t\big\} - \mathbb{P}\big\{R_0^2(s,p)/\{2f_\varepsilon(0)\} \leq t\big\}\big| \\ \leq C_1\, n^{1-\kappa} + C_2\big[\{s\log(\gamma_s pn)\}^{7/8} n^{-1/8} + \gamma_s^{1/4}\{s\log(\gamma_s pn)\}^{3/2} n^{-1/4}\big],\end{aligned} \tag{20}$$

where $\mathcal{LR}_n(s,p)$ is given by (8), $C_1 > 0$ is a constant depending on a_2, κ, $\mathbb{E}|\varepsilon|$, $\mathbb{E}|\varepsilon|^\kappa$ and $C_2 > 0$ is a constant depending on a_2, A_0, A_2 and A_3 in Conditions 3.1 and 3.4.

Remark 7 Under the null model, the unknown parameter $f_\varepsilon(0)$ can be consistently estimated by the kernel density estimator $\widehat{f_\varepsilon}(0) = (nh)^{-1}\sum_{i=1}^n K(Y_i/h)$, where $K(\cdot)$ is a kernel function and $h = h_n > 0$ is the bandwidth. For simplicity, we may use the Epanechnikov kernel function $K_{\mathrm{Epa}}(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$ along with the rule-of-thumb bandwidth $h_{\mathrm{ROT}} = 2.34\,\widehat{\sigma}_0 n^{-1/5}$, where $\widehat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$.

3.5 Multiplier bootstrap procedure

The distribution of the random variable $R_0(s,p)$ given by (14) depends on the unknown covariance matrix $\boldsymbol{\Sigma}$. In practice, it is natural to replace $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ and $\boldsymbol{G} \sim N(0, \boldsymbol{\Sigma})$ by $\widehat{\boldsymbol{G}} \sim N(\mathbf{0}, \widehat{\boldsymbol{\Sigma}})$ in the definition of $R_0(s,p)$. With this substitution, the distribution of $R_0(s,p)$ can be simulated. In particular, $\widehat{\boldsymbol{G}}$ can be simulated as $n^{-1/2}\sum_{i=1}^n e_i \boldsymbol{X}_i$, where $e_1, \ldots, e_n$ are i.i.d. standard normal random variables that are independent of $\{\boldsymbol{X}_i\}_{i=1}^n$. The resulting estimator is

$$R_n(s,p) = \max_{S\subseteq[p]:|S|=s} \|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\widehat{\boldsymbol{G}}_S\|_2, \tag{21}$$

which is a multiplier bootstrap version of $R_0(s,p)$. The following proposition follows directly from Theorem 3.2 in Fan, Shao and Zhou (2015).

Proposition 8 *Assume that Condition (3.1) holds, $1 \leq s \leq \min(p,n)$ and $s\log(\gamma_s pn) = o(n^{1/5})$ as $n \to \infty$. Then $\sup_{t\geq 0}|\mathbb{P}\{R_0(s,p) \leq t\} - \mathbb{P}\{R_n(s,p) \leq t\,|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}| \to 0$ in probability.*

The computation of $R_n(s,p)$ requires solving a combinatorial optimization. This can be alleviated by using the LAMM algorithm in Section 2.3. To begin with, by Remark 5, we write $R_n(s,p)$ in (21) as

$$R_n^2(s,p) = \max_{S\subseteq[p]:|S|=s} \mathbf{e}^{\mathrm{T}}\mathbb{X}_S(\mathbb{X}_S^{\mathrm{T}}\mathbb{X}_S)^{-1}\mathbb{X}_S^{\mathrm{T}}\mathbf{e} = \|\boldsymbol{e}\|_2^2 - \min_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} \|\boldsymbol{e}-\mathbb{X}\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{e} = (e_1,\ldots,e_n)^{\mathrm{T}}$ and $\mathbb{X}_S = (\boldsymbol{X}_{1S},\ldots,\boldsymbol{X}_{nS})^{\mathrm{T}}$ for every subset $S\subseteq[p]$. This can be computed approximately by the LAMM algorithm in Section 2.3, resulting in the solution $\widehat{\boldsymbol{\beta}}(s)$. Finally, we set $R_n^2(s,p) = \|\boldsymbol{e}\|_2^2 - \|\boldsymbol{e}-\mathbb{X}\widehat{\boldsymbol{\beta}}(s)\|_2^2$.

The numerical performance may be improved by employing mixed integer optimization formulations (Bertsimas, King and Mazumder, 2016). Such an attempt, however, is beyond the scope of the paper and we leave it for future research.

4. Spurious discoveries and model selection

Based on the theoretical developments in Section 3, here we address the question whether discoveries by machine learning and data mining techniques for GLIM are any better than by chance. For simplicity, we focus on the Lasso. Let $q_\alpha(s,p)$ be the upper α-quantile of the random variable $R_0(s,p)$ defined by (14). Assume that the dispersion parameter ϕ in (5) equals 1. By Theorem 2, we see that for any prespecified $\alpha\in(0,1)$,

$$\mathbb{P}\big\{2\mathcal{LR}_n(s,p) \le q_\alpha^2(s,p)\big\} \to 1-\alpha, \tag{22}$$

where $\mathcal{LR}_n(s,p)$ is as in (6).

Let $\widehat{\boldsymbol{\beta}}_\lambda = \mathrm{argmin}_{\boldsymbol{\beta}}\{L_n(\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1\}$ be the ℓ_1-penalized maximum likelihood estimator with $\widehat{s}_\lambda = |\widehat{S}_\lambda| = |\mathrm{supp}(\widehat{\boldsymbol{\beta}}_\lambda)|$, where $\lambda>0$ is the regularization parameter. The goodness of fit is likelihood ratio $L_n(\mathbf{0})-L_n(\widehat{\boldsymbol{\beta}}_\lambda)$. Since $\widehat{s}_\lambda$ covariates are selected, it should be compared with the distribution of GOSF $\mathcal{LR}_n(s,p)$ by taking $s=\widehat{s}_\lambda$. In view of (22), if

$$L_n(\widehat{\boldsymbol{\beta}}_\lambda) \ge L_n(\mathbf{0}) - q_\alpha^2(\widehat{s}_\lambda,p)/2 = nb(0) - q_\alpha^2(\widehat{s}_\lambda,p)/2,$$

then we may regard the discovery of variables $\widehat{S}_\lambda$ as unimpressive, no better than fitting by chance, or simply spurious.

In practice, the unknown quantile $q_\alpha(s,p)$ should be replaced by its bootstrap version $q_{n,\alpha}(s,p)$, the upper α-quantile of $R_n(s,p)$ defined by (21). This leads to the following data-driven criteria for judging where the discovery $\widehat{S}(\lambda)$ is spurious:

$$L_n(\widehat{\boldsymbol{\beta}}_\lambda) \ge nb(0) - q_{n,\alpha}^2(\widehat{s}_\lambda,p)/2. \tag{23}$$

The theoretical justification is given by Theorem 2 and Proposition 8. In particular, when the loss is quadratic, this reduces to the case studied by Fan, Shao and Zhou (2015).

The concept of GOSF and its theoretical quantile provide important guidelines for model selection. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{cv}}$ be a cross-validated Lasso estimator, which selects $\widehat{s}_{\mathrm{cv}} = \|\widehat{\boldsymbol{\beta}}_{\mathrm{cv}}\|_0$ important variables. Due to the bias of the ℓ_1 penalty, the Lasso typically selects far larger model size since the visible bias in Lasso forces the cross-validation procedure to choose a smaller value of λ. This phenomenon is documented in the simulations studies. See Table 1 in

Section 5.2. With an over-selected model, both the goodness of fit $\widehat{\mathcal{LR}}_\lambda = L_n(\mathbf{0}) - L_n(\widehat{\boldsymbol{\beta}}_\lambda)$ and the spurious fit can be very large, and so is the finite sample Wilks approximation error. To avoid over-selecting, we suggest an alternative procedure that uses the quantity $q_{n,\alpha}(s,p)$ as a guidance to choose the tuning parameter, which guards us from spurious discoveries. More specifically, for each λ in the Lasso solution path, we compute $\widehat{\mathcal{LR}}_\lambda$ and $q_{n,\alpha}(s,p)|_{s=\widehat{s}_\lambda}$ with a prespecified α. Starting from the largest λ, we stop the Lasso path the first time that the sign of $2\widehat{\mathcal{LR}}_\lambda - q^2_{n,\alpha}(\widehat{s}_\lambda, p)$ is changed from positive to negative, and let $\widehat{\lambda}_{\text{fit}}$ be the smallest λ satisfying $2\widehat{\mathcal{LR}}_\lambda \geq q^2_{n,\alpha}(\widehat{s}_\lambda, p)$. Denote by $\widehat{s}_{\text{fit}}$ the corresponding selected model size. This value can be regarded as the maximum model size for Lasso (or any other variable selection technique such as SCAD) to choose from. Another viable alternative is to only select the best cross-validated model among those whose fit are better than GOSF. We will show in Section 5.2 by simulation studies that this procedure selects much smaller model size which is closer to the truth.

5. Numerical studies

5.1 Accuracy of the Gaussian approximation

First we ran a simulation study to examine how accurate the Gaussian approximation $R^2_0(s,p)$ is to the generalized likelihood ratio statistic $2\mathcal{LR}_n(s,p)$ in the null model. To illustrate the method, we focus on the logistic regression model: $\mathbb{P}(Y=1|\boldsymbol{X}) = \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)/\{1+\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\}$. Under the null model $\boldsymbol{\beta}^* = 0$, $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli random variables with success probability $1/2$. Independent of Y_i's, we generate $\boldsymbol{X}_i \sim N(0, \boldsymbol{\Sigma})$ with two different covariance matrices: $\boldsymbol{\Sigma}_1 = (\rho^{|j-k|})_{1\leq j,k\leq p}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{2,jk})_{1\leq j,k\leq p}$, where

$$\sigma_{2,jk} = \big(\big||j-k|+1\big|^{2\rho} + \big||j-k|-1\big|^{2\rho} - 2|i-j|^{2\rho}\big)/2, \quad 1\leq j,k\leq p.$$

The first design has an AR(1) correlation structure (a short-memory process), whereas the second design reflects strong long memory dependence. We take $\rho = 0.8$ in both cases.

Figure 2 reports the distributions of generalized likelihood ratios (GLRs) and their Gaussian approximations (GARs) when $n = 400$, $p = 1000$ and $s \in \{1, 2, 5, 10\}$. The results show that the accuracy of Gaussian approximation is fairly reasonable and is affected by the size of s as well as the dependence between the coordinates of $\boldsymbol{X}$.

5.2 Detection of spurious discoveries

In this section, we conduct a moderate scale simulation study to examine how effective the multiplier bootstrap quantile $q_{n,\alpha}(s,p)$ serves as a benchmark for judging whether the discovery is spurious. To illustrate the main idea, again we restrict our attention to the logistic regression model and the Lasso procedure.

The results reported here are based on 200 simulations with the ambient dimension $p = 400$ and the sample size n taken values in $\{120, 160, 200\}$. The true regression coefficient vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is $(3, -1, 3, -1, 3, 0, \ldots, 0)^{\mathrm{T}}$. We consider two random designs: $\boldsymbol{\Sigma} = \mathbf{I}_p$ (independent) and $\boldsymbol{\Sigma} = (0.5^{|j-k|})_{1\leq j,k\leq p}$ (dependent).

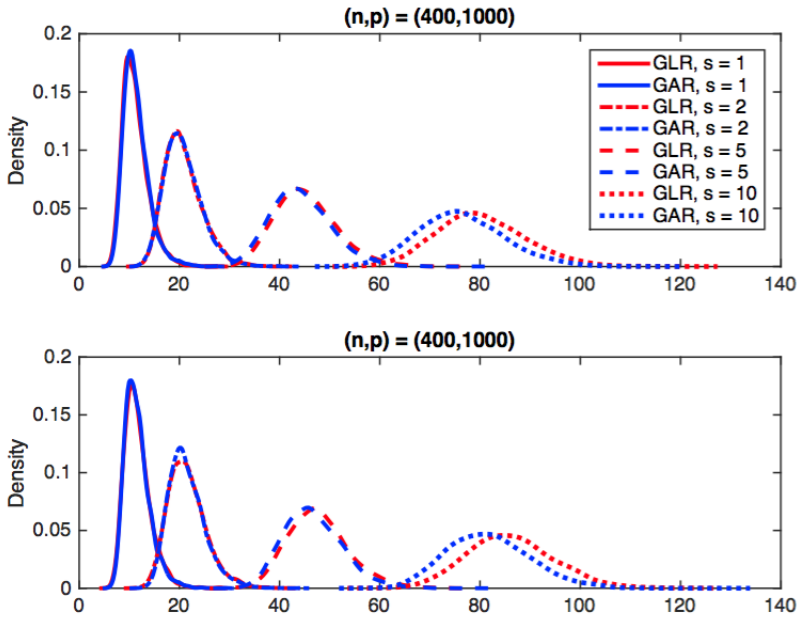

Figure 2: Distributions of generalized likelihood ratios (red) and Gaussian approximations (blue) based on 5000 simulations for $n = 400$, $p = 1000$ and $s = 1, 2, 5, 10$ when $\mathbf{\Sigma}$ is equal to $\mathbf{\Sigma}_1$ (upper panel) or $\mathbf{\Sigma}_2$ (lower panel).

Let $\widehat{\boldsymbol{\beta}}_{\rm cv}$ be the five-fold cross-validated Lasso estimator, which selects a model of size $\widehat{s}_{\rm cv} = \|\widehat{\boldsymbol{\beta}}_{\rm cv}\|_0$. For a given $\alpha \in (0, 1)$, consider the spurious discovery probability (SDP)

$$\mathbb{P}\big\{n \log(2) - L_n(\widehat{\boldsymbol{\beta}}_{\rm cv}) \le q^2_{n,\alpha}(\widehat{s}_{\rm cv}, p)/2\big\},$$

which is basically the probability of the type II error since the simulated model is not null. We take $\alpha = 0.1$ and compute the empirical SDP based on 200 simulations. For each simulated data set, $q_{n,\alpha}(s, p)|_{s=\widehat{s}_{\rm cv},\, p=400}$ is computed based on 1000 bootstrap replications. The results are depicted in Table 1 below.

Table 1: The empirical power and the median size of the selected models with its robust standard deviation (RSD) in the parenthesis based on 200 simulations when $p = 400$ and $\alpha = 10\%$. RSD is the interquantile range divided by 1.34.

	$n = 120$		$n = 160$		$n = 200$	
	Ind.	Dep.	Ind.	Dep.	Ind.	Dep.
Power	0.595	0.750	0.925	0.980	1.000	1.000
$\widehat{s}_{\rm cv}$	32.0 (13.43)	24.5 (11.94)	40.0 (13.81)	25.5 (12.69)	42.0 (14.18)	29.0 (14.18)

As reflected by Table 1, the empirical power, which is one minus the empirical SDP, increases rapidly as the sample size n grows. This is in line with our intuition that the more data we have, the less likely that the discovery by a variable selection method is spurious. When the sample size is small, the SDP can be high and hence the discovery

$\widehat{S}_{\rm cv} = {\rm supp}(\widehat{\boldsymbol{\beta}}_{\rm cv})$ should be interpreted with caution. We need either more samples or more powerful variable selection methods.

We see from Table 1 that the Lasso with cross-validation selects far larger model size than the true one, which is 5. This is because the intrinsic bias in Lasso forces the cross-validation procedure to choose a smaller value of λ. We now use our procedure in Section 4 to choose the tuning parameter from the Lasso solution path. As before, we take $\alpha = 0.1$ in $q_{n,\alpha}(s,p)$ to provide an upper bound on the model size from perspective of guarding against spurious discoveries. The empirical median of $\widehat{s}_{\rm fit}$ and its robust standard deviation are 9 and 1.87 over 200 simulations when $(n,p) = (200, 400)$ and $\boldsymbol{\Sigma} = (0.5^{|j-k|})_{1\le j,k\le p}$. The feature over-selection phenomenon is considerably alleviated.

5.3 Neuroblastoma data

In this section, we apply the idea of detecting spurious discoveries to the neuroblastoma data reported in Oberthuer et al. (2006). This data set consists of 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. The complete data set, obtained via the MicroArray Quality Control phase-II (MAQC-II) project (Shi et al., 2010), includes gene expression over 10,707 probe sites. There are 246 subjects with 3-year event-free survival information available (56 positive and 190 negative). See Oberthuer et al. (2006) for more details about the data sets.

For each $\lambda > 0$, we apply Lasso using the logistic regression model to select $\widehat{s}_\lambda$ genes. In particular, ten-fold cross-validated Lasso selects $\widehat{s}_{\rm cv} = 40$ genes. Then we calculate the goodness of fit $\widehat{\mathcal{LR}}_\lambda := L_n(\mathbf{0}) - L_n(\widehat{\boldsymbol{\beta}}_\lambda) = n\log(2) - L_n(\widehat{\boldsymbol{\beta}}_\lambda)$. Along the Lasso path, we record in Table 2 the number of selected probes, the corresponding square-root the goodness of fit $(2\widehat{\mathcal{LR}}_\lambda)^{1/2}$ and upper α-quantiles of the multiplier bootstrap approximations $R_0(s,p)|_{s=\widehat{s}_\lambda,\, p=10{,}707}$ with $\alpha = 10\%$ and 5% based on 2000 bootstrap replications. For illustrative purposes, we only display partial Lasso solutions with selected model size $\widehat{s}_\lambda$ lying between 20 and 40. From Table 2, we observe that only the discovery of 17 probes has a generalized measure of the goodness of fit better than GOSF at $\alpha = 5\%$, whereas the finding (of the 40 probes) via the cross-validation procedure is likely to over-select.

6. Proofs

We now turn to the proofs of Theorems 2 and 6. In each proof, we provide the primary steps, with more technical details stated as lemmas and proved in the appendix.

6.1 Proof of Theorem 2

Throughout, we work with the quasi-likelihood $\mathcal{L}_n(\boldsymbol{\beta}) = -L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \{Y_i \boldsymbol{X}_i^{\rm T}\boldsymbol{\beta} - b(\boldsymbol{X}_i^{\rm T}\boldsymbol{\beta})\}$ and consider the general case where the dispersion parameter ϕ in (5) is specified (not necessarily equals 1 to facilitate the derivations for the normal case). For a given $s \in [p]$, define

$$Q_n(s,p) = \max_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} \mathcal{L}_n(\boldsymbol{\beta}) \quad \text{and} \quad Q_n^* = \mathcal{L}_n(\mathbf{0}).$$

We divide the proof into three steps. First, for each s-subset $S \subseteq [p]$, we prove Wilks's result for the S-restricted model where only a subset of the covariates indexed by S are

Table 2: Lasso fitted square-root likelihood ratio statistic, the mean cross-validated error, and upper 0.1- and 0.05-quantiles of the multiplier bootstrap approximation based on 2000 bootstrap samples.

λ	$\widehat{s}_\lambda$	$(2\widehat{\mathcal{LR}}_\lambda)^{1/2}$	$q_{n,0.1}(\widehat{s}_\lambda, p)$	$q_{n,0.05}(\widehat{s}_\lambda, p)$	Mean Cross-Validated Error
0.2117	3	9.1389	6.4898	6.6519	1.0641
0.1929	4	9.4753	7.2464	7.4353	1.0450
0.1841	6	9.7273	8.4241	8.6061	1.0346
0.1678	7	10.1670	8.8959	9.0750	1.0092
0.1601	8	10.3675	9.3121	9.5102	0.9974
0.1459	9	10.7263	9.7115	9.9097	0.9751
0.1329	11	11.0739	10.3954	10.6071	0.9543
0.1269	12	11.2376	10.7042	10.9207	0.9452
0.1211	13	11.4330	10.9875	11.2085	0.9359
0.1104	14	11.7764	11.2576	11.4849	0.9186
0.1006	15	12.0756	11.5084	11.7407	0.9006
0.0960	17	**12.2096**	**11.9664**	**12.2000**	0.8934
0.0875	20	**12.4788**	**12.5543**	**12.7891**	0.8815
0.0761	25	12.9535	13.3824	13.6022	0.8651
0.0575	31	13.8675	14.1407	14.3703	0.8361
0.0456	40	14.5588	14.9712	15.2099	0.8255

included. Specifically, we show that the square root deviation of the S-restricted maximum log-likelihood from its baseline value under the null model can be well approximated by the ℓ_2-norm of the normalized score vector. Second, based on a high-dimensional invariance principle, we prove the Gaussian/chi-squared approximation for the maximum of the ℓ_2-norms of normalized score vectors. Finally, we apply an anti-concentration argument to construct non-asymptotic Wilks approximation for $2\{Q_n(s,p) - Q_n^*\}$.

Step 1: Wilks approximation. In the null model where Y and $\boldsymbol{X}$ are independent, the true parameter $\boldsymbol{\beta}^*$ in (5) is zero, and thus the density function of Y has the form $f(y) = \exp\{-\phi^{-1}b(0) + c(y,\phi)\}$. Moreover, we have

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n(\boldsymbol{\beta})\} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{X}}\{Y_i\, \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}.$$

To this see, note that in model (5) with $\boldsymbol{\beta}^* = \mathbf{0}$, $\mathbb{E}(Y) = b'(0)$ and $\mathrm{Var}(Y) = \phi b''(0)$. This implies that $\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n(\boldsymbol{\beta})\} = \sum_{i=1}^n\{b'(0)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}$. This function is strictly concave with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\beta} = \mathbf{0}$ satisfies its first order condition, and hence is its maximizer.

For each s-subset $S \subseteq [p]$, define the S-restricted log-likelihood $\mathcal{L}_n^S(\boldsymbol{\theta}) = \sum_{i=1}^n\{Y_i\, \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} - b(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\}$ and the score function $\nabla\mathcal{L}_n^S(\boldsymbol{\theta}) = \sum_{i=1}^n\{Y_i - b'(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\}\boldsymbol{X}_{iS}$, $\boldsymbol{\theta} \in \mathbb{R}^s$. In this

notation, it can be seen from (6) that

$$Q_n(s,p) = \max_{S\subseteq[p]:|S|=s} \max_{\boldsymbol{\theta}\in\mathbb{R}^s} \mathcal{L}_n^S(\boldsymbol{\theta}) = \max_{S\subseteq[p]:|S|=s} \mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S), \tag{24}$$

where

$$\widehat{\boldsymbol{\theta}}_S = (\widehat{\theta}_{S,1},\ldots,\widehat{\theta}_{S,s})^{\mathrm{T}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^s} \mathcal{L}_n^S(\boldsymbol{\theta}) \tag{25}$$

denotes the maximum likelihood estimate of the target parameter for the S-restricted model, which is given by $\boldsymbol{\theta}_S^* := \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^s} \mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = \mathbf{0}$.

Given the i.i.d. observations $\{(Y_i,\boldsymbol{X}_i)\}_{i=1}^n$, $\nabla\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = \sum_{i=1}^n\{b'(0)-b'(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\}\boldsymbol{X}_{iS}$ and $\mathbf{H}_S(\boldsymbol{\theta}) := -\nabla^2\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = \sum_{i=1}^n b''(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}$ for $\boldsymbol{\theta}\in\mathbb{R}^s$. In particular, write

$$\mathbf{H}_S^* := \mathbf{H}_S(\mathbf{0}) = nb''(0)\,\widehat{\boldsymbol{\Sigma}}_{SS} \tag{26}$$

for $\boldsymbol{\Sigma}_{SS}$ as in (12). Further, define the S-restricted normalized score

$$\widehat{\boldsymbol{\xi}}_S = \mathbf{H}_S^{*-1/2}\nabla\mathcal{L}_n^S(\mathbf{0}) = \{nb''(0)\}^{-1/2}\,\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\sum_{i=1}^n \varepsilon_i\boldsymbol{X}_{iS}, \qquad \varepsilon_i = Y_i - b'(0). \tag{27}$$

The following result is a conditional analogue of Corollary 1.12 in the supplement of Spokoiny (2012), which provides an exponential inequality for the ℓ_2-norm of $\widehat{\boldsymbol{\xi}}_S$ given $\{\boldsymbol{X}_i\}_{i=1}^n$. The proofs of this Lemma and other lemmas can be found in the appendix.

Lemma 9 *Assume that Conditions 3.1 and 3.2 hold. Then, for every $t\geq 0$,*

$$\mathbb{P}_{\boldsymbol{X}}\big\{\|\widehat{\boldsymbol{\xi}}_S\|_2^2 \geq a_0\phi\Delta(s,t)\big\} \leq 2e^{-t} \tag{28}$$

holds almost surely on the event $\{\widehat{\boldsymbol{\Sigma}}_{SS}\succ\mathbf{0}\}$, where

$$\Delta(s,t) := \begin{cases} s+(8ts)^{1/2}, & \text{if } 0\leq t\leq \frac{1}{18}(2s)^{1/2},\\ s+6t, & \text{if } t>\frac{1}{18}(2s)^{1/2}. \end{cases} \tag{29}$$

The following lemma characterizes the Wilks phenomenon from a non-asymptotic perspective. Recall that $\widehat{\boldsymbol{\theta}}_S$ at (25) is the S-restricted maximum likelihood estimator, and in the null model, $\mathcal{L}_n^S(\mathbf{0}) = \mathcal{L}_n(\mathbf{0}) = -nb(0)$, $\sigma_Y^2 = \mathrm{Var}(Y) = \phi b''(0)$. For every $\tau>0$, define the event

$$\mathcal{E}_0(\tau) = \bigcap_{S\subseteq[p]:|S|=s}\Big\{\widehat{\boldsymbol{\Sigma}}_{SS}\succ\mathbf{0},\ \max_{1\leq i\leq n}\boldsymbol{X}_{iS}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\boldsymbol{X}_{iS}\leq\tau\Big\}. \tag{30}$$

Lemma 10 *Assume that Conditions 3.1 and 3.2 hold. Then, on the event $\mathcal{E}_0(\tau)$, for any $\tau>0$,*

$$\mathbb{P}_{\boldsymbol{X}}\bigg(\max_{S\subseteq[p]:|S|=s}\Big|[2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S)-\mathcal{L}_n(\mathbf{0})\}]^{1/2}-\|\widehat{\boldsymbol{\xi}}_S\|_2\Big| \leq C_1\,\phi\tau^{1/2}\frac{s\log(pn)}{\sqrt{n}}\bigg)\leq 5n^{-1} \tag{31}$$

whenever $n\geq C_2\,\phi\tau s\log(pn)$, where C_1 and C_2 are positive constants depending only on a_0,a_1,A_1 and $b''(0)$.

To apply Lemma 10, we need to show first that for properly chosen τ, the event $\mathcal{E}_0(\tau)$ occurs with high probability. First, applying Theorem 5.39 in Vershynin (2012) to the random vectors $\boldsymbol{\Sigma}_{SS}^{-1/2}\boldsymbol{X}_{1S},\ldots,\boldsymbol{\Sigma}_{SS}^{-1/2}\boldsymbol{X}_{nS}$ yields that, for every $t\geq 0$,

$$\big\|\boldsymbol{\Sigma}_{SS}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{SS}\boldsymbol{\Sigma}_{SS}^{-1/2}-\mathbf{I}_s\big\| = \big\|n^{-1}\boldsymbol{\Sigma}_{SS}^{-1/2}\mathbb{X}_S^{\mathrm{T}}\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-1/2}-\mathbf{I}_s\big\| \leq \max(\delta,\delta^2) \tag{32}$$

holds with probability at least $1-2e^{-t}$, where $\delta = C_3(s\vee t)^{1/2}n^{-1/2}$, and $C_3>0$ is a constant depending only on A_0. This, together with Boole's inequality implies by taking $t = s\log\frac{ep}{s}+\log n$ that, with probability at least $1-2n^{-1}$,

$$\max_{S\subseteq[p]:|S|=s}\big\|\boldsymbol{\Sigma}_{SS}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{SS}\boldsymbol{\Sigma}_{SS}^{-1/2}-\mathbf{I}_s\big\| \leq C_3\bigg(\frac{s\log\frac{ep}{s}+\log n}{n}\bigg)^{1/2}\leq\frac{1}{2} \tag{33}$$

whenever $n\geq 4C_3^2(s\log\frac{ep}{s}+\log n)$. Providing (33) holds, the smallest eigenvalue of $\boldsymbol{\Sigma}_{SS}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{SS}\boldsymbol{\Sigma}_{SS}^{-1/2}$ is bounded from below by $\frac{1}{2}$ so that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{SS})\geq\frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{SS})$. Moreover,

$$\boldsymbol{X}_{iS}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\boldsymbol{X}_{iS}\leq 2\lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{SS})\|\boldsymbol{X}_{iS}\|_2^2\leq 2s\lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{SS})\max_{j\in S}X_{ij}^2. \tag{34}$$

For the last term on the right-hand side of (34), let $\boldsymbol{e}_j=(0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$ be the unit vector in $\mathbb{R}^p$ with 1 at the jth position and note that $X_{ij}=\boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{X}_i=\boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}_{SS}^{1/2}\boldsymbol{U}_i$ with $\|\boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\|_2=1$, where $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ are i.i.d. p-dimensional random vectors with covariance matrix $\mathbf{I}_p$. By Condition 3.1, $\|X_{ij}\|_{\psi_2}=\|\boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_i\|_{\psi_2}\leq A_0$ and hence for every $t\geq 0$,

$$\mathbb{P}\bigg(\max_{1\leq i\leq n}\max_{1\leq j\leq p}X_{ij}^2\geq t\bigg)\leq 2\sum_{i=1}^n\sum_{j=1}^p\exp(-C_4^{-1}t)\leq 2\exp\{\log(pn)-C_4^{-1}t\},$$

where $C_4>0$ is a constant depending only on A_0. This, together with (34) implies by taking $t=2C_4\log(pn)$ that, with probability at least $1-3n^{-1}$,

$$\max_{1\leq i\leq n}\max_{S\subseteq[p]:|S|=s}\boldsymbol{X}_{iS}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\boldsymbol{X}_{iS}\leq 2\lambda_{\min}^{-1}(s)\{1+2C_4\,s\log(pn)\}. \tag{35}$$

Now, by (30) and (35), we take $\tau_0=2\lambda_{\min}^{-1}(s)\{1+2C_4\,s\log(pn)\}$ such that the event $\mathcal{E}_0(\tau_0)$ occurs with probability greater than $1-3n^{-1}$ as long as $n\geq 4C_3^2(s\log\frac{ep}{s}+\log n)$. This, together with Lemma 10 yields that with probability at least $1-8n^{-1}$,

$$\max_{S\subseteq[p]:|S|=s}\Big|[2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S)-\mathcal{L}_n(\mathbf{0})\}]^{1/2}-\|\widehat{\boldsymbol{\xi}}_S\|_2\Big|\leq C_5\,\phi\lambda_{\min}^{-1/2}(s)\{s\log(pn)\}^{3/2}n^{-1/2} \tag{36}$$

whenever $n\geq C_6(1\vee\phi)\lambda_{\min}^{-1}(s)\{s\log(pn)\}^2$, where $C_5,C_6>0$ are constants depending only on a_0,a_1,A_0,A_1 and $b''(0)$.

Step 2: Gaussian approximation. For any $i=1,\ldots,n$ and $S\subseteq[p]$, define $\boldsymbol{Z}_i=\{b''(0)\}^{-1/2}\varepsilon_i\boldsymbol{X}_i$ and $\boldsymbol{Z}_{iS}=\{b''(0)\}^{-1/2}\varepsilon_i\boldsymbol{X}_{iS}$ such that $\widehat{\boldsymbol{\xi}}_S=n^{-1/2}\sum_{i=1}^n\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\boldsymbol{Z}_{iS}$. Moreover, define

$$\boldsymbol{\xi}=n^{-1/2}\sum_{i=1}^n\boldsymbol{Z}_i\quad\text{and}\quad\boldsymbol{\xi}_S=n^{-1/2}\sum_{i=1}^n\boldsymbol{\Sigma}_{SS}^{-1/2}\boldsymbol{Z}_{iS}. \tag{37}$$

The following result shows that for each s-subset $S \subseteq [p]$, the ℓ_2-norm of the S-restricted normalized score $\widehat{\boldsymbol{\xi}}_S$ is close to that of $\boldsymbol{\xi}_S$ with overwhelmingly high probability.

Lemma 11 *Assume that Condition 3.1 holds. Then, for every s-subset $S \subseteq [p]$ and for every $0 \le t \le \frac{3}{4}(n-2s)$,*

$$\mathbb{P}\Big[\Big|\|\widehat{\boldsymbol{\xi}}_S\|_2 - \|\boldsymbol{\xi}_S\|_2\Big| > C_7\{(s+t)\phi\Delta(s,t)\}^{1/2} n^{-1/2}\Big] \le 12.4\, e^{-t}, \tag{38}$$

provided that $n \ge C_8(s+t)$, where $\Delta(s,t)$ is as in (29) and $C_7, C_8 > 0$ are constants depending only on a_0 and A_0.

Using the union bound and taking $t = s\log\frac{ep}{s} + \log n$ in Lemma 11, we see that with probability at least $1 - 12.4\, n^{-1}$,

$$\max_{S\subseteq[p]:|S|=s} \big|\|\widehat{\boldsymbol{\xi}}_S\|_2 - \|\boldsymbol{\xi}_S\|_2\big| \le C_7\, \phi^{1/2}(s\log\tfrac{ep}{s} + \log n)\, n^{-1/2} \tag{39}$$

whenever $n \ge C_9(s\log\frac{ep}{s} + \log n)$.

Note that, the random vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_S, S \subseteq [p]$ defined in (37) satisfy $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}) = \phi\boldsymbol{\Sigma}$, $\mathbb{E}(\boldsymbol{\xi}_S) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{\xi}_S\boldsymbol{\xi}_S^{\mathrm{T}}) = \phi\mathbf{I}_s$. The following lemma provides a coupling inequality, showing that the random variable $\max_{S\subseteq[p]:|S|=s}\|\phi^{-1/2}\boldsymbol{\xi}_S\|_2$ can be well approximated, with high probability, by some random variable which is distributed as the maximum of the ℓ_2-norms of a sequence of normalized Gaussian random vectors, that is, $\{\|\boldsymbol{\Sigma}_{SS}^{-1/2}\boldsymbol{G}_S\|_2 : S \subseteq [p], |S| = s\}$.

Lemma 12 *Assume that Condition 3.1 holds. Then, there exists a random variable $T_0 \stackrel{d}{=} R_0(s,p)$ such that for any $\delta \in (0,1]$,*

$$\Big|\max_{S\subseteq[p]:|S|=s}\|\phi^{-1/2}\boldsymbol{\xi}_S\|_2 - T_0\Big| \le C_{10}\big[\delta + \{s\log(\gamma_s pn)\}^{1/2}n^{-1/2} + \{s\log(\gamma_s pn)\}^2 n^{-3/2}\big] \tag{40}$$

holds with probability greater than $1 - C_{11}\big[\delta^{-3}n^{-1/2}\{s\log(\gamma_s pn)\}^2 \vee \delta^{-4}n^{-1}\{s\log(\gamma_s pn)\}^5\big]$, where $C_{10}, C_{11} > 0$ are constants depending only on a_0 and A_0

Step 3: Completion of the proof. We now apply an anti-concentration argument to construct the Berry-Esseen bound for the square root of the excess $2\phi^{-1}\{Q_n(s,p) - Q_n^*\}$. To this end, taking $\delta = \{s\log(\gamma_s pn)\}^{3/8}n^{-1/8}$ in Lemma 12 leads to that, with probability at least $1 - C_{11}\{s\log(\gamma_s pn)\}^{7/8}n^{-1/8}$,

$$\Big|\max_{S\subseteq[p]:|S|=s}\|\phi^{-1/2}\boldsymbol{\xi}_S\|_2 - T_0\Big| \le C_{12}\{s\log(\gamma_s pn)\}^{3/8}n^{-1/8} \tag{41}$$

whenever $n \ge \{s\log(\gamma_s pn)\}^3$. Further, for $R_0(s,p)$ in (14), note that

$$R_0^2(s,p) = \max_{S\subseteq[p]:|S|=s}\max_{\boldsymbol{u}\in\mathbb{S}^{s-1}}\frac{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{G}_S)^2}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{SS}\boldsymbol{u}} = \max_{\boldsymbol{u}\in\mathcal{F}(s,p)}\frac{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{G})^2}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u}},$$

where $\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\mathcal{F}(s,p) := \{\boldsymbol{x} \mapsto \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} : \boldsymbol{u} \in \mathbb{S}^{p-1}, \|\boldsymbol{u}\|_0 \le s\}$ is a class of linear functions $\mathbb{R}^p \mapsto \mathbb{R}$. Hence, it follows from Lemma 7.3 in Fan, Shao and Zhou (2015) with slight modification and Lemma A.1 in the supplement of Chernozhukov, Chetverikov and Kato (2014) that, for every $t > 0$,

$$\sup_{u\ge 0} \mathbb{P}(|T_0 - u| \le t) = \sup_{u\ge 0} \mathbb{P}\big\{|R_0(s,p) - u| \le t\big\} \le C_{13}\big(s \log \tfrac{\gamma_s e p}{s}\big)^{1/2} t, \tag{42}$$

where $C_{13} > 0$ is an absolute constant. Combining (42) with the preceding results (36), (39) and (41) proves (16). ■

6.2 Proof of Theorem 6

The main strategy of the proof is similar to that of Theorem 2 but technical details are substantially different. As before, we define the quasi-likelihood $\mathcal{L}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}|$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and observe that $\max_{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\le s} \mathcal{L}_n(\boldsymbol{\beta}) = \max_{S\subseteq[p]:|S|=s} \max_{\boldsymbol{\theta}\in\mathbb{R}^s} \mathcal{L}_n^S(\boldsymbol{\theta})$, where $\mathcal{L}_n^S(\boldsymbol{\theta}) = -\sum_{i=1}^n |Y_i - \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}|$. In the null model (7) with $\boldsymbol{\beta}^* = \boldsymbol{0}$, we have for each s-subset $S \subseteq [p]$, $\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = \boldsymbol{0}$ by the first order condition and concavity, and the S-restricted least absolute deviation estimator can be written as

$$\widehat{\boldsymbol{\theta}}_S = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^s} \mathcal{L}_n^S(\boldsymbol{\theta}). \tag{43}$$

We first establish in Lemma 13 an upper bound for the maximum ℓ_2-risks of $\widehat{\boldsymbol{\theta}}_S$.

Lemma 13 *Assume that* (18) *holds and that* $\mathbb{E}|\varepsilon|^\kappa < \infty$ *for some* $1 < \kappa \le 2$. *Then, on the event* $\mathcal{E}_0(\tau)$ *for* $\tau > 0$, *the sequence of LAD estimators* $\{\widehat{\boldsymbol{\theta}}_S : S \subseteq [p], |S| = s\}$ *satisfies*

$$\max_{S\subseteq[p]:|S|=s} \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2}\widehat{\boldsymbol{\theta}}_S\|_2 \le C_1\, a_2^{-1}\{s\log(pn)\}^{1/2} n^{-1/2} \tag{44}$$

with conditional probability (over the randomness of $\{\varepsilon_i\}_{i=1}^n$*) greater than* $1 - c_1 n^{-1} - c_2 n^{1-\kappa}$, *where* $C_1, c_1 > 0$ *are absolute constants and* $c_2 > 0$ *is a constant depending only on* a_2, κ, $\mathbb{E}|\varepsilon|$ *and* $\mathbb{E}|\varepsilon|^\kappa$.

Based on Lemma 13, we further study the concentration property of the Wilks expansion for the excess $\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\boldsymbol{0})$. Since the function $\mathcal{L}_n^S(\cdot)$ is concave, we use $\nabla\mathcal{L}_n^S(\cdot)$ to denote its subgradient. For $\boldsymbol{\theta} \in \mathbb{R}^s$, let $\boldsymbol{\zeta}^S(\boldsymbol{\theta}) = \mathcal{L}_n^S(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta})$ be the stochastic component of $\mathcal{L}_n^S(\boldsymbol{\theta})$. Then, it is easy to see that

$$\nabla\boldsymbol{\zeta}^S(\boldsymbol{\theta}) = -2\sum_{i=1}^n w_i^S(\boldsymbol{\theta})\boldsymbol{X}_{iS}, \quad \nabla\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) = -\sum_{i=1}^n \{2\mathbb{P}_{\boldsymbol{X}}(Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - 1\}\boldsymbol{X}_{iS}, \tag{45}$$

where $w_i^S(\boldsymbol{\theta}) := I(Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - \mathbb{P}_{\boldsymbol{X}}(Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})$. In particular, we have $\nabla\boldsymbol{\zeta}^S(\boldsymbol{0}) = -\sum_{i=1}^n\{2I(\varepsilon_i \le 0) - 1\}\boldsymbol{X}_{iS}$. Recall that f_ε and F_ε denote, respectively, the density function and the cumulative distribution function of ε. By the second expression in (45), $\nabla\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) = -\sum_{i=1}^n\{2F_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - 1\}\boldsymbol{X}_{iS}$ and

$$\mathbf{H}_S(\boldsymbol{\theta}) := -\nabla^2\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) = 2\sum_{i=1}^n f_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}. \tag{46}$$

In line with (26), we have $\mathbf{H}_S^* = \mathbf{H}_S(\mathbf{0}) = 2nf_\varepsilon(0)\,\widehat{\mathbf{\Sigma}}_{SS}$, which is the negative Hessian of $\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\mathbf{0})$. As in (27), define the normalized score

$$\widehat{\boldsymbol{\xi}}_S = \mathbf{H}_S^{*-1/2}\nabla\mathcal{L}_n^S(\mathbf{0}) = \{2nf_\varepsilon(0)\}^{-1/2}\,\widehat{\mathbf{\Sigma}}_{SS}^{-1/2}\sum_{i=1}^n\{2I(\varepsilon_i \le 0) - 1\}\boldsymbol{X}_{iS}. \tag{47}$$

The following result is a non-asymptotic, conditional version of the Wilks theorem, saying that with high probability, the square root of the excess $\max_{\boldsymbol{\theta}}\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathcal{L}_n^S(\mathbf{0})$ and the ℓ_2-norm of the normalized score $\widehat{\boldsymbol{\xi}}_S$ are sufficiently close uniformly over all s-subsets $S \subseteq [p]$.

Lemma 14 *Assume that Conditions 3.1 and 3.4 are satisfied. Then*

$$\begin{aligned}\max_{S\subseteq[p]:|S|=s}&\Big|[2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0})\}]^{1/2} - \|\widehat{\boldsymbol{\xi}}_S\|_2\Big| \\ &\le C_2\{f_\varepsilon(0)\}^{-1/2}\big[\lambda_{\min}^{-1/2}(s)\{s\log(pn)\}^{3/2}n^{-1/2} + \lambda_{\min}^{-1/4}(s)s\log(pn)n^{-1/4}\big]\end{aligned} \tag{48}$$

holds with probability greater than $1 - c_2n^{1-\kappa} - c_3n^{-1}$ whenever $n \ge C_3\,\lambda_{\min}^{-1}(s)\{s\log(pn)\}^2$, where $C_2 > 0$ is a constant depending only on a_2, A_2 and A_3, c_2 is as in Lemma 13, $c_3 > 0$ is an absolute constant and $C_3 > 0$ is a constant depending only on a_2 and A_2.

Further, write $\widetilde{\varepsilon}_i = 2I(\varepsilon_i \le 0) - 1$ and $\widetilde{\boldsymbol{X}_i} = \widetilde{\varepsilon}_i\boldsymbol{X}_i$. Note that $\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n$ are i.i.d. Rademacher random variables and thus $\widetilde{\boldsymbol{X}_1}, \ldots, \widetilde{\boldsymbol{X}_n}$ are sub-exponential random vectors. In this notation, we have $\widehat{\boldsymbol{\xi}}_S = \{2nf_\varepsilon(0)\}^{-1/2}\sum_{i=1}^n\widehat{\mathbf{\Sigma}}_{SS}^{-1/2}\widetilde{\boldsymbol{X}}_{iS}$. For each $S \subseteq [p]$, define

$$\boldsymbol{\xi}_S = \{2nf_\varepsilon(0)\}^{-1/2}\sum_{i=1}^n\mathbf{\Sigma}_{SS}^{-1/2}\boldsymbol{X}_{iS}.$$

Then, applying Lemma 11 with slight modification and the union bound we obtain that, with probability at least $1 - c_4n^{-1}$,

$$\max_{S\subseteq[p]:|S|=s}\big|\|\widehat{\boldsymbol{\xi}}_S\|_2 - \|\boldsymbol{\xi}_S\|_2\big| \le C_4\{f_\varepsilon(0)\}^{-1/2}s\log(pn)\,n^{-1/2} \tag{49}$$

for all $n \ge C_5\,s\log(pn)$, where $c_4 > 0$ is an absolute constant and $C_4, C_5 > 0$ are constants depending only on A_0.

Observe that $\mathbb{E}(\widetilde{\boldsymbol{X}_i}) = \mathbb{E}[\boldsymbol{X}_i\{2\mathbb{P}(\varepsilon_i \le 0|\boldsymbol{X}_i) - 1\}] = 0$ and $\mathbb{E}(\widetilde{\boldsymbol{X}_i}\widetilde{\boldsymbol{X}_i}^{\mathrm{T}}) = \mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}) = \mathbf{\Sigma}$. Hence, it follows from Lemma 12 that there exists a random variable $T_0 \overset{d}{=} R_0(s,p)$ such that for any $\delta \in (0,1]$,

$$\Big|\sqrt{2f_\varepsilon(0)}\max_{S\subseteq[p]:|S|=s}\|\boldsymbol{\xi}_S\|_2 - T_0\Big| \le C_6\big[\delta + \{s\log(\gamma_spn)\}^{1/2}n^{-1/2} + \{s\log(\gamma_spn)\}^2n^{-3/2}\big] \tag{50}$$

holds with probability at least $1 - C_7\big[\delta^{-3}n^{-1/2}\{s\log(\gamma_spn)\}^2 \vee \delta^{-4}n^{-1}\{s\log(\gamma_spn)\}^5\big]$, where $C_6, C_7 > 0$ are constants depending only on A_0.

Finally, combining (48), (49), (50) and (42) proves (20). ∎

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation Grants DMS-1206464 and DMS-1406266 and the National Institutes of Health Grant R01-GM072611-10.

References

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer-Verlag, Berlin Heidelberg, 2011.

T. Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39, 2012.

T. Tony Cai, Jianqing Fan, and Tiefeng, Jiang. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14:1837–1864, 2013.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

Jianqing Fan and Runze Li. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. TAC for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*, 2015.

Jianqing Fan, Qi-Man Shao, and Wen-Xin Zhou. Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *arXiv preprint arXiv:1502.04237*, 2015.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 2015.

Roger Koenker. *Quantile Regression.* Cambridge University Press, Cambridge, 2005.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

Gangadharrao S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge, 1983.

Peter McCullagh and John A. Nelder. *Generalized Linear Models.* Chapman & Hall/CRC, London, 1989.

André Oberthuer, Frank Berthold, Patrick Warnat, Barbara Hero, Yvonne Kahlert, Rüdiger Spitz, Karen Ernestus, Rainer König, Stefan Haas, Roland Eils, Manfred Schwab, Benedikt Brors, Frank Westermann, and Matthias Fischer. Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, 24(31):5070–5078, 2006.

Leming Shi, et al. (MAQC Consortium). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–841, 2010.

Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

Vladimir Spokoiny. Bernstein-von Mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*, 2013.

Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Gudmundur A. Thorisson, Albert V. Smith, Lalitha Krishnan, and Lincoln D. Stein. The International HapMap Project Web site. *Genome Research*, 15:1592–1593, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268, Cambridge University Press, Cambridge, 2012.

Lie Wang. The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

Samuel S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.

Appendix A. Appendix A.

In this appendix we prove the technical lemmas appeared in Section 6.

A.1 Proof of Lemma 9

Define the loss function $\ell(y,z) = yz - b(z)$ for $y, z \in \mathbb{R}$. For each s-subset $S \subseteq [p]$ and $\boldsymbol{\theta} \in \mathbb{R}^s$, define $\boldsymbol{\zeta}^S(\boldsymbol{\theta}) = \mathcal{L}_n^S(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{\zeta}_i^S(\boldsymbol{\theta})$, where $\boldsymbol{\zeta}_i^S(\boldsymbol{\theta}) = \ell(Y_i, \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\ell(Y_i, \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})$. Note that $\nabla\boldsymbol{\zeta}_i^S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}} = \varepsilon_i \boldsymbol{X}_{iS}$ with $\varepsilon_i = Y_i - b'(0)$. Thus, we have $\mathbf{V}_0^2 := \mathrm{Var}_{\boldsymbol{X}}\{\nabla\boldsymbol{\zeta}^S(\mathbf{0})\} = n\phi b''(0)\,\widehat{\boldsymbol{\Sigma}}_{SS}$.

For every $\boldsymbol{u} \in \mathbb{R}^s \setminus \{\mathbf{0}\}$ and $u \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}} \exp\left\{u\frac{\boldsymbol{u}^{\mathrm{T}}\nabla\boldsymbol{\zeta}^S(\mathbf{0})}{\|\mathbf{V}_0\boldsymbol{u}\|_2}\right\} &= \prod_{i=1}^n \mathbb{E}_{\boldsymbol{X}} \exp\left(u\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}}{\|\mathbf{V}_0\boldsymbol{u}\|_2}\varepsilon_i\right) \\
&= \prod_{i=1}^n \mathbb{E}_{\boldsymbol{X}} \exp\left\{\frac{u}{\sqrt{n}} \times \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}}{(\boldsymbol{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{SS}\boldsymbol{u})^{1/2}} \times \frac{\varepsilon_i}{(\mathrm{Var}\,\varepsilon_i)^{1/2}}\right\} \\
&\le \exp\left\{\frac{1}{2}a_0u^2 \times \frac{1}{n}\sum_{i=1}^n \frac{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS})^2}{\boldsymbol{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{SS}\boldsymbol{u}}\right\} = \exp(a_0u^2/2).
\end{aligned}$$

This verifies condition (ED_0) with $\nu_0^2 = a_0$ in Theorem B.3 from the supplement of Spokoiny and Zhilova (2015). Consequently, taking $\mathbb{B}^2 = \mathbf{H}_S^{*-1/2}\mathbf{V}_0^2\mathbf{H}_S^{*-1/2} = \phi\mathbf{I}_s$ and $\mathrm{g} = \{C\mathrm{tr}(\mathbb{B}^2)\}^{1/2}$ for some $C \ge 2$ there, we have $\lambda_{\max}(\mathbb{B}^2) = \phi$, $\mathrm{tr}(\mathbb{B}^2) = \phi s$, $\mathrm{tr}(\mathbb{B}^4) = \phi^2 s$ and $\mathrm{x}_c = \frac{1}{2}(\frac{3}{2}C - 1 - \log 3)s \ge \frac{3}{4}(C-2)s$. This implies that almost surely on the event $\{\widehat{\boldsymbol{\Sigma}}_{SS} \succ \mathbf{0}\}$, with conditional probability at least $1 - 2e^{-t} - 8.4\,e^{-\mathrm{x}_c}$,

$$\|\widehat{\boldsymbol{\xi}}_S\|_2^2 \le a_0\phi \times \begin{cases} s + (8ts)^{1/2}, & \text{if } 0 \le t \le \frac{1}{18}(2s)^{1/2}, \\ s + 6t, & \text{if } \frac{1}{18}(2s)^{1/2} < t \le \mathrm{x}_c. \end{cases}$$

Finally, letting $C \to \infty$ proves (28). ∎

A.2 Proof of Lemma 10

We prove this lemma by applying the conditional version of Theorem 2.3 in Spokoiny (2013). To this end, we need to verify conditions (ED_0), (ED_2), $(\mathcal{L}_0)$, $(\mathcal{I})$ and $(\mathcal{L})$. In line with the notation used therein, we fix $S \subseteq [p]$ and write

$$\mathbf{D}^2(\boldsymbol{\theta}) = -\nabla^2\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = \sum_{i=1}^n b''(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}, \quad \mathbf{D}_0^2 = \mathbf{D}^2(\mathbf{0}) = nb''(0)\,\widehat{\boldsymbol{\Sigma}}_{SS}.$$

The validity of (ED_0) is guaranteed from the proof of Lemma 9, and (ED_2) is automatically satisfied with $\omega \equiv 0$ since $\nabla^2\boldsymbol{\zeta}^S(\boldsymbol{\theta})$ vanishes for all $\boldsymbol{\theta} \in \mathbb{R}^s$. Turning to $(\mathcal{L}_0)$, observe

that

$$\begin{aligned}
&\left\|\mathbf{D}_0^{-1}\mathbf{D}^2(\boldsymbol{\theta})\mathbf{D}_0^{-1} - \mathbf{I}_s\right\| \\
&= \left\|\mathbf{D}_0^{-1}\sum_{i=1}^n \{b''(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - b''(0)\}\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}\mathbf{D}_0^{-1}\right\| \\
&= \left\|\mathbf{D}_0^{-1}\sum_{i=1}^n b'''(\eta_i)\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}\, \boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}\mathbf{D}_0^{-1}\right\|, \qquad (51)
\end{aligned}$$

where η_i lies between 0 and $\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}$. For $r > 0$, define $\Theta_0(r) = \{\boldsymbol{\theta} \in \mathbb{R}^s : \|\mathbf{D}_0\boldsymbol{\theta}\|_2 \le r\}$. On the event $\mathcal{E}_0(\tau)$ for some $\tau > 0$ and for $\boldsymbol{\theta} \in \Theta_0(r)$,

$$|\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}| = |\boldsymbol{\theta}^{\mathrm{T}}\mathbf{D}_0\mathbf{D}_0^{-1}\boldsymbol{X}_{iS}| \le \|\mathbf{D}_0^{-1}\boldsymbol{X}_{iS}\|_2 \le \{nb''(0)\}^{-1/2}\tau^{1/2}r. \qquad (52)$$

This together with (51) implies that

$$\left\|\mathbf{D}_0^{-1}\mathbf{D}^2(\boldsymbol{\theta})\mathbf{D}_0^{-1} - \mathbf{I}_s\right\| \le \frac{\max_{|t|\le\{nb''(0)\}^{-1/2}\tau^{1/2}r}|b'''(t)|}{\{b''(0)\}^{3/2}}\frac{\tau^{1/2}r}{n^{1/2}} := \delta(\tau, r). \qquad (53)$$

Recalling that $\mathbf{V}_0^2 = \mathrm{Var}_{\boldsymbol{X}}\{\boldsymbol{\zeta}^S(\mathbf{0})\} = \phi\mathbf{D}_0^2$, $(\mathcal{I})$ is satisfied with $a = \phi^{1/2}$.

To verify $(\mathcal{L}\mathrm{r})$, define $g(t) = b'(0)t - b(t)$ so that $g'(t) = b'(0) - b'(t)$ and $g''(t) = -b''(t)$. Then, for any $\boldsymbol{\theta} \in \mathbb{R}^s$ satisfying $\|\mathbf{D}_0\boldsymbol{\theta}\|_2 = r > 0$, it follows from the second-order Taylor expansion that

$$\begin{aligned}
&-2\{\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\mathbf{0})\} = -2\sum_{i=1}^n\{g(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - g(0)\} \\
&= -2\sum_{i=1}^n\left\{g'(0)\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} + \tfrac{1}{2}g''(\eta_i)(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})^2\right\} = \sum_{i=1}^n b''(\eta_i)(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})^2, \qquad (54)
\end{aligned}$$

where η_i is a point lying between 0 and $\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}$. On the event $\mathcal{E}_0(\tau)$, the right-hand side of (54) is further bounded from below by

$$r^2\{b''(0)\}^{-1}\min_{|t|\le\{nb''(0)\}^{-1/2}\tau^{1/2}r} b''(t).$$

When $\|\mathbf{D}_0\boldsymbol{\theta}\|_2 = r \le \{nb''(0)/\tau\}^{1/2}$, $-2\{\mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\mathcal{L}_n^S(\mathbf{0})\}$ is bounded from below by a_1r^2 for a_1 as in (15). Further, from the convexity of the function $\boldsymbol{\theta} \mapsto -\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathcal{L}_n^S(\mathbf{0})\}$, we see that $-\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathcal{L}_n^S(\mathbf{0})\} \ge a_1r\{nb''(0)/\tau\}^{1/2}$, for all $\boldsymbol{\theta}$ satisfying $\|\mathbf{D}_0\boldsymbol{\theta}\|_2 = r \ge \{nb''(0)/\tau\}^{1/2}$. Define the function $r \mapsto b(r)$ as

$$b(r) = \begin{cases} a_1 & \text{if } 0 \le r \le \{nb''(0)/\tau\}^{1/2}, \\ a_1r^{-1}\{nb''(0)/\tau\}^{1/2} & \text{if } r > \{nb''(0)/\tau\}^{1/2}. \end{cases} \qquad (55)$$

By definition, $rb(r)$ is non-decreasing in $r \ge 0$ and for $\boldsymbol{\theta} \in \mathbb{R}^s$ satisfying $\|\mathbf{D}_0\boldsymbol{\theta}\|_2 = r$,

$$\frac{-2\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta}) - \mathcal{L}_n^S(\mathbf{0})\}}{\|\mathbf{D}_0\boldsymbol{\theta}\|_2^2} \ge b(r). \qquad (56)$$

With the above preparations, we apply Theorem 2.3 in Spokoiny (2013) with slight modification on the constant. In view of (29) and (55), set

$$r_0 = 2(\phi a_0)^{1/2} a_1^{-1} \big[s + 6\big(s \log \tfrac{ep}{s} + \log n\big)\big]^{1/2}, \tag{57}$$

such that Condition 2.3 there is satisfied on $\mathcal{E}_0(\tau)$ whenever $n \geq \{b''(0)\}^{-1} r_0^2 \tau$. Hence, it follows from Theorem 2.3 in Spokoiny (2013) and the union bound that, conditional on the event $\mathcal{E}_0(\tau)$,

$$\mathbb{P}_{\boldsymbol{X}}\bigg(\max_{S\subseteq[p]:|S|=s} \big| [2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0})\}]^{1/2} - \|\widehat{\boldsymbol{\xi}}_S\|_2 \big| \leq 5\delta(\tau, r_0) r_0 \bigg) \leq 5n^{-1}, \tag{58}$$

where $\delta(\tau, r)$ and r_0 are as in (53) and (57), respectively. This proves (31) by properly choosing C_1 and C_2. ∎

A.3 Proof of Lemma 11

To begin with, note that for each s-subset $S \subseteq [p]$, $\boldsymbol{Z}_{1S}, \ldots, \boldsymbol{Z}_{nS}$ are i.i.d. s-dimensional random vectors with mean zero and covariance matrix $\phi \boldsymbol{\Sigma}_{SS}$. By (27) and (37),

$$\|\widehat{\boldsymbol{\xi}}_S\|_2^2 - \|\boldsymbol{\xi}_S\|_2^2 = \boldsymbol{\xi}_S^{\mathrm{T}} \big(\boldsymbol{\Sigma}_{SS}^{1/2} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \boldsymbol{\Sigma}_{SS}^{1/2} - \mathbf{I}_s \big) \boldsymbol{\xi}_S.$$

Write $\mathbb{X}_S = (\boldsymbol{X}_{1S}, \ldots, \boldsymbol{X}_{nS})^{\mathrm{T}} \in \mathbb{R}^{n\times s}$, then $\mathbb{X}_S \boldsymbol{\Sigma}_{SS}^{-1/2}$ is an $n \times s$ matrix whose rows are independent sub-Gaussian random vectors in $\mathbb{R}^s$. Further, observe that $\boldsymbol{X}_{iS} = \mathbf{P}_S \boldsymbol{X}_i$ and $\boldsymbol{\Sigma}_{SS} = \mathbf{P}_S \boldsymbol{\Sigma} \mathbf{P}_S^{\mathrm{T}}$, where $\mathbf{P}_S \in \mathbb{R}^{s\times p}$ is a projection matrix. Under Condition 3.1, $\|\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{X}_{iS}\|_{\psi_2} = \|\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \mathbf{P}_S \boldsymbol{\Sigma}_{SS}^{1/2} \boldsymbol{U}\|_{\psi_2} \leq A_0 \|\boldsymbol{\Sigma}_{SS}^{1/2} \mathbf{P}_S^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{u}\|_2 = A_0$ for $\boldsymbol{u} \in \mathbb{S}^{s-1}$. Then, it follows from (32) that for all sufficient large n so that $\delta \leq \frac{1}{2}$, $\|\boldsymbol{\Sigma}_{SS}^{1/2} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \boldsymbol{\Sigma}_{SS}^{1/2} - \mathbf{I}_s\| \leq 2\delta$ and hence,

$$\begin{aligned} |\|\widehat{\boldsymbol{\xi}}_S\|_2 - \|\boldsymbol{\xi}_S\|_2| &= \frac{|\|\widehat{\boldsymbol{\xi}}_S\|_2^2 - \|\boldsymbol{\xi}_S\|_2^2|}{\|\widehat{\boldsymbol{\xi}}_S\|_2 + \|\boldsymbol{\xi}_S\|_2} \\ &\leq \|\boldsymbol{\xi}_S\|_2^{-1} \times |\|\widehat{\boldsymbol{\xi}}_S\|_2^2 - \|\boldsymbol{\xi}_S\|_2^2| \leq 2C_3 (s \vee t)^{1/2} n^{-1/2} \times \|\boldsymbol{\xi}_S\|_2. \end{aligned} \tag{59}$$

Next we upper bound the quadratic term $\|\boldsymbol{\xi}_S\|_2$. First we show that $\boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{Z}_{iS} = \phi^{1/2}\, \boldsymbol{\Sigma}_{SS}^{-1/2} \widetilde{\varepsilon}_i \boldsymbol{X}_i$ are sub-exponential random vectors, where $\widetilde{\varepsilon}_i := \varepsilon_i / (\operatorname{Var} \varepsilon_i)^{1/2}$. In fact, for every $\boldsymbol{u} \in \mathbb{S}^{s-1}$, $\|\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{Z}_{iS}\|_{\psi_1} \leq 2 \|\widetilde{\varepsilon}_i\|_{\psi_2} \|\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{X}_{iS}\|_{\psi_2} \leq 2A_0' A_0$, where $A_0' > 0$ is a constant depending only on a_0 in Condition 3.1. Following the proof of Lemma 5.15 in Vershynin (2012), we derive that for every $\boldsymbol{u} \in \mathbb{R}^s$ satisfying $\|\boldsymbol{u}\|_2 \leq \phi^{-1/2} (4eA_0' A_0)^{-1} \sqrt{n}$,

$$\begin{aligned} \log \mathbb{E} \exp(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{\xi}_S) &= \sum_{i=1}^n \log \mathbb{E} \exp(n^{-1/2} \boldsymbol{u}^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{Z}_{iS}) \\ &\leq 2e^2 \|\boldsymbol{u}\|_2^2\, n^{-1} \sum_{i=1}^n \big\| (\boldsymbol{u}/\|\boldsymbol{u}\|_2)^{\mathrm{T}} \boldsymbol{\Sigma}_{SS}^{-1/2} \boldsymbol{Z}_{iS} \big\|_{\psi_1}^2 \\ &\leq (4eA_0' A_0)^2 \phi \frac{\|\boldsymbol{u}\|_2^2}{2}. \end{aligned}$$

Consequently, applying Corollary 1.12 in the supplement of Spokoiny (2012) with $\mathrm{g} = \sqrt{n}$, $\mathbb{B} = \mathbf{I}_s$ and $\mathrm{x}_c = \frac{3}{4}n - \frac{1}{2}(1+\log 3)s \geq \frac{3}{4}n - \frac{3}{2}s$ to the random vector $(4eA_0'A_0)^{-1}\phi^{-1/2}\boldsymbol{\xi}_S$ yields that, for every $0 \leq t \leq \mathrm{x}_c$,

$$\mathbb{P}\big[\|\boldsymbol{\xi}_S\|_2 \geq 4eA_0'A_0\{\phi\Delta(s,t)\}^{1/2}\big] \leq 2e^{-t} + 8.4\, e^{-\mathrm{x}_c}. \tag{60}$$

Finally, combining (59) and (60) completes the proof of (38). ■

A.4 Proof of Lemma 12

First, observe that

$$\max_{S\subseteq[p]:|S|=s} \|\boldsymbol{\xi}_S\|_2 = \max_{\boldsymbol{u}\in\mathcal{F}(s,p)} n^{-1/2} \sum_{i=1}^n \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}_i}{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u})^{1/2}},$$

where $\mathcal{F}(s,p) = \{\boldsymbol{x} \mapsto \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} : \boldsymbol{u} \in \mathbb{S}^{p-1}, \|\boldsymbol{u}\|_0 \leq s\}$. Recall that $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are i.i.d. p-dimensional centered random vectors with covariance matrix $\mathbb{E}(\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}) = \phi\boldsymbol{\Sigma}$. As in the proof of Lemma 11, we have for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$,

$$\|\phi^{-1/2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}_i\|_{\psi_1} \leq 2\|\varepsilon_i/(\operatorname{Var}\varepsilon_i)^{1/2}\|_{\psi_2}\, \|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_i\|_{\psi_2} \leq 2A_0'A_0(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u})^{1/2}.$$

Consequently, it follows from Lemma 7.5 in Fan, Shao and Zhou (2015) that there exists a random variable $T_0 \overset{d}{=} R_0(s,p) = \max_{\boldsymbol{u}\in\mathcal{F}(s,p)} \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{G}}{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u})^{1/2}}$ for $\boldsymbol{G} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ such that, for any $\delta \in (0,1]$,

$$\begin{aligned}\mathbb{P}\bigg\{\bigg|\max_{S\subseteq[p]:|S|=s} \|\phi^{-1/2}\boldsymbol{\xi}_S\|_2 - T_0\bigg| \geq C_1A_0'A_0\bigg(\delta + \frac{\gamma_{s,p,n}^{1/2}}{\sqrt{n}} + \frac{\gamma_{s,p,n}^2}{n^{3/2}}\bigg)\bigg\} \\ \leq C_2\bigg[\frac{\{s\log(\gamma_s pn)\}^2}{\delta^3\sqrt{n}} + \frac{\{s\log(\gamma_s pn)\}^5}{\delta^4 n}\bigg],\end{aligned}$$

where $\gamma_{s,p,n} = s\log\frac{\gamma_s ep}{s} + \log n$ and $C_1, C_2 > 0$ are absolute constants. This proves (40). ■

A.5 Proof of Lemma 13

The proof employs techniques from empirical process theory which modify the arguments used in Wang (2013). To begin with, note that

$$\widehat{\boldsymbol{\theta}}_S = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^s} f(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^s} \|\boldsymbol{Y} - \mathbb{X}_S\boldsymbol{\theta}\|_1.$$

Under the null model, $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \mathbb{X}_S\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\theta}^* = \mathbf{0}$. Then the sub-differential of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \mathbf{0}$ can be written as $\nabla f(\mathbf{0}) = -\mathbb{X}_S^{\mathrm{T}}\operatorname{sgn}(\boldsymbol{\varepsilon})$, where $\operatorname{sgn}(\boldsymbol{\varepsilon}) = (\operatorname{sgn}(\varepsilon_1), \ldots, \operatorname{sgn}(\varepsilon_n))^{\mathrm{T}}$ with $\operatorname{sgn}(u) := I(u > 0) - I(u < 0)$. Define $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}} = \operatorname{sgn}(\boldsymbol{\varepsilon})$, and note that $z_1, \ldots, z_n$ are i.i.d. random variables satisfying $\mathbb{P}(z_i = 1) = \mathbb{P}(z_i = -1) = 1/2$.

Since $\widehat{\boldsymbol{\theta}}_S$ minimizes $\|\boldsymbol{Y} - \mathbb{X}_S\boldsymbol{\theta}\|_1$ over $\mathbb{R}^s$, we have the following basic inequality

$$\|\boldsymbol{Y} - \mathbb{X}_S\widehat{\boldsymbol{\theta}}_S\|_1 = \|\mathbb{X}_S\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\varepsilon}\|_1 \leq \|\boldsymbol{\varepsilon}\|_1. \tag{61}$$

Further, define a random process $\{Q(\boldsymbol{\theta})\}$ indexed by $\boldsymbol{\theta} \in \mathbb{R}^s$:

$$Q(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} \big(|\boldsymbol{X}_{iS}^{\mathrm{T}} \boldsymbol{\theta} - \varepsilon_i| - |\varepsilon_i| \big). \tag{62}$$

In what follows, we prove that with overwhelmingly high probability , $Q(\boldsymbol{\theta})$ is concentrated around its expectation $Q_{\mathbb{X}}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{X}}\{Q(\boldsymbol{\theta})\}$ uniformly over $\boldsymbol{\theta} \in \mathbb{R}^s$ via a straightforward adaptation of the peeling argument.

For $\delta_1 > 0$ and $\ell = 1, 2, \ldots$, consider the following sequence of events

$$\mathcal{G}(\delta_1) = \big\{ \boldsymbol{\theta} \in \mathbb{R}^s : \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2} \boldsymbol{\theta}\|_2 \geq \delta_1 \big\}, \quad \mathcal{G}_\ell(\delta_1) = \big\{ \boldsymbol{\theta} \in \mathbb{R}^s : \alpha^{\ell-1} \delta_1 \leq \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2} \boldsymbol{\theta}\|_2 \leq \alpha^{\ell} \delta_1 \big\}, \tag{63}$$

where $\alpha = \sqrt{2}$. Here, δ_1 can be regarded as a tolerance parameter, and it is easy to see that $\mathcal{G}(\delta_1) = \cup_{\ell=1}^{\infty} \mathcal{G}_\ell(\delta_1)$. For $R > 0$, set $\mathcal{V}(R) = \{\boldsymbol{\theta} \in \mathcal{G}(\delta_1) : \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2} \boldsymbol{\theta}\|_2 \leq R\}$ and let $\Delta(R)$ be the maximum deviation over the elliptic vicinity $\mathcal{V}(R)$:

$$\Delta(R) = \max_{\boldsymbol{\theta} \in \mathcal{V}(R)} |Q(\boldsymbol{\theta}) - Q_{\mathbb{X}}(\boldsymbol{\theta})|. \tag{64}$$

For every $\boldsymbol{\theta} \in \mathbb{R}^s$, define the rescaled vector $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\Sigma}}_{SS}^{1/2} \boldsymbol{\theta}$ such that

$$\Delta(R) = \max_{\delta_1 \leq \|\widetilde{\boldsymbol{\theta}}\|_2 \leq R} \Big| Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) \Big|.$$

For every $0 < \epsilon \leq R$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of the Euclidean ball $\mathbb{B}_2^s(R)$ with cardinality bounded by $\big(1 + \frac{2R}{\epsilon}\big)^s$. For $\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2 \in \mathbb{B}_2^s(R)$ satisfying $\|\widetilde{\boldsymbol{\theta}}_1 - \widetilde{\boldsymbol{\theta}}_2\|_2 \leq \epsilon$, observe that

$$\begin{aligned} \Big| Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}_1) - Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}_2) \Big| &\leq n^{-1/2} \sum_{i=1}^{n} \Big| \boldsymbol{X}_{iS}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} (\widetilde{\boldsymbol{\theta}}_1 - \widetilde{\boldsymbol{\theta}}_2) \Big| \\ &\leq \Big\| \mathbb{X}_S \widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} (\widetilde{\boldsymbol{\theta}}_1 - \widetilde{\boldsymbol{\theta}}_2) \Big\|_2 \leq \epsilon n^{1/2}. \end{aligned}$$

Then, it is easy to see that

$$\Delta(R) \leq \max_{\widetilde{\boldsymbol{\theta}} \in \mathcal{N}_\epsilon} \Big| Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) \Big| + 2\epsilon n^{1/2}. \tag{65}$$

For each $\widetilde{\boldsymbol{\theta}} \in \mathbb{B}_2^s(R)$ fixed, $Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}})$ is a sum of independent random variables with zero means and for $i = 1, \ldots, n$, $||\boldsymbol{X}_{iS}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}} - \varepsilon_i| - |\varepsilon_i|| \leq |\boldsymbol{X}_{iS}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}|$. Therefore, it follows from Hoeffding's inequality that for every $t > 0$,

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{X}} \Big\{ \Big| Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}}) \Big| \geq t \Big\} \\ &\leq 2 \exp \left\{ - \frac{nt^2}{2 \sum_{i=1}^{n} (\boldsymbol{X}_{iS}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2} \widetilde{\boldsymbol{\theta}})^2} \right\} = 2 \exp \Big(- \frac{t^2}{2 \|\widetilde{\boldsymbol{\theta}}\|_2^2} \Big). \end{aligned}$$

In other words, for every $\widetilde{\boldsymbol{\theta}} \in \mathbb{B}_2^s(R)$ and $\delta > 0$,

$$\Big|Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\widetilde{\boldsymbol{\theta}})\Big| \le (2\delta)^{1/2}\|\widetilde{\boldsymbol{\theta}}\|_2 \le (2\delta)^{1/2}R$$

holds with probability at least $1 - 2e^{-\delta}$. This, together with the union bound yields

$$\mathbb{P}_{\boldsymbol{X}}\bigg\{\max_{\widetilde{\boldsymbol{\theta}}\in\mathcal{N}_\epsilon}\Big|Q(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\widetilde{\boldsymbol{\theta}}) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1/2}\widetilde{\boldsymbol{\theta}})\Big| \ge (2\delta)^{1/2}R\bigg\} \le \exp\big\{s\log\big(1+\tfrac{2R}{\epsilon}\big) - \delta\big\}. \tag{66}$$

In particular, by taking $\epsilon = Rn^{-1}$ in (65) and $\delta = s\log(1+\frac{2R}{\epsilon}) + t \le 2s\log n + t$ in (66) we conclude that

$$\mathbb{P}_{\boldsymbol{X}}\big\{\Delta(R) \ge R(2t)^{1/2} + 2R(s\log n)^{1/2} + 2Rn^{-1/2}\big\} \le 2e^{-t} \tag{67}$$

holds almost surely on the event $\mathcal{E}_0(\tau)$ for any $\tau > 0$.

In particular, by taking $t = cnR^2$ in (67) for some $c > 0$ to be specified below (72) and the union bound, we have

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{X}}\Big[\exists\,\boldsymbol{\theta}\in\mathcal{G}(\delta_1), \text{s.t.}\,|Q(\boldsymbol{\theta}) - Q_{\mathbb{X}}(\boldsymbol{\theta})| \ge 2^{3/2}\|\widetilde{\boldsymbol{\theta}}\|_2\big\{\|\widetilde{\boldsymbol{\theta}}\|_2(cn)^{1/2} + (s\log n)^{1/2} + n^{-1/2}\big\}\Big]\\
&\le \sum_{\ell=1}^{\infty}\mathbb{P}_{\boldsymbol{X}}\Big[\exists\,\boldsymbol{\theta}\in\mathcal{G}_\ell(\delta_1), \text{s.t.}\,|Q(\boldsymbol{\theta}) - Q_{\mathbb{X}}(\boldsymbol{\theta})| \ge (\alpha^\ell\delta_1)^2(2cn)^{1/2} + 2\alpha^\ell\delta_1\big\{(s\log n)^{1/2} + n^{-1/2}\big\}\Big]\\
&\le \sum_{\ell=1}^{\infty}\mathbb{P}_{\boldsymbol{X}}\Big[\Delta(\alpha^\ell\delta_1) \ge (\alpha^\ell\delta_1)^2(2cn)^{1/2} + 2\alpha^\ell\delta_1\big\{(s\log n)^{1/2} + n^{-1/2}\big\}\Big]\\
&\le 2\sum_{\ell=1}^{\infty}\exp\{-cn(\alpha^\ell\delta_1)^2\} \le 2\sum_{\ell=1}^{\infty}\exp\{-2c\ell\log(\alpha)n\delta_1^2\} \le \frac{2\exp(-c_0n\delta_1^2)}{1-\exp(-c_0n\delta_1^2)},
\end{aligned}$$

where $c_0 = c\log 2$. This implies that with probability at least $1 - 4\exp(-c_0n\delta_1^2)$,

$$|Q(\boldsymbol{\theta}) - Q_{\mathbb{X}}(\boldsymbol{\theta})| \le 2^{3/2}\sqrt{c}\,\|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2}\boldsymbol{\theta}\|_2^2 + 2^{3/2}\|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2}\boldsymbol{\theta}\|_2\big\{(s\log n)^{1/2} + n^{-1/2}\big\} \tag{68}$$

holds for all $\boldsymbol{\theta} \in \mathcal{G}(\delta_1)$ whenever $n \ge c^{-1}\delta_1^{-2}$.

For the (conditional) expectation

$$Q_{\mathbb{X}}(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{X}}\big(|\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} - \varepsilon_i| - |\varepsilon_i|\big) = n^{-1/2}\big(\mathbb{E}_{\boldsymbol{X}}\|\mathbb{X}_S\boldsymbol{\theta} - \boldsymbol{\varepsilon}\|_1 - \mathbb{E}\|\boldsymbol{\varepsilon}\|_1\big),$$

applying Lemmas 5 and 6 in Wang (2013) with slight modifications gives

$$Q_{\mathbb{X}}(\boldsymbol{\theta}) \ge \begin{cases} \frac{1}{4\sqrt{n}}\|\mathbb{X}_S\boldsymbol{\theta}\|_1 = \frac{\sqrt{n}}{4}\|n^{-1}\mathbb{X}_S\boldsymbol{\theta}\|_1 & \text{if } \|\mathbb{X}_S\boldsymbol{\theta}\|_1 \ge \frac{2n}{a_2},\\ \frac{a_2}{8\sqrt{n}}\|\mathbb{X}_S\boldsymbol{\theta}\|_2^2 = \frac{a_2\sqrt{n}}{8}\|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2}\boldsymbol{\theta}\|_2^2 & \text{if } \|\mathbb{X}_S\boldsymbol{\theta}\|_1 < \frac{2n}{a_2}, \end{cases} \tag{69}$$

where a_2 is as in Condition 3.4. For the sequence of LAD estimators $\{\widehat{\boldsymbol{\theta}}_S : S \subseteq [p], |S| = s\}$, from (61) it can be seen that $\|\mathbb{X}_S\widehat{\boldsymbol{\theta}}_S\|_1 \le \|\mathbb{X}_S\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\varepsilon}\|_1 + \|\boldsymbol{\varepsilon}\|_1 \le 2\|\boldsymbol{\varepsilon}\|_1$, and hence

$$\max_{S\subseteq[p]:|S|=s}\|n^{-1}\mathbb{X}_S\widehat{\boldsymbol{\theta}}_S\|_1 \le 2\bigg\{\mathbb{E}|\varepsilon| + n^{-1}\sum_{i=1}^{n}(|\varepsilon_i| - \mathbb{E}|\varepsilon_i|)\bigg\}.$$

For every $t > 0$ and $1 < \kappa \le 2$, by Markov's inequality we have

$$\mathbb{P}\Big\{ \sum_{i=1}^{n} (|\varepsilon_i| - \mathbb{E}|\varepsilon_i|) \ge t \Big\} \le t^{-\kappa}\, \mathbb{E} \Big| \sum_{i=1}^{n} (|\varepsilon_i| - \mathbb{E}|\varepsilon_i|) \Big|^{\kappa} \le 4^{2-\kappa} t^{-\kappa} n\, \mathbb{E}|\varepsilon|^{\kappa},$$

where we used the inequality $|1 + x|^{\kappa} \le 1 + \kappa x + 2^{2-\kappa}|x|^{\kappa}$ for $1 < \kappa \le 2$ and $x \in \mathbb{R}$. The last two displays together imply that, with probability at least $1 - \delta_2$,

$$\max_{S \subseteq [p]: |S| = s} \|n^{-1} \mathbb{X}_S \widehat{\boldsymbol{\theta}}_S\|_1 \le 2\mathbb{E}|\varepsilon| \big\{ 1 + 4^{(2-\kappa)/\kappa} (\mathbb{E}|\varepsilon|)^{-1} (\mathbb{E}|\varepsilon|^{\kappa})^{1/\kappa} \delta_2^{-1/\kappa} n^{-1+1/\kappa} \big\}.$$

By Condition 3.4, we have $a_2 \mathbb{E}|\varepsilon| < 1$. Therefore, as long as the sample size n satisfies

$$n \ge \left\{ \frac{4^{2-q} a_2^{\kappa}\, \mathbb{E}|\varepsilon|^{\kappa}}{(1 - a_2\, \mathbb{E}|\varepsilon|)^{\kappa}} \right\}^{1/(\kappa-1)} \delta_2^{-1/(\kappa-1)}, \tag{70}$$

the event

$$\mathcal{E}_1 := \left\{ \max_{S \subseteq [p]: |S| = s} \|n^{-1} \mathbb{X}_S \widehat{\boldsymbol{\theta}}_S\|_1 \le 2a_2^{-1} \right\} \tag{71}$$

occurs with probability at least $1 - \delta_2$.

Now, by (61), we have $Q(\widehat{\boldsymbol{\theta}}_S) \le 0$ and thus $-\{Q(\widehat{\boldsymbol{\theta}}_S) - Q_{\mathbb{X}}(\widehat{\boldsymbol{\theta}}_S)\} \ge Q_{\mathbb{X}}(\widehat{\boldsymbol{\theta}}_S)$ holds for every s-subset $S \subseteq [p]$. Together with (68)–(71) and the union bound, this implies that on the event $\mathcal{E}_0(\tau) \cap \mathcal{E}_1$ for any $\tau > 0$,

$$\max_{S \subseteq [p]: |S| = s} \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2} \widehat{\boldsymbol{\theta}}_S\|_2 \le \min \left[\delta_1, 32\sqrt{2}\, a_2^{-1} \left\{ \left(\frac{s \log n}{n} \right)^{1/2} + \frac{1}{n} \right\} \right] \tag{72}$$

holds with (conditional) probability $1 - 4\binom{p}{s} \exp(-c_0 n \delta_1^2) - \delta_2$, provided that the sample size n satisfies $n \ge 2 \cdot 32^2 (a_2 \delta_1)^{-2}$ and (70).

Finally, taking

$$\delta_1 = \frac{32}{a_2} \sqrt{\frac{2}{\log(2)}} \left(\frac{s \log \frac{ep}{s} + \log n}{n} \right)^{1/2} \quad \text{and} \quad \delta_2 = \frac{4^{2-q} a_2^{\kappa}\, \mathbb{E}|\varepsilon|^{\kappa}}{(1 - a_2\, \mathbb{E}|\varepsilon|)^{\kappa}} \frac{1}{n^{\kappa-1}}$$

in (72) proves (44). ■

A.6 Proof of Lemma 14

We prove this lemma by employing the arguments similar to those used in Spokoiny (2013), where the likelihood function $\mathcal{L}(\boldsymbol{\theta})$ is assumed to be twice differentiable with respect to $\boldsymbol{\theta}$. It is worth noticing that both Conditions $(\mathcal{L})$ and (ED_2) in Spokoiny (2013) are not satisfied in the current situation. We provide here a self-contained proof in which Lemma 13 also plays an important role.

Step 1: Local linear approximation of $\nabla \mathcal{L}_n^S(\boldsymbol{\theta})$. Let $\chi_1^S(\boldsymbol{\theta})$ be the normalized residual of the local linear approximation of $\nabla \mathcal{L}_n^S(\boldsymbol{\theta})$ given by

$$\begin{aligned} \chi_1^S(\boldsymbol{\theta}) &= \mathbf{D}_0^{-1} \{ \nabla \mathcal{L}_n^S(\boldsymbol{\theta}) - \nabla \mathcal{L}_n^S(\mathbf{0}) + \mathbf{D}_0^2 \boldsymbol{\theta} \} \\ &= \mathbf{D}_0^{-1} \{ \boldsymbol{U}(\boldsymbol{\theta}) + \nabla \mathbb{E}_{\boldsymbol{X}} \mathcal{L}_n^S(\boldsymbol{\theta}) - \nabla \mathbb{E}_{\boldsymbol{X}} \mathcal{L}_n^S(\mathbf{0}) + \mathbf{D}_0^2 \boldsymbol{\theta} \}, \end{aligned} \tag{73}$$

where $\boldsymbol{U}(\boldsymbol{\theta}) = \nabla\boldsymbol{\zeta}^S(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^S(\mathbf{0})$ and $\mathbf{D}_0^2 = -\nabla^2\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\mathbf{0})\} = 2f_\varepsilon(0)\sum_{i=1}^n \boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}$. Then it follows from the mean value theorem that

$$\mathbb{E}_{\boldsymbol{X}}\{\chi_1^S(\boldsymbol{\theta})\} = \{\mathbf{I}_s - \mathbf{D}_0^{-1}\mathbf{D}^2(\widetilde{\boldsymbol{\theta}})\mathbf{D}_0^{-1}\}\mathbf{D}_0\boldsymbol{\theta}, \tag{74}$$

where $\mathbf{D}^2(\boldsymbol{\theta}) = -\nabla^2\mathbb{E}_{\boldsymbol{X}}\{\mathcal{L}_n^S(\boldsymbol{\theta})\} = 2\sum_{i=1}^n f_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}} = \lambda\boldsymbol{\theta}$ for some $0 \le \lambda \le 1$. As before, for every $r \ge 0$, define the local elliptic neighborhood of $\mathbf{0}$ as

$$\Theta_0(r) = \{\boldsymbol{\theta} \in \mathbb{R}^s : \|\mathbf{D}_0\boldsymbol{\theta}\|_2 \le r\}.$$

On the event $\mathcal{E}_0(\tau)$ for some $\tau > 0$,

$$|\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}| \le \|\mathbf{D}_0\boldsymbol{\theta}\|_2\|\mathbf{D}_0^{-1}\boldsymbol{X}_{iS}\|_2 \le \{2nf_\varepsilon(0)\}^{-1/2}\tau^{1/2}r \tag{75}$$

for all $\boldsymbol{\theta} \in \Theta_0(r)$. Thus it follows from the Taylor expansion that for $r \le \{2nf_\varepsilon(0)/\tau\}^{1/2}$,

$$\begin{aligned}&\big\|\mathbf{I}_s - \mathbf{D}_0^{-1}\mathbf{D}^2(\widetilde{\boldsymbol{\theta}})\mathbf{D}_0^{-1}\big\| \\ &= 2\bigg\|\mathbf{D}_0^{-1}\sum_{i=1}^n\{f_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\widetilde{\boldsymbol{\theta}}) - f_\varepsilon(0)\}\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^{\mathrm{T}}\mathbf{D}_0^{-1}\bigg\| \le \frac{A_3}{\sqrt{2}f_\varepsilon^{3/2}(0)}\frac{\tau^{1/2}r}{n^{1/2}} := \delta(\tau, r).\end{aligned} \tag{76}$$

Together, (74) and (76) imply that under the same constraint for (76),

$$\|\mathbb{E}_{\boldsymbol{X}}\{\chi_1^S(\boldsymbol{\theta})\}\|_2 \le \delta(\tau, r)r. \tag{77}$$

Turning to the stochastic component $\mathbf{D}_0^{-1}\boldsymbol{U}(\boldsymbol{\theta}) = \chi_1^S(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{X}}\{\chi_1^S(\boldsymbol{\theta})\}$, we aim to bound $\max_{\boldsymbol{\theta}\in\Theta_0(r)}\|\mathbf{D}_0^{-1}\boldsymbol{U}(\boldsymbol{\theta})\|_2$, which can be written as

$$\max_{\boldsymbol{\theta}\in\Theta_0(r),\|\boldsymbol{u}\|_2\le 1}\boldsymbol{u}^{\mathrm{T}}\mathbf{D}_0^{-1}\boldsymbol{U}(\boldsymbol{\theta}) = r^{-1}\max_{\boldsymbol{u},\boldsymbol{\theta}\in\Theta_0(r)}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{\theta}). \tag{78}$$

Note that $\{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{\theta}) : \boldsymbol{v}, \boldsymbol{\theta} \in \mathbb{R}^s\}$ is a bivariate process indexed by $(\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{2s}$. Define

$$\begin{gathered}\bar{\boldsymbol{\theta}} = (\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{2s}, \quad \bar{\mathbf{D}}_0 = \begin{pmatrix}\mathbf{D}_0 & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_0\end{pmatrix} \in \mathbb{R}^{(2s)\times(2s)}, \\ \bar{\boldsymbol{U}}(\bar{\boldsymbol{\theta}}) = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{\theta}), \quad \bar{\Theta}_0(r) = \{\bar{\boldsymbol{\theta}} \in \mathbb{R}^{2s} : \|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2 \le r\}.\end{gathered}$$

In this notation, from (78) and the identity $\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}} = \mathbf{D}_0\boldsymbol{v} + \mathbf{D}_0\boldsymbol{\theta}$, it is easy to see that

$$\max_{\boldsymbol{\theta}\in\Theta_0(r)}\|\mathbf{D}_0^{-1}\boldsymbol{U}(\boldsymbol{\theta})\|_2 \le r^{-1}\max_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}_0(2r)}\bar{\boldsymbol{U}}(\bar{\boldsymbol{\theta}}). \tag{79}$$

Recall that $\nabla\boldsymbol{\zeta}^S(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^S(\mathbf{0}) = -2\sum_{i=1}^n\{I(Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - I(Y_i \le 0) + 1/2 - F_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})\}\boldsymbol{X}_{iS}$, where for $i = 1, \ldots, n$, $I(Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - I(Y_i \le 0) + 1/2 - F_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})$ is equal to

$$\begin{cases}I(0 < Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - \mathbb{P}_{\boldsymbol{X}}(0 < Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) & \text{if } \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} \ge 0,\\ -I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < Y_i \le 0) + \mathbb{P}_{\boldsymbol{X}}(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < Y_i \le 0) & \text{if } \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < 0.\end{cases}$$

For $\boldsymbol{\theta} \in \mathbb{R}^s$, define random variables $\varepsilon_{i,\boldsymbol{\theta}} = I(0 < Y_i \le \boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta}) - I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < Y_i \le 0)$ satisfying

(i) conditional on $\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} \geq 0$, $\varepsilon_{i,\boldsymbol{\theta}} = 1$ with probability $P_{i,\boldsymbol{\theta}} - 1/2$ and $\varepsilon_{i,\boldsymbol{\theta}} = 0$ with probability $3/2 - P_{i,\boldsymbol{\theta}}$;

(ii) conditional on $\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < 0$, $\varepsilon_{i,\boldsymbol{\theta}} = -1$ with probability $1/2 - P_{i,\boldsymbol{\theta}}$ and $\varepsilon_{i,\boldsymbol{\theta}} = 0$ with probability $1/2 + P_{i,\boldsymbol{\theta}}$,

where $P_{i,\boldsymbol{\theta}} = F_\varepsilon(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta})$. In this notation, $\nabla\boldsymbol{\zeta}^S(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^S(\mathbf{0}) = -2\sum_{i=1}^n(\mathrm{Id} - \mathbb{E}_{\boldsymbol{X}})\varepsilon_{i,\boldsymbol{\theta}}\boldsymbol{X}_{iS}$. For every $\lambda \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^s$, we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}} \exp[\lambda \boldsymbol{u}^{\mathrm{T}}\{\nabla\boldsymbol{\zeta}^S(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^S(\mathbf{0})\}] \\
&= \prod_{i=1}^n \Big[\mathbb{E}_{\boldsymbol{X}}\{e^{-2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(I-\mathbb{E}_{\boldsymbol{X}})\varepsilon_{i,\boldsymbol{\theta}}}\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} \geq 0) + \mathbb{E}_{\boldsymbol{X}}\{e^{-2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(I-\mathbb{E}_{\boldsymbol{X}})\varepsilon_{i,\boldsymbol{\theta}}}\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < 0)\Big] \\
&= \prod_{i=1}^n \Big[\big\{e^{-2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(3/2-P_{i,\boldsymbol{\theta}})}(P_{i,\boldsymbol{\theta}} - 1/2) + e^{2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(P_{i,\boldsymbol{\theta}}-1/2)}(3/2 - P_{i,\boldsymbol{\theta}})\big\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} \geq 0) \\
&\qquad + \big\{e^{2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(1/2+P_{i,\boldsymbol{\theta}})}(1/2 - P_{i,\boldsymbol{\theta}}) + e^{2\lambda\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}(P_{i,\boldsymbol{\theta}}-1/2)}(1/2 + P_{i,\boldsymbol{\theta}})\big\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < 0)\Big].
\end{aligned}
$$

Further, using the inequalities $|e^u - 1 - u| \leq \frac{1}{2}u^2e^{u\vee 0}$ and $1 + u \leq e^u$ which hold for all $u \in \mathbb{R}$, the last term above can be bounded by

$$
\begin{aligned}
&\prod_{i=1}^n \Big[\big\{1 + 2\lambda^2(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS})^2(P_{i,\boldsymbol{\theta}} - 1/2)(3/2 - P_{i,\boldsymbol{\theta}})e^{2\lambda|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}|}\big\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} \geq 0) \\
&\qquad + \big\{1 + 2\lambda^2(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS})^2(1/2 - P_{i,\boldsymbol{\theta}})(1/2 + P_{i,\boldsymbol{\theta}})e^{2\lambda|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}|}\big\}I(\boldsymbol{X}_{iS}^{\mathrm{T}}\boldsymbol{\theta} < 0)\Big] \\
&\leq \prod_{i=1}^n \big\{1 + 2\lambda^2(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS})^2|P_{i,\boldsymbol{\theta}} - 1/2|e^{2\lambda|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}|}\big\} \\
&\leq \prod_{i=1}^n \exp\big\{2\lambda^2(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS})^2|P_{i,\boldsymbol{\theta}} - 1/2|e^{2\lambda|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_{iS}|}\big\}.
\end{aligned}
$$

Consequently, for every $\bar{\boldsymbol{\theta}} = (\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}} \in \bar{\Theta}_0(2r)$,

$$
\begin{aligned}
&\log \mathbb{E}_{\boldsymbol{X}} \exp\left\{\lambda\frac{\bar{\boldsymbol{U}}(\bar{\boldsymbol{\theta}}) - \bar{\boldsymbol{U}}(\mathbf{0})}{\|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2}\right\} = \log \mathbb{E}_{\boldsymbol{X}} \exp\left\{\lambda\frac{\boldsymbol{v}^{\mathrm{T}}\{\boldsymbol{\zeta}^S(\boldsymbol{\theta}) - \boldsymbol{\zeta}^S(\mathbf{0})\}}{\|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2}\right\} \\
&\leq \frac{2\lambda^2}{\|\mathbf{D}_0\boldsymbol{v}\|_2^2 + \|\mathbf{D}_0\boldsymbol{\theta}\|_2^2}\sum_{i=1}^n(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{iS})^2|P_{i,\boldsymbol{\theta}} - 1/2|\exp\left(\frac{2\lambda|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{iS}|}{\|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2}\right). \qquad (80)
\end{aligned}
$$

On the event $\mathcal{E}_0(\tau)$ for some $\tau > 0$, we have $|P_{i,\boldsymbol{\theta}} - 1/2| \leq 2A_2\{2nf_\varepsilon(0)\}^{-1/2}\tau^{1/2}r$ and $|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{iS}| \leq \|\mathbf{D}_0\boldsymbol{v}\|_2\|\mathbf{D}_0^{-1}\boldsymbol{X}_{iS}\|_2 \leq \|\mathbf{D}_0\boldsymbol{v}\|_2\{2nf_\varepsilon(0)\}^{-1/2}\tau^{1/2}$. Together with (80), this yields that for all $|\lambda| \leq \{2nf_\varepsilon(0)/\tau\}^{1/2}$,

$$
\log \mathbb{E}_{\boldsymbol{X}} \exp\left\{\lambda\frac{\bar{\boldsymbol{U}}(\bar{\boldsymbol{\theta}}) - \bar{\boldsymbol{U}}(\mathbf{0})}{\|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2}\right\} \leq \frac{\lambda^2}{2}\frac{4e^2A_2\,r}{f_\varepsilon(0)}\sqrt{\frac{\tau}{2nf_\varepsilon(0)}}. \qquad (81)
$$

In view of (81), define

$$w_0(\tau) = 2e\sqrt{\frac{A_2\, r_0}{f_\varepsilon(0)}}\left\{\frac{\tau}{2nf_\varepsilon(0)}\right\}^{1/4} \tag{82}$$

for some $r_0 > 0$ to be specified (see (88) below), such that for any $\bar{\boldsymbol{\theta}} = (\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}} \in \bar{\Theta}_0(2r)$ with $0 \le r \le r_0$,

$$\mathbb{E}_{\boldsymbol{X}} \exp\left\{\frac{\lambda}{w_0(\tau)}\frac{\bar{\boldsymbol{U}}(\bar{\boldsymbol{\theta}}) - \bar{\boldsymbol{U}}(\mathbf{0})}{\|\bar{\mathbf{D}}_0\bar{\boldsymbol{\theta}}\|_2}\right\} \le \exp(\lambda^2/2) \tag{83}$$

holds almost surely on $\mathcal{E}_0(\tau)$ for all

$$|\lambda| \le 2e\sqrt{\frac{A_2\, r_0}{f_\varepsilon(0)}}\left\{\frac{2nf_\varepsilon(0)}{\tau}\right\}^{1/4} := g_0(\tau). \tag{84}$$

By (83), it follows from Corollary 2.2 in the supplement of Spokoiny (2012) and (79) that, for any $\tau > 0$, $0 \le r \le r_0$ and $0 < t \le \frac{1}{2}g_0^2(\tau) - 2s$,

$$\mathbb{P}_{\boldsymbol{X}}\left\{\max_{\boldsymbol{\theta}\in\Theta_0(r)} \|\mathbf{D}_0^{-1}\boldsymbol{U}(\boldsymbol{\theta})\|_2 \ge 6w_0(\tau)(2t+4s)^{1/2}\right\} \le e^{-t} \tag{85}$$

holds almost surely on $\mathcal{E}_0(\tau)$, where g_0 is given at (84).

Combining (74) and (85) we obtain that for any $\tau > 0$, $0 \le r \le r_0 \le \{2nf_\varepsilon(0)/\tau\}^{1/2}$ and $0 < t \le \frac{1}{2}g_0^2(\tau) - 2s$,

$$\mathbb{P}_{\boldsymbol{X}}\left\{\max_{\boldsymbol{\theta}\in\Theta_0(r)} \|\chi_1^S(\boldsymbol{\theta})\|_2 \ge \delta(\tau, r)r + 6w_0(\tau)(2t+4s)^{1/2}\right\} \le e^{-t} \tag{86}$$

almost surely on $\mathcal{E}_0(\tau)$. For a given triplet (τ, r, t), define the event

$$\Omega_0^S(\tau, r, t) = \left\{\max_{\boldsymbol{\theta}\in\Theta_0(r)} \|\chi_1^S(\boldsymbol{\theta})\|_2 \le \delta(\tau, r)r + 6w_0(\tau)(2t+4s)^{1/2}\right\}. \tag{87}$$

Step 2: Fisher approximation. By Lemma 13,

$$\begin{aligned}&\max_{S\subseteq[p]:|S|=s} \|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2 \\ &= \{2nf_\varepsilon(0)\}^{1/2} \max_{S\subseteq[p]:|S|=s} \|\widehat{\boldsymbol{\Sigma}}_{SS}^{1/2}\widehat{\boldsymbol{\theta}}_S\|_2 \le C_1\, a_2^{-1}\{2f_\varepsilon(0)s\log(pn)\}^{1/2} := r_0\end{aligned} \tag{88}$$

holds with probability at least $1 - c_1n^{-1} - c_2n^{1-\kappa}$. Moreover, since $\widehat{\boldsymbol{\theta}}_S$ maximizes $\mathcal{L}_n^S(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \mathbb{R}^s$ for each s-subset $S \subseteq [p]$, we have $\nabla\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) = \mathbf{0}$ and $\chi_1^S(\widehat{\boldsymbol{\theta}}) = \mathbf{D}_0\widehat{\boldsymbol{\theta}}_S - \widehat{\boldsymbol{\xi}}_S$. This, together with (87) implies that on the event $\{\widehat{\boldsymbol{\theta}}_S \in \Theta_0(r_0)\} \cap \Omega_0^S(\tau, r_0, t)$,

$$\|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S - \widehat{\boldsymbol{\xi}}_S\|_2 \le \delta(\tau, r_0)r_0 + 6w_0(\tau)(2t+4s)^{1/2} \tag{89}$$

whenever $n \ge \{2f_\varepsilon(0)\}^{-1}\tau r_0^2$.

Step 3: Wilks approximation. For $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_0(r)$, define

$$\chi_2^S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathcal{L}_n^S(\boldsymbol{\theta}) - \mathcal{L}_n^S(\boldsymbol{\theta}_2) - (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^{\mathrm{T}} \nabla \mathcal{L}_n^S(\boldsymbol{\theta}_2) + \frac{1}{2}\|\mathbf{D}_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2^2. \tag{90}$$

Noting that $\nabla_{\boldsymbol{\theta}_1} \chi_2^S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \nabla \mathcal{L}_n^S(\boldsymbol{\theta}_1) - \nabla \mathcal{L}_n^S(\boldsymbol{\theta}_2) + \mathbf{D}_0^2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) = \mathbf{D}_0\{\chi_1^S(\boldsymbol{\theta}_1) - \chi_1^S(\boldsymbol{\theta}_2)\}$, we have

$$|\chi_2^S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| = |\chi_2^S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - \chi_2^S(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2)| \le 2\|\mathbf{D}_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2 \max_{\boldsymbol{u} \in \Theta_0(r)} \|\chi_1^S(\boldsymbol{u})\|_2, \tag{91}$$

where $\widetilde{\boldsymbol{\theta}} = \lambda \boldsymbol{\theta}$ for some $0 \le \lambda \le 1$. Let $r_0 > 0$ be as in (88). Then, it follows from (91) that on $\Omega_0^S(\tau, r_0, t)$ with $n \ge \{2f_\varepsilon(0)\}^{-1}\tau r_0^2$,

$$\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_0(r_0)} \frac{|\chi_2^S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|}{\|\mathbf{D}_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2} \le 2\delta(\tau, r_0) r_0 + 12 w_0(\tau)(2t + 4s)^{1/2}.$$

In view of (90), $\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0}) - \frac{1}{2}\|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2^2 = -\chi_2^S(\mathbf{0}, \widehat{\boldsymbol{\theta}}_S)$. Therefore, on the event $\{\widehat{\boldsymbol{\theta}}_S \in \Theta_0(r_0)\} \cap \Omega_0^S(\tau, r_0, t)$ we have

$$\begin{aligned}
&\Big|[2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0})\}]^{1/2} - \|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2\Big| \\
&\le \frac{|2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0})\} - \|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2^2|}{\|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2} \le \frac{2|\chi_2^S(\mathbf{0}, \widehat{\boldsymbol{\theta}}_S)|}{\|\mathbf{D}_0\widehat{\boldsymbol{\theta}}_S\|_2} \le 4\big\{\delta(\tau, r_0) r_0 + 6 w_0(\tau)(2t + 4s)^{1/2}\big\},
\end{aligned}$$

provided that $n \ge \{2f_\varepsilon(0)\}^{-1}\tau r_0^2$. Together with (89), this implies that conditional on the event $\cap_{S \subseteq [p]: |S| = s}\{\widehat{\boldsymbol{\theta}}_S \in \Theta_0(r_0)\} \cap \Omega_0^S(\tau, r_0, t)$,

$$\max_{S \subseteq [p]: |S| = s} \Big|[2\{\mathcal{L}_n^S(\widehat{\boldsymbol{\theta}}_S) - \mathcal{L}_n^S(\mathbf{0})\}]^{1/2} - \|\widehat{\boldsymbol{\xi}}_S\|_2\Big| \le 5\big\{\delta(\tau, r_0) r_0 + 6 w_0(\tau)(2t + 4s)^{1/2}\big\} \tag{92}$$

whenever $n \ge \{2f_\varepsilon(0)\}^{-1} r_0^2 \tau$, where $\delta(\tau, r)$, r_0 and $w_0(\tau)$ are as in (76), (88) and (82).

Finally, taking $\tau = \tau_0 \asymp \lambda_{\min}^{-1}(s) s \log(pn)$ as in (36) and setting $t = s \log \frac{ep}{s} + \log n$ in the concentration bound (86) prove (48) using Boole's inequality. ■