Concentration Based Inference for High Dimensional (Generalized) Regression Models: New Phenomena in Hypothesis Testing

Zhu, Ying

Purdue University

17 August 2018

Concentration Based Inference for High Dimensional (Generalized) Regression Models: New Phenomena in Hypothesis Testing

Ying Zhu*

September 30, 2018

Abstract

We develop simple and non-asymptotically justified methods for hypothesis testing about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional (generalized) regression models where p can exceed the sample size n. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, we consider $H_0 : h(\theta^*) = \mathbf{0}_m$ against the alternative hypothesis $H_1 : h(\theta^*) \neq \mathbf{0}_m$, where m can be as large as p and h can be nonlinear in θ^*. Our test statistics is based on the sample score vector evaluated at an estimate $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. We provide nonasymptotic control on the Type I and Type II errors for the score test. In addition, confidence regions are constructed in terms of the score vectors. By exploiting the concentration phenomenon in Lipschitz functions, the key component reflecting the "dimension complexity" in our non-asymptotic thresholds uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies between the coordinates. The novelty of our methods is that their validity does not rely on good behavior of $\left\|\hat{\theta}_\alpha - \theta^*\right\|_2$ or even $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ nonasymptotically or asymptotically. Most interestingly, we discover phenomena that are opposite from the existing literature: (1) More restrictions (larger m) in H_0 make our procedures more powerful; (2) whether θ^* is sparse or not, it is possible for our procedures to detect alternatives with probability at least $1 -$ Type II error when $p \geq n$ and $m > p - n$; (3) the coverage probability of our procedures is not affected by how sparse θ^* is. The proposed procedures are evaluated with simulation studies, where the empirical evidence supports our key insights.

*Email: yingzhu@purdue.edu. Assistant Professor of Statistics and Computer Science. Purdue University. West Lafayette, Indiana. A start-up fund from Purdue University partially supported this research. An earlier draft of this manuscript was prepared during my appointment at Michigan State University (Department of Economics, Social Science Data Analytics Initiative) that also provided financial support.

1 Introduction

A common feature of the existing procedures that are deemed "practical" for inference of high dimensional regression coefficients is that they all hinge on asymptotic validity to some extent. This occurrence is perhaps not coincidental as asymptotic analysis often allows one to focus on the "leading" term(s) by assuming the "remainder" terms approach to zero faster, which can be quite convenient for determining the threshold in a test. However, many real-world applications (in psychology, for example) have a limited sample size which renders any asymptotic argument questionable.

Our primary goal is to find situations where effective non-asymptotic methods can be developed for hypothesis testing about the coefficients in high dimensional regression models. We illustrate the key insight with the linear regression model

$$Y_i = X_i\theta^* + W_i, \qquad i = 1, ..., n, \tag{1}$$

where $W = \{W_i\}_{i=1}^n \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ and $\mathbf{0}_n$ denotes an $n-$dimensional vector of zeros; θ^* is a $p-$dimensional vector of unknown coefficients and p is allowed to exceed the sample size n; $Y = \{Y_i\}_{i=1}^n$ is an $n-$dimensional vector of responses; $X = \{X_i\}_{i=1}^n \in \mathbb{R}^{n\times p}$ is the design matrix with the ith row specified by X_i. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, let

$$H_0 : h(\theta^*) = \mathbf{0}_m \; vs. \; H_1 : h(\theta^*) \neq \mathbf{0}_m,$$

where m can be as large as p and h can be nonlinear in θ^*. Relative to existing literature, we will look at these broader forms of hypotheses and the impact of m, the number of restrictions in the null hypothesis. By making simple changes in the notations, we can also test $H_0 : h(\theta^*) \leq \mathbf{0}_m$ or $H_0 : h(\theta^*) \geq \mathbf{0}_m$ using the procedures and analysis developed later in the paper.

Our secondary goal is to seek some general nonasymptotic theory for inference in high dimensional models that involve non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients (including the binary response models and certain nonlinear regressions). Throughout the paper, we make our argument by conditioning on X; in addition, we assume $\{\theta \in \mathbb{R}^p : h(\theta) = \mathbf{0}_m\} \neq \emptyset$ and H_0 does not contain any redundant restrictions.

This work is initially inspired by an important problem from intervention studies – testing for heterogeneity in treatment effects. Suppose V_i is a binary variable which equals 1 if individual i receives treatment and 0 otherwise; Z_i is a $p-$dimensional vector of covariates such that $\mathbb{E}(Z_i) = \mathbf{0}_p$ (this zero-mean condition can be relaxed but is assumed here to lighten the notations). We use Y_i^A to denote the (potential) outcome upon receiving treatment, Y_i^B to denote the (potential) outcome without treatment, and Y_i to denote the observed outcome; note that $Y_i = (1-V_i)Y_i^B + V_iY_i^A$.

A commonly studied model (see, e.g., [19]) takes the form

$$Y_i = \pi_0^* + \pi_1^* V_i + \sum_{j=1}^{p} \gamma_j^* V_i Z_{ij} + \sum_{j=1}^{p} \alpha_j^* Z_{ij} + W_i \tag{2}$$

where

$$TE(Z_i) := \mathbb{E}\left(Y_i^A - Y_i^B | Z_i\right) = \pi_1^* + \sum_{j=1}^{p} \gamma_j^* Z_{ij}, \tag{3}$$

$$ATE := \mathbb{E}\left(Y_i^A - Y_i^B\right) = \pi_1^*. \tag{4}$$

The heterogeneity in the treatment effect $TE(Z_i)$ corresponds to $\sum_{j=1}^{p} \gamma_j^* Z_{ij}$. Taking the expectation of $TE(Z_i)$ in (3) over Z_i gives (4), referred to as the Average Treatment Effect (ATE). We are often interested in testing

$$H_0 : \gamma_j^* = \gamma_j^0 \ \forall j \in \{1, 2, ..., p\}. \tag{5}$$

Such a hypothesis can be handled by the methods developed in this paper since it is a special case of our H_0. Note that when $\gamma_j^0 = 0$ for all j, the hypothesis above implies there is no heterogeneity in the treatment effect.

Before this paper, some tests have been proposed in the literature of high dimensional inference. For example, [7] establish asymptotic consistency for testing

$$H_{0,G} : \theta_j^* = 0 \ \forall j \in G \subseteq \{1, 2, ..., p\} \tag{6}$$

in $Y_i = X_i\theta^* + W_i$, where they require $\log(|G|) = o(n^{1/7})$ and the sparsity parameter s_0 of θ^* to satisfy $n^{-1}(s_0 \log p)^2 \log(|G|) = o(1)$; [23] allow $G = \{1, 2, ..., p\}$ but require $n^{-1}(\log(pn))^7 = o(1)$ and $n^{-1}(s_0 \log p)^2 \log p = o(1)$ (which essentially restricts $|G|$ through p). [23] note that the smaller $|G|$ gets, the more powerful their procedure becomes (see equation (13) in [23]); furthermore, their simulation results suggest that the coverage probability decreases as θ^* gets less sparse.

In our view, the aforementioned findings are counterintuitive: First, more restrictions (larger $|G|$) on θ^* in H_0 result in fewer parameters to be "determined" and thus should only make the testing problem easier; second, if $|G|$ is large enough, the power of a test should not rely on whether θ^* is sparse or not. With these questions in mind, we offer a new testing method and statistical analysis, which does not require the conditions mentioned in the previous paragraph and works for any finite (n, p). We reveal phenomena that are opposite from the existing literature: (1) More restrictions (larger m) in H_0 make our procedures more powerful; (3) whether θ^* is sparse or not, it is possible for our procedures to detect alternatives with probability at least $1 -$ Type II error when $p \geq n$ and $m > p - n$; (3) the coverage probability of our procedures is not affected by how sparse θ^* is.

As suggested by the title, this paper studies nonasymptotic inference by exploiting the sharp concentration phenomenon in Lipschitz functions, which should be

distinguished from another line of literature based on normal approximations using the Stein's Method, for example, [6] and [11] (also see [23], whose method is justified by the theory in [6]). In particular, [11] studies similar models (as this paper) and develops results for hypothesis testing in the regime of $n \gg p$; by contrast, our focus is on the regime of $p \geq n$ and possibly $p \gg n$. In [11], some of the results are still only asymptotically valid and the other results (even though nonasymptotically justified) come with probabilistic guarantees that contain rather loose constants and dimension-dependent components.

For the mean of a high-dimensional random vector, [1] study bootstrap confidence regions with the concentration approach. Beyond the inference for the mean of a high-dimensional random vector, is it possible to adapt a concentration approach for testing about the coefficients in a high-dimensional regression problem? At first glance, there seems no lack of non-asymptotic bounds on the l_p-error (often $p \in [1, 2]$ or $p = \infty$) of some (regularized) estimator concerning (1). However, these bounds (even in the sharpest forms) tend to involve quite a few unknown nuisance parameters that are hard to estimate in practice. In order to adapt the existing bounds for the purpose of inference, prior knowledge on the sparsity of θ^* would be needed at a minimum; see, e.g., [10].

For this reason, we choose our test statistics to base on the sample score vector evaluated at $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. By definition, the resulting procedure is a score test. Our test statistics take the form

$$\Psi_q(\hat{\theta}_\alpha) := \left\| \frac{1}{n} X^T \left(Y - X\hat{\theta}_\alpha \right) \right\|_q, \tag{7}$$

where $\hat{\theta}_\alpha$ is obtained by solving the following program:

$$\begin{aligned} \left(\hat{\theta}_\alpha, \, \hat{\mu}_\alpha \right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}^p} & \left\| \mu_\alpha \right\|_{\tilde{q}} \\ \text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^n X_i \left(Y_i - X_i \theta_\alpha \right) - \mu_\alpha \right\|_q & \leq r_{\alpha,q}, \\ h(\theta_\alpha) & = \mathbf{0}_m, \end{aligned} \tag{8}$$

with $q, \, \tilde{q} \in [1, \infty]$ chosen by the users. For $1 \leq q \leq \infty$, we write $\|v\|_q$ to mean the l_q-norm of a $k-$dimensional vector v, where $\|v\|_q := \left(\sum_{i=1}^k |v_i|^q \right)^{1/q}$ when $1 \leq q < \infty$ and $\|v\|_q := \max_{i=1,\ldots,k} |v_i|$ when $q = \infty$. The choice for $r_{\alpha,q}$ in the first constraint is to be specified in the subsequent section.

We can also work with an alternative formulation:

$$\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha,\mu_\alpha)\in\mathbb{R}^p\times\mathbb{R}} \mu_\alpha$$
$$\text{subject to: } \left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\theta_\alpha\right)\right\|_q \leq r_{\alpha,q} + \mu_\alpha, \tag{9}$$
$$h(\theta_\alpha) = \mathbf{0}_m,$$
$$\mu_\alpha \geq 0.$$

Throughout this paper, we will slightly abuse the notations as in the above, where $\hat{\mu}_\alpha$ (also μ_α) in (8) is a vector and in (9) is a scalar. In addition, we suppress the dependence of $(\hat{\theta}_\alpha, \hat{\mu}_\alpha)$ in (8) on $(q,\, \tilde{q})$ and the dependence of $(\hat{\theta}_\alpha, \hat{\mu}_\alpha)$ in (9) on q for notational simplicity.

A solution $\hat{\theta}_\alpha$ to either (8) or (9) may not necessarily be unique: that is, there might be different $\hat{\theta}_\alpha$s that satisfy (8) (or (9)) while delivering the same (minimal) objective value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$). We refer to the vector μ_α in (8) (and the scalar μ_α in (9)) as the "slack" vector (respectively, the "slack" variable) that fills the "gap" between $\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\theta^*\right)\right\|_q$ and $\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\theta_\alpha\right)\right\|_q$ where $h(\theta_\alpha) = \mathbf{0}_m$. When the null hypothesis is true, i.e., $h(\theta^*) = \mathbf{0}_m$, the optimal value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$) must be zero with probability at least $1-\alpha$. This fact does not imply that $\hat{\theta}_\alpha$ would necessarily be "close" to θ^* under H_0, but rather,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\hat{\theta}_\alpha\right)\right\|_q \leq r_{\alpha,q}, \qquad \text{(under } H_0\text{)}$$

with the same probability guarantee $1-\alpha$ for the event

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\theta^*\right)\right\|_q \leq r_{\alpha,q}.$$

In the paper, we establish statistical guarantees (stated in terms of $(\alpha, \tilde{q}, q)$) for (8), and statistical guarantees (stated in terms of (α, q)) for (9).

To compare (8) with (9) from the computational perspective, we let $\mathcal{F}_1^\alpha$ denote the set of $(\theta_\alpha, \mu_\alpha)$ that are feasible for (8) and $\mathcal{F}_{1,\theta}^\alpha$ denote the set of θ_α from $\mathcal{F}_1^\alpha$; similarly, $\mathcal{F}_2^\alpha$ and $\mathcal{F}_{2,\theta}^\alpha$ are defined with regard to (9). Note that an element $(\tilde{\theta}_\alpha, \tilde{\mu}_\alpha)$ in $\mathcal{F}_1^\alpha$ implies

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\tilde{\theta}_\alpha\right)\right\|_q \leq r_{\alpha,q} + \|\tilde{\mu}_\alpha\|_q\,;$$

that is, $(\tilde{\theta}_\alpha, \|\tilde{\mu}_\alpha\|_q) \in \mathcal{F}_2^\alpha$. Consequently, $\mathcal{F}_{1,\theta}^\alpha \subseteq \mathcal{F}_{2,\theta}^\alpha$. On the other hand, the objective function in (8) is minimized over a $p-$dimensional vector as opposed to a scalar in (9). However, (8) does not require the entries in the slack vector to be

positive while (9) require the slack variable to be positive. These facts suggest that the choice between (8) and (9) incurs some trade-offs in terms of computational cost.

Compared to basing the test statistics on a consistent estimator for θ^*, such as the existing Lasso estimators, Dantzig selectors, or the new variant (10) with $\tilde{q} = 1$ and $q = \infty$ (to be discussed later), the score statistics (7) using $\hat{\theta}_\alpha$ from (8) or (9) allow us to bypass the sparsity assumption on θ^* and the inherent challenges in an inverse problem. As a consequence, our thresholds or confidence regions do not involve unknown parameters related to sparsity.

In terms of relaxing sparsity assumptions, this paper shares slight similarity as [24] although our method is drastically different from what is proposed in [24]. Also, [24] deal with $H_0 : a^T\theta^* = b^0$ for some prespecified $a \in \mathbb{R}^p$ and $b^0 \in \mathbb{R}$ while the form of our null hypothesis is much more general and can impose up to p restrictions on θ^*; moreover, the statistical guarantees in [24] are asymptotic while our procedures are nonasymptotically valid and found to work well for small n (such as 15) in simulations; finally, [24] show that their test can attain certain optimality in detecting alternatives as long as the sparsity parameters of θ^* and a are in the order $o\left(\frac{\sqrt{n}}{\log p}\right)$, while we find the power of our tests depends on the number of restrictions in H_0 (whether θ^* is sparse or not).

If we choose $q = \infty$, then (7) is reduced to

$$\Psi_\infty\left(\hat{\theta}_\alpha\right) := \left\| \frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i\hat{\theta}_\alpha\right)\right\|_\infty .$$

This statistics shares some resemblance to the score-based correction term in the debiased Lasso literature (see, e.g., [7, 12, 17, 22, 23]) as well as the decorrelated score in [15]. Unlike the debiased and decorrelated procedures which require an initial (consistent) estimator for (the sparse) θ^* in the correction term, our $\hat{\theta}_\alpha$ here need not be consistent and is directly used in the test statistics (requiring no further debiasing or decorrelating step). In addition, our methods are nonasymptotically valid and do not require θ^* to be sparse, whereas the aforementioned papers hinge on the asymptotic normality of the debiased or decorrelated procedure and require θ^* to be sufficiently sparse.

We derive implementable (non-asymptotic) thresholds $r_{\alpha,q}$ such that

$$\begin{aligned} \mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}\right\} &\leq \alpha, \qquad \text{(Type I Error)} \\ \mathbb{P}_1\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}\right\} &\leq \beta, \qquad \text{(Type II Error)} \end{aligned}$$

where $\mathbb{P}_0$ means under H_0, $\mathbb{P}_1$ means under H_1 and a "Level$-\beta$ Separation Requirement" imposed upon the l_q-distances between the population score vectors evaluated at θ^* and θ_αs satisfying $h(\theta_\alpha) = \mathbf{0}_m$. Our decision rule is that if $\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}$, we reject the null hypothesis H_0. In addition to the guarantees on the Type I and Type II errors, we also construct confidence regions in terms of the score vectors.

Our non-asymptotic thresholds $r_{\alpha,q}$ consist of data-driven components which reflect the "dimension complexity", as well as components which are free of p. This form is a direct result of the concentration phenomenon in Lipschitz functions. The key data-driven component in our $r_{\alpha,q}$ uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies across coordinates. These facts put our framework in sharp contrast with the Bonferroni approach used in the estimation literature (e.g., [10]). In this perspective, our results share some similarity as those in [1] except that [1] concern inference for the mean of a random vector while we consider inference about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional regression models.

Beyond the context of hypothesis testing, as a secondary contribution, the data-driven approach proposed in this paper for setting the thresholds $r_{a,q}$ also suggests a new class of regularized estimators:

$$\hat{\theta}_\alpha^{new} \in \arg\min_{\theta_\alpha \in \mathbb{R}^p} \|\theta_\alpha\|_{\tilde{q}} \ \text{ subject to } \left\| \frac{1}{n} X^T (Y - X\theta_\alpha) \right\|_q \leq r_{\alpha,q}. \tag{10}$$

When $\tilde{q} = 1$ and $q = \infty$, (10) can be viewed as a variant of the Dantzig selector, for which we establish a complementary l_2-error bound. In contrast to (10), (8) and (9) involve a second constraint $h(\theta_\alpha) = \mathbf{0}_m$ and a slack vector (or variable) μ_α in the first constraint, as well as a different objective function (minimizing the $l_{\tilde{q}}-$norm of the slack vector or minimizing the slack variable, instead of minimizing $\|\theta_\alpha\|_{\tilde{q}}$). Consequently, the resulting solution to (10) is not constrained to satisfy $h(\hat{\theta}_\alpha^{new}) = \mathbf{0}_m$, whereas $\hat{\theta}_\alpha$s in (8) and (9) satisfy $h(\hat{\theta}_\alpha) = \mathbf{0}_m$.

The rest of the paper is organized as follows. In Section 2, we focus on the Gaussian regression models and establish nonasymptotic control on the Type I and Type II errors for the proposed score test. Implementations for some natural choices of q (relevant to both (8) and (9)) and $\tilde{q}$ (relevant to (8)) are also discussed.

We demonstrate numerical evidence through simulation studies in Section 3, where the computational performance of (8) and (9) as well as different choices of ($\tilde{q}$, q) in (8) and q in (9) are also compared. We look at a "small sample" setup ($n = 15$, $p = 50$) and a "larger sample" setup ($n = 100$, $p = 300$). Our designs range from highly dense θ^* to highly sparse θ^* and our null hypotheses take either the form (6) or $H_0 : A\theta^* = \mathbf{0}_m$, for some prespecified $A \in \mathbb{R}^{m \times p}$ and $m \in \{p, p-3, p-9\}$. The second form of hypotheses is motivated by real world applications in marketing and more detail is described in Section 3. To the best of our knowledge, this paper is the first that studies $H_0 : A\theta^* = \mathbf{0}_m$ with "large" m, which cannot be handled by existing approaches in the literature.

The remaining sections are about various extensions. Section 4 provides some general nonasymptotic justifications for inference in high dimensional models that involve non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients (including the binary response models and certain nonlinear regressions). Motivated by the data-driven feature of our concentration approach,

Section 5 proposes a new class of regularized estimators along with a complementary l_2-error bound. Section 6 concludes the paper and all technical details are deferred to the supplementary materials.

2 Gaussian Regressions

For the linear regression model (1), we first consider the scenario where σ^2 is known, and then look at the scenario where σ^2 is not known *a priori*. Throughout this section, we use $\mathbb{E}_W[\cdot]$ to denote the expectation over W only, conditioning on X.

By considering the concentration of $\left\|\frac{1}{n}X^TW\right\|_q$ around $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$, our first result establishes an "ideal" confidence region for the l_q-distance between the score vectors evaluated at θ^* and a "theoretical" optimal solution, $\hat{\theta}^*_\alpha$; that is,

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}^*_\alpha)\right\|_q = \left\|\left[\frac{1}{n}X^T(Y-X\hat{\theta}^*_\alpha)\right]-\left[\frac{1}{n}X^T(Y-X\theta^*)\right]\right\|_q.$$

This "theoretical" optimal solution above, $\hat{\theta}^*_\alpha$, is obtained by setting $r_{\alpha,q}$ in (8) (and (9)) to $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ plus a deviation. In practice, $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ may be bounded with its Monte Carlo approximation and a "small" remainder term. This approach results in a "practical" optimal solution, $\hat{\theta}_\alpha$, which can then be used to construct test statistics and a "practical" confidence region.

To state the first result, we introduce the following notation (which will appear in many places throughout this paper):

$$\begin{aligned}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q &= \sqrt[q]{\sum_{j=1}^p\left(\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2}\right)^q}, \qquad q\in[1,\infty)\\ \left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q &= \max_{j\in\{1,\ldots,p\}}\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2}, \qquad q=\infty.\end{aligned}$$

Proposition 2.1. *Assume (1) where $W\sim\mathcal{N}(\mathbf{0}_n,\,\sigma^2\boldsymbol{I}_n)$ and is independent of X. Then for any $q\in[1,\,\infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+t\right\}\leq\exp\left(\frac{-nt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{11}$$

Moreover, for $\alpha\in(0,\,1)$, let

$$r_{\alpha,q}=r^*_{\alpha,q}:=\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+\sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha}} \tag{12}$$

in (8) (or (9)). Then, an optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ *to (8) must satisfy*

$$\left\|\frac{1}{n}X^TX(\theta^* - \hat{\theta}^*_\alpha)\right\|_{\tilde{q}} \;\geq\; \|\hat{\mu}^*_\alpha\|_{\tilde{q}}\,, \tag{13}$$

$$\left\|\frac{1}{n}X^TX(\theta^* - \hat{\theta}^*_\alpha) - \hat{\mu}^*_\alpha\right\|_{q} \;\leq\; 2r^*_{\alpha,q}, \tag{14}$$

with probability at least $1-\alpha$*. Similarly, an optimal solution* $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ *to (9) must satisfy*

$$\left\|\frac{1}{n}X^TX(\theta^* - \hat{\theta}^*_\alpha)\right\|_{q} \;\geq\; \hat{\mu}^*_\alpha, \tag{15}$$

$$\left\|\frac{1}{n}X^TX(\theta^* - \hat{\theta}^*_\alpha)\right\|_{q} \;\leq\; 2r^*_{\alpha,q} + \hat{\mu}^*_\alpha, \tag{16}$$

with probability at least $1-\alpha$*.*

2.1 Hypothesis Testing

For the moment, suppose we set $r_{\alpha,q} = r^*_{\alpha,q}$ in (8) (or (9)) according to (12) as in Proposition 2.1. Under H_0, $(\theta^*,\, \mathbf{0}_p)$ $((\theta^*,\, 0))$ is an optimal solution to (8) (respectively, (9)). Consequently, given the test statistics (7) and a chosen $\alpha \in (0,\, 1)$, an optimal solution to (8) (and (9)) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}^*_\alpha) \geq r^*_{\alpha,q}\right\} \leq \alpha \tag{17}$$

where $\mathbb{P}_0$ means under H_0.

The claim in (17) suggests a test (with level α) based on the statistics $\Psi_q(\hat{\theta}^*_\alpha)$ and an "ideal" critical value, $r^*_{\alpha,q}$, given in (12). When $W \sim \mathcal{N}(\mathbf{0}_n,\, \sigma^2\boldsymbol{I}_n)$ and σ^2 is known, the first term $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ in $r^*_{\alpha,q}$ can be approximated by Monte-Carlo as follows. Let $Z \in \mathbb{R}^{n\times R}$ be a matrix consisting of independent entries randomly drawn from $\mathcal{N}(0,\, 1)$ and the rth column of Z is denoted by Z_r. By (75) and (76), note that $\sigma R^{-1}\sum_{r=1}^R \left\|\frac{1}{n}X^TZ_r\right\|_q - \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ is sub-Gaussian with parameter at most $(nR)^{-1/2}\sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q$. Consequently, (73) yields the following concentration

$$\mathbb{P}\left\{\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] \geq \frac{\sigma}{R}\sum_{r=1}^R\left\|\frac{1}{n}X^TZ_r\right\|_q + t\right\} \leq \exp\left(\frac{-nRt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{18}$$

Combining (11) and (18) yields

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q + t_1 + t_2\right\}$$
$$\leq \exp\left(\frac{-nt_1^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right) + \exp\left(\frac{-nRt_2^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \quad (19)$$

2.1.1 Construction of Critical Values ($r_{\alpha,q}$) and Type I Error

For some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$, we let in (19),

$$t_1 = \sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}} := \tau_{\alpha_1,q}, \quad (20)$$
$$t_2 = \sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\sqrt{\frac{2}{nR}\log\frac{1}{\alpha_2}} := \sqrt{\frac{1}{R}}\tau_{\alpha_2,q}.$$

Based on (19) along with the choices of t_1 and t_2 above, the RHS of the first constraint in (8) (or (9)) is set to

$$r_{\alpha,q} = \frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q}. \quad (21)$$

Note that we can draw as many columns in Z as we want, to make $\sqrt{\frac{1}{R}}\tau_{\alpha_2,q}$ in (21) small; for a given α, we can let α_2 be smaller than α_1 because of the additional "$\sqrt{\frac{1}{R}}$".

Under H_0, $(\theta^*,\, \mathbf{0}_p)$ $((\theta^*,\, 0))$ is an optimal solution to (8) (respectively, (9)) with $r_{\alpha,q}$ specified in (21). Consequently, a (practical) optimal solution to (8) (and (9)) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}\right\} \leq \alpha \qquad \textit{(Type I Error).} \quad (22)$$

Remarks. In terms of control on the Type I error, the l_q-norm in (7), (8) and (9) can be generalized to the function $\zeta_q : \mathbb{R}^p \mapsto \mathbb{R}$ that satisfies:

- for all $z \in \mathbb{R}^p$ and $a \in \mathbb{R}^+$, $\zeta_q(az) = a\zeta_q(z)$,
- for all $z,\, z' \in \mathbb{R}^p$, $\zeta_q\left(z + z'\right) \leq \zeta_q(z) + \zeta_q(z')$,
- for all $z \in \mathbb{R}^p$, $|\zeta_q(z)| \leq \|z\|_q$ for $q \in [1,\, \infty]$.

In this case, we simply let

$$r_{\alpha,q} = \frac{\sigma}{R}\sum_{r=1}^{R}\zeta_q\left(\frac{1}{n}X^T Z_r\right) + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q},$$

and obtain

$$\mathbb{P}_0\left\{\zeta_q\left(\frac{1}{n}X^T\left(Y - X\hat{\theta}_\alpha\right)\right) \geq r_{\alpha,q}\right\} \leq \alpha \qquad \text{(Type I Error)},$$

where $\hat{\theta}_\alpha$ is a solution to (8) (or (9)) with the l_q-norm in the first constraint replaced by ζ_q. Given ζ_q is subadditive and bounded by the l_q-norm, the results above follow from the simple fact that

$$\begin{aligned}\left|\zeta_q\left(\frac{1}{n}X^T W\right) - \zeta_q\left(\frac{1}{n}X^T W'\right)\right| &\leq \left\|\frac{1}{n}X^T\left(W - W'\right)\right\|_q \\ &\leq \frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q \left\|W - W'\right\|_2.\end{aligned}$$

Consequently, we can establish bounds that are identical to (11), (18), (19) in terms of $\zeta_q\left(\frac{1}{n}X^T W\right)$, $\mathbb{E}_W\left[\zeta_q\left(\frac{1}{n}X^T W\right)\right]$, $\frac{\sigma}{R}\sum_{r=1}^{R}\zeta_q\left(\frac{1}{n}X^T Z_r\right)$, and then follow the same argument as what is used to show (22).

2.1.2 Practical Confidence Regions

Let $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ be an optimal solution to (8) with $r_{\alpha,q}$ specified in (21). Our previous analysis implies that

$$\begin{aligned}&\left\|\frac{1}{n}X^T X(\theta^* - \hat{\theta}_\alpha) - \hat{\mu}_\alpha\right\|_q \\ \leq\ &\left\|\frac{1}{n}X^T(Y - X\hat{\theta}_\alpha) - \hat{\mu}_\alpha\right\|_q + \left\|\frac{1}{n}X^T W\right\|_q \\ \leq\ &\frac{2\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^T Z_r\right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}}\tau_{\alpha_2,q} \qquad (23)\end{aligned}$$

and

$$\left\|\frac{1}{n}X^T X(\theta^* - \hat{\theta}_\alpha)\right\|_{\tilde{q}} \geq \|\hat{\mu}_\alpha\|_{\tilde{q}} \qquad (24)$$

with probability at least $1-\alpha$; similarly, in terms of (9), we have

$$\begin{aligned} & \left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha)\right\|_q - \hat{\mu}_\alpha \\ \leq & \left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha)\right\|_q - \hat{\mu}_\alpha + \left\|\frac{1}{n}X^TW\right\|_q \\ \leq & \frac{2\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}}\tau_{\alpha_2,q} \end{aligned} \tag{25}$$

and

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha)\right\|_q \geq \hat{\mu}_\alpha \tag{26}$$

with probability at least $1-\alpha$. The argument for (24) and (26) is identical to what is used to show (13) and (15). As we have pointed out in the introduction, there might be different $\hat{\theta}_\alpha$s that satisfy (8) (or (9)) while producing the same (minimal) objective value $\|\hat{\mu}_\alpha\|_{\tilde{q}}$ (respectively, $\hat{\mu}_\alpha$). Consequently, there is more than one confidence region in the form of (23)-(24) or (25)-(26). In view of (25)-(26), the length of the confidence interval is naturally

$$CI-Length = \frac{2\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}}\tau_{\alpha_2,q}. \tag{27}$$

If $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ can be known exactly and we were able to set $r_{\alpha_1,q} = r^*_{\alpha_1,q}$ in (8) (or (9)) as in Proposition 2.1, then any resulting optimal solution $\left(\hat{\theta}^*_\alpha, \hat{\mu}^*_\alpha\right)$ to (8) (respectively, (9)) should satisfy

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}^*_\alpha) - \hat{\mu}^*_\alpha\right\|_q \leq 2\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + 2\tau_{\alpha_1,q}, \tag{28}$$

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}^*_\alpha)\right\|_q - \hat{\mu}^*_\alpha \leq 2\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + 2\tau_{\alpha_1,q}, \tag{29}$$

both with probability at least $1-\alpha_1$. Comparing (23) with (28) and (25) with (29), note that the difference in terms of the right hand sides is

$$2\left(\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q - \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]\right) + 2\sqrt{\frac{1}{R}}\tau_{\alpha_2,q},$$

which can be made arbitrarily small with a large number of random draws in the Monte-Carlo approximation. Because of such an approximation, the probabilistic guarantees for (23) and (25) are bounded from below by $1-\alpha$ instead of $1-\alpha_1$.

Given the statistics $\Psi_q(\hat{\theta}_\alpha)$ in (7) based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (21), we have constructed a test with level α as shown in (22). For

some $\beta \in (0,\, 1)$, when can this test correctly detect an alternative with probability at least $1-\beta$? To answer this question, we introduce the "Separation Requirement" in the following section.

2.1.3 Separation Requirement and Type II Error

Letting $\Theta_0 := \{\theta \in \mathbb{R}^p \,:\, h(\theta) = \mathbf{0}_m\}$, we choose $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, and assume

$$\inf_{\theta \in \Theta_0} \left\| \frac{1}{n} X^T X(\theta^* - \theta) \right\|_q \geq \delta_{\alpha,\beta,q} \tag{30}$$

with

$$\delta_{\alpha,\beta,q} = 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}} \tau_{\alpha_2,q} + \sqrt{\frac{1}{R}} \tau_{\beta_1,q} + \tau_{\beta_2,q} \tag{31}$$

for the prespecified $\alpha_1,\, \alpha_2 > 0$ (as used in (21)) such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$. We will refer to (30) as the "Separation Requirement" (SR) at the level β. In view of

$$\begin{aligned} & \left\| \frac{1}{n} X^T X(\theta^* - \theta) \right\|_q \\ = \; & \left\| \mathbb{E}_W \left[\frac{1}{n} X^T (Y - X\theta) \right] - \mathbb{E}_W \left[\frac{1}{n} X^T (Y - X\theta^*) \right] \right\|_q \\ = \; & \left\| \mathbb{E}_W \left[\frac{1}{n} X^T (Y - X\theta) \right] - \mathbb{E}_W \left[\frac{1}{n} X^T W \right] \right\|_q, \end{aligned}$$

note that the SR is imposed upon the l_q-distance between the population score vectors evaluated at θ^* and $\theta (\in \Theta_0)$.

Our next result concerns the Type II error of the test based on $\Psi_q(\hat{\theta}_\alpha)$ in (7) and $r_{\alpha,q}$ defined in (21). For completeness, we also include the claim for the Type I error.

Theorem 2.1. *Assume (1) where $W \sim \mathcal{N}(\mathbf{0}_n,\, \sigma^2 \boldsymbol{I}_n)$ and is independent of X. For some chosen $\alpha_1,\, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$, consider the statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (21). For any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \textit{(Type I Error)} \tag{32}$$

where $\mathbb{P}_0$ means under H_0. For the same $r_{\alpha,q}$ used in (32) and some $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, if $h(\theta^*) \neq \mathbf{0}_m$ and (30) is satisfied, we have

$$\mathbb{P}_1 \left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \beta, \qquad \textit{(Type II Error)} \tag{33}$$

where $\mathbb{P}_1$ means under H_1 and (30).

2.1.4 Implications of Our Results

Some interesting observations can be made from the results we have established so far. First, our guarantees do not rely on good behavior of $\left\|\hat{\theta}_\alpha - \theta^*\right\|_2$ or even $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ nonasymptotically or asymptotically. As a consequence, θ^* need not be sparse for the results to hold.

Second, ***the number of restrictions (i.e., m) in*** H_0 plays a significant role in the power of our procedures. If $p \geq n$, $m \leq p - n$, and $\Theta_0 \neq \emptyset$, we can always find a solution $\hat{\theta}$ such that $X\hat{\theta} = Y$. Consequently, we have

$$\left\|\frac{1}{n}X^T X(\theta^* - \hat{\theta})\right\|_q = \left\|\left[\frac{1}{n}X^T(Y - X\hat{\theta})\right] - \left[\frac{1}{n}X^T(Y - X\theta^*)\right]\right\|_q = \left\|\frac{1}{n}X^T W\right\|_q. \tag{34}$$

By (11),

$$\mathbb{P}\left(\left\|\frac{1}{n}X^T W\right\|_q \leq \mathbb{E}_W\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + \tau_{\alpha,q}\right) \geq 1 - \alpha,$$

which implies that $\mathbb{P}\left(\left\|\frac{1}{n}X^T X(\theta^* - \hat{\theta})\right\|_q \geq \delta_{\alpha,\beta,q}\right)$ is small and there is not enough separation for our procedures to detect the alternatives. Note that $\left(\hat{\theta},\, \mathbf{0}_p\right)$ $(\left(\hat{\theta},\, 0\right))$ also solves (8) (respectively, (9)) with probability 1 for any $r_{\alpha,q} \geq 0$. Comparing with (27), the length of $\left\|\frac{1}{n}X^T X(\theta^* - \hat{\theta})\right\|_q$ here can be bounded from above by $\mathbb{E}_W\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + \tau_{\alpha,q}$.

As m gets larger relative to $p - n$, it becomes easier for (30) to be satisfied. In general, the more restrictions in H_0 we have (i.e., the larger m is), the more powerful our procedures will be (whether θ^* is sparse or not). This phenomenon is opposite from what have been shown in the existing literature (cf. the discussions in the sixth paragraph of Section 1) and exists not only in the linear regression models here, but also in the models considered in Sections 2.3 and 4.

Third, we observe from (31) and (20) that the quantities taking the form of $\sqrt{\log\frac{1}{\gamma}}$ in $\delta_{\alpha,\beta,q}$ are dimension free. Instead, the leading term $\mathbb{E}_W\left[\left\|\frac{1}{n}X^T W\right\|_q\right]$ in (31) and (11) reflects the "dimension complexity" and automatically takes into consideration the dependencies between the coordinates. This result is a direct consequence of the concentration phenomenon in Lipschitz functions of Gaussians. Take $q = \infty$, $W \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$ and consider the extreme example where X consists of p copies of the same column X_0. Then, we have

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^T W\right\|_\infty\right] = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_{0i}^2}$$

and (31) becomes

$$\delta_{\alpha,\beta,\infty} = 2\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_{0i}^2} + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q} + \sqrt{\frac{1}{R}}\tau_{\beta_1,q} + \tau_{\beta_2,q}, \tag{35}$$

which involves no dimension complexity (as desired). In terms of practical implementation, we have demonstrated that $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ can be well approximated by the data-driven threshold $\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q$; see (18) and (79).

Beyond the extreme example, more generally for $q = \infty$ and $W \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$ (without much loss of generality by assuming $\sigma = 1$), we show in Section A.4 that

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \geq \frac{1}{2}\left(1-\frac{1}{e}\right)\sqrt{\frac{\log p}{4n^2}\min_{j,l\in\{1,\ldots,p\}}\sum_{i=1}^{n}(X_{ij}-X_{il})^2} \tag{36}$$

for all $p \geq 20$, and

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \leq \sqrt{\frac{2\log p}{n^2}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n}X_{ij}^2} + \sqrt{\frac{8}{n^2\log p}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n}X_{ij}^2} \tag{37}$$

for all $p \geq 2$. While the nonasymptotic validity of our testing procedures does not require any growth restrictions on the dimensionality, we see from (36) that $\delta_{\alpha,\beta,\infty}$ can tend to zero only when $\frac{\log p}{n} = o(1)$ (if X does not contain identical columns).

As an alternative, the Bonferroni approach can also be used to construct a testing procedure. In particular, we can solve (8) (or (9)) with $q = \infty$ and

$$r_{\alpha,\infty} = \sqrt{\max_{j\in\{1,\ldots,p\}}\frac{2\sigma^2}{n}\sum_{i=1}^{n}X_{ij}^2}\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}}. \tag{38}$$

Consequently, the separation distance in (30) that allows us to correctly detect an alternative with probability at least $1-\beta$ takes the form

$$\begin{aligned}\delta_{\alpha,\beta,\infty} &= r_{\alpha,\infty} + r_{\beta,\infty} \\ &= \sqrt{\max_{j\in\{1,\ldots,p\}}\frac{2\sigma^2}{n}\sum_{i=1}^{n}X_{ij}^2}\left(\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}} + \sqrt{\frac{1}{n}\log\frac{2p}{\beta}}\right).\end{aligned} \tag{39}$$

In contrast to our previous concentration approach, the Bonferroni alternative derives the upper bound (38) from a simple union bound on $\left\|\frac{1}{n}X^TW\right\|_\infty$; as a consequence, the resulting threshold $r_{\alpha,\infty}$ depends on p and fails to capture the dependencies between the coordinates. In the extreme example discussed previously, note that $\delta_{\alpha,\beta,\infty}$ for the Bonferroni approach can be substantially bigger than (35) due to the extra "$\log p$" term.

2.2 Unknown Noise Variance

When there is no prior information on σ, $\sqrt{\mathrm{Var}\,(Y_i)}$ may be used as an upper bound. We can easily estimate $\sqrt{\mathrm{Var}\,(Y_i)}$ by $\hat{\sigma}_Y = \sqrt{n^{-1}\sum\left(Y_i - \bar{Y}\right)^2}$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. If $\{X_i\}_{i=1}^n$ consists of i.i.d. Gaussian vectors (so that Y is Gaussian), Proposition 4.1 in [1] implies that

$$\sqrt{\mathrm{Var}\,(Y_i)} \leq \left(C_n - \frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\tau}{2}\right)\right)^{-1}\hat{\sigma}_Y := \bar{B}_\tau$$

with probability at least $1-\tau$, where $C_n = \sqrt{\frac{2}{n}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)} = 1 - O(n^{-1})$.

In problems where $\mathrm{Var}(W_i)$ is a constant over i, X is fixed, and the only source of randomness in Y comes from W, replacing σ with $\bar{B}_\tau$ does not make $r_{\alpha,q}$ a more conservative threshold for constructing confidence regions. In problems with a random design, using $\bar{B}_\tau$ could result in confidence regions that are more conservative.

We find it rather challenging to estimate σ precisely and obtain a sharp threshold simultaneously within the non-asymptotic framework. The main issue is that our procedure does not guarantee a small $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ with high probability, which seems to be needed for consistent estimation of σ. On the other hand, if we were able to ensure a small error with respect to the prediction norm, our nonasymptotic control is likely to become less sharper and also involves unknown nuisance parameters that are hard to estimate.

2.3 Gaussian Nonlinear Regressions

The procedures and theory established in Section 2.1 can be easily extended to the Gaussian nonlinear regression models

$$Y_i = \Upsilon\left(X_i;\theta^*\right) + W_i, \qquad i = 1, ..., n, \tag{40}$$

where $W = \{W_i\}_{i=1}^n \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n)$ as before, the functional form of $\Upsilon\left(X_i;\theta^*\right)$ is assumed to be known and possibly nonlinear in θ^*. Our test statistics (7) then takes the form

$$\Psi_q\left(\hat{\theta}_\alpha\right) := \left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - \Upsilon\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_q,$$

which we will refer to as the "quasi-score" evaluated at $\hat{\theta}_\alpha$, a solution to (8) (or (9)) where $X_i\theta_\alpha$ is replaced with $\Upsilon\left(X_i;\theta_\alpha\right)$ for each i. Note that (11), (18), (19), and (79) still hold. As a result, if we replace (30) with

$$\inf_{\theta\in\Theta_0}\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[\Upsilon\left(X_i;\theta^*\right) - \Upsilon\left(X_i;\theta\right)\right]\right\|_q \geq \delta_{\alpha,\beta,q},$$

the statements in Theorem 2.1 and its implications in Section 2.1.4 (with the linear index replaced by Υ for each i) can be carried over to the case of Gaussian nonlinear regressions.

2.4 Implementation

We discuss implementations for some natural choices of q (relevant to both (8) and (9)) and $\tilde{q}$ (relevant to (8)).

For example, letting $q = \infty$ leads to a "Dantzig Selector like" constraint and we simply rewrite it as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(Y_i - X_i\theta_\alpha\right) - \mu_\alpha &\leq r_{\alpha,q} \qquad \forall j = 1, ..., p, \\ \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(-Y_i + X_i\theta_\alpha\right) + \mu_\alpha &\leq r_{\alpha,q} \qquad \forall j = 1, ..., p, \end{aligned}$$

for (8) and as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(Y_i - X_i\theta_\alpha\right) &\leq r_{\alpha,q} + \mu_\alpha \qquad \forall j = 1, ..., p, \\ \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(-Y_i + X_i\theta_\alpha\right) &\leq r_{\alpha,q} + \mu_\alpha \qquad \forall j = 1, ..., p, \end{aligned}$$

for (9). Note that the constraints above are linear whereas for $q = 2$, the first constraint in (8) (respectively, (9)) becomes nonlinear and is implemented without further manipulation in our simulations.

In the cases where $q = \tilde{q} = \infty$ in (8), we work with the following (equivalent) program:

$$\begin{aligned} & \min_{(\theta_\alpha, \mu_\alpha, z) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}} \quad z \\ \text{subject to: } & \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(Y_i - X_i\theta_\alpha\right) - \mu_\alpha \leq r_{\alpha,q} \qquad \forall j = 1, ..., p, \\ & \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(-Y_i + X_i\theta_\alpha\right) + \mu_\alpha \leq r_{\alpha,q} \qquad \forall j = 1, ..., p, \\ & \mu_{j,\alpha} \leq z \qquad \forall j = 1, ..., p, \\ & -\mu_{j,\alpha} \leq z \qquad \forall j = 1, ..., p, \\ & h(\theta_\alpha) = \mathbf{0}_m. \end{aligned} \tag{41}$$

In the cases where $q = \infty$, $\tilde{q} = 1$ in (8), we work with the following (equivalent) program:

$$\min_{(\theta_\alpha,\mu_\alpha,z)\in\mathbb{R}^p\times\mathbb{R}^p\times\mathbb{R}^p} \quad \sum_{j=1}^{p} z_j$$

$$\begin{aligned}
\text{subject to: } \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(Y_i - X_i\theta_\alpha\right) - \mu_\alpha &\leq r_{\alpha,q} \qquad \forall j = 1,...,p, \\
\frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(-Y_i + X_i\theta_\alpha\right) + \mu_\alpha &\leq r_{\alpha,q} \qquad \forall j = 1,...,p, \\
\mu_{j,\alpha} &\leq z_j \qquad \forall j = 1,...,p, \\
-\mu_{j,\alpha} &\leq z_j \qquad \forall j = 1,...,p, \\
h(\theta_\alpha) &= \mathbf{0}_m,
\end{aligned} \tag{42}$$

where $z = \{z_j\}_{j=1}^{p}$. The tricks employed to formulate the programs (41) and (42) are standard in the literature on linear programming (see, e.g., [2]).

3 Simulations

In this section, we evaluate the performance of our procedures through simulation studies. The following choices of q and $\tilde{q}$ are considered:

(I) $q = \tilde{q} = \infty$ in (8),

(II) $q = \infty$, $\tilde{q} = 1$ in (8),

(III) $q = \infty$ in (9),

(IV) $q = 2$, $\tilde{q} = \infty$ in (8),

(V) $q = \tilde{q} = 2$ in (8),

(VI) $q = 2$ in (9).

The optimization problems above are solved with the "interior point" algorithm.

The matrix $X \in \mathbb{R}^{n\times p}$ consists of n rows, which are fixed i.i.d. realizations from the normal distribution $\mathcal{N}(\mathbf{0}_p, \Sigma)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jj'} = 0.3$ for $j \neq j'$ and $j, j' \in \{1,...,p\}$. Our null hypotheses take either the form (6) or

$$H_0 : A\theta^* = \mathbf{0}_m \tag{43}$$

for some prespecified $A \in \mathbb{R}^{m\times p}$ consisting of m rows, which are fixed i.i.d. realizations from the normal distribution $\mathcal{N}(\mathbf{0}_p, \Sigma)$. In the simulations, we assign a seed number, different from what is used to draw the rows in X, to generate the rows in A.

The second form of hypotheses above is motivated by real world applications in marketing, where firms usually can choose or have information about the covariates but lack observations on the outcome. For example, a startup company may only be able to perform experiments over a small set of customers and record their responses. On the other hand, there could be numerous product attributes to be chosen freely by the company; it might also have rich data on the characteristics of potential customers. In these applications, the researchers can often "simulate" the matrix A of their interest (where m can be as large as p). Recalling (2) and (3), one problem is to test $H_0 : A\gamma^* = \mathbf{0}_m$; that is, there is no heterogeneity in the treatment effect for the simulated profiles.

We first look at the case $n = 15$ and $p = 50$ to examine the "small sample" performance of our procedures. A setup with such a small n (but larger p) is rarely seen among existing simulation studies for regression models. In the end we look at $n = 100$ and $p = 300$ to see the improvement. For the form (6), we consider the following scenarios:

(a) $G = \{1, ..., p\} \setminus \{1, \frac{p}{2}, p\}$ with $\theta_j^* = \frac{c_a}{p}$ for all $j = 1, ..., p$,

(b) $G = \{1, ..., p\} \setminus \{1, 2, ..., 9\}$ with $\theta_j^* = \frac{c_b}{p}$ for all $j = 1, ..., p$,

(c) $G = \{1, ..., p\}$ with $\theta_2^* = c_c$ and $\theta_j^* = 0$ for all $j \neq 2$,

(d) $G = \{1, ..., p\} \setminus \{1, \frac{p}{2}, p\}$ with $\theta_2^* = c_d$ and $\theta_j^* = 0$ for all $j \neq 2$.

For the form (43), we let A consist of:

(e) $p - 3$ rows with $\theta_j^* = \frac{c_e}{p}$ for all $j = 1, ..., p$,

(f) $p - 9$ rows with $\theta_j^* = \frac{c_f}{p}$ for all $j = 1, ..., p$.

Our coverage probabilities and rejection probabilities are calculated based on 100 repetitions. The subscripted cs are the approximate cutoff values that make the rejection probabilities at 95% and may vary among (a)-(f). The rejection probabilities decrease as the subscripted cs decrease and vice versa. We first apply methods (I) to (VI) to each of the scenarios listed above under $n = 15$ and $p = 50$.

For each of the 100 repetitions, the noise vector W is drawn from $\mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ with $\sigma = 0.5$; we take $R = 10000$ i.i.d. draws (Z_rs) from $\mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$ and choose $\alpha_1 = 0.049$, $\alpha_2 = 0.001$ (i.e., $\alpha = 0.05$) to balance between $\tau_{\alpha_1,q}$ and $\sqrt{\frac{1}{R}}\tau_{\alpha_2,q}$ in (21). We set $\beta_1 = 0.001$, $\beta_2 = 0.049$ (i.e., $\beta = 0.05$) in (31) and approximate $\delta_{\alpha,\beta,q}$ with

$$\hat{\delta}_{\alpha,\beta,q} = \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q} + \sqrt{\frac{1}{R}}\tau_{\beta_1,q} + \tau_{\beta_2,q},$$

which is compared with the actual separation $\left\| \frac{1}{n} X^T X \left(\theta^* - \hat{\theta}_\alpha \right) \right\|_q$. Tables 3.1-3.7 exhibit:

(i) θ^* that makes the rejection probability (Item v below) at 95%,

(ii) the average of $\left\|\frac{1}{n}X^TX\left(\theta^*-\hat{\theta}_\alpha\right)\right\|_q$ over 100 repetitions,

(iii) the average of $\hat{\delta}_{\alpha,\beta,q}$ over 100 repetitions,

(iv) the coverage probability,

(v) the rejection probability (i.e., 95%).

The evidence from our simulation studies supports the main points of this paper. A rejection can happen when a subscripted c is shared equally over all $j = 1, ..., p$ (which gives approximately sparse θ^* or non-sparse θ^*) or over a single coefficient (which gives exactly sparse θ^*). All it takes is sufficient separation in terms of $\left\|\frac{1}{n}X^TX\left(\theta^*-\hat{\theta}_\alpha\right)\right\|_q$. When the number of restrictions (m) in H_0 decreases from $p-3$ to $p-9$ (Tables 3.1 to 3.2, Tables 3.5 to 3.6), the values of the subscripted cs needed to make the rejection probabilities at 95% become substantially larger, as shown in row (i); that is, our procedures become less powerful. Similar patterns are also observed when m decreases from p to $p-3$ (Tables 3.3 to 3.4). This behavior is opposite from what has been noted in [23]: The smaller $|G|$ gets, the more powerful their procedure becomes in detecting sparse alternatives (such as Scenarios (c) and (d)).

In view of (23)-(24) for (8) and (25)-(26) for (9), it is not surprising that the coverage probabilities of our procedures are not affected by how sparse θ^* is. In contrast to "undercoverage" commonly reported in many asymptotic procedures, the coverage probabilities shown in Tables 3.1-3.7 suggest our method is conservative. The actual separation (ii) needed to achieve a power of 95% is somewhat smaller than $\hat{\delta}_{\alpha,\beta,q}$ (iii). This result is plausible given (33) only states that β is an upper bound on the probability of our procedures failing to reject H_0, under H_1, (30), and (31); also, the proposed separation in (31) is only sufficient but not necessary. Establishing the minimax separation in terms of (30) for (6) and (43) under $p \geq n$, $m > p - n$ could be a useful endeavor for future research.

For the same choice of q, methods (I)-(III) (respectively, methods (IV)-(VI)) give very similar performance as shown by rows (i) and (ii). Between the choices of $q = \infty$ and $q = 2$, methods (I)-(III) appear more powerful than methods (IV)-(VI), evidenced by (i) where the subscripted cs are smaller. From the computational aspect, method (III) is much faster than the rest. It also yields a more natural looking confidence interval (see, (25)-(26)) whose length is simply (27), corresponding to the "bold" numbers in the tables for the various scenarios and (n, p) combinations. Note that the lengths here are the same as $\hat{\delta}_{\alpha,\beta,q}$ for Method III because $\beta_1 = \alpha_2$ and $\beta_2 = \alpha_1$. But of course, for more general β_1 and β_2 as shown in Section 2, the lengths of the confidence intervals do not have to coincide with $\hat{\delta}_{\alpha,\beta,q}$s.

To compare the small sample performance ($n = 15$, $p = 50$) with the larger sample performance ($n = 100$, $p = 300$), we repeat the same exercise with methods

(I) to (VI). For conciseness, we only exhibit the improvement for method (III) under $n = 100$, $p = 300$ in Table 3.7. The improvement for other methods is very similar; the same patterns discussed in the previous paragraph (when $n = 15$, $p = 50$) are also observed under $n = 100$, $p = 300$.

As shown in Table 3.7, the subscripted cs that make the rejection probabilities at 95% decrease drastically in the larger sample experiment. The difference in the subscripted cs between Scenario (a) ($m = p - 3$) and Scenario (b) ($m = p - 9$), between Scenario (e) ($m = p - 3$) and Scenario (f) ($m = p - 9$), between Scenario (c) ($m = p$) and Scenario (d) ($m = p - 3$), respectively, gets smaller as p and n increase. This finding is intuitive and can be reasoned as follows: Relative to the setup $n = 15$ and $p = 50$, the number of "free" parameters remains the same (3 for Scenarios (a), (d), and (e); 9 for Scenario (b) and (f); 0 for Scenario (c)) while n is increased to 100. Consequently, the impact of more restrictions in H_0 on the power becomes less substantial.

Table 3.1: $n = 15$, $p = 50$, **Scenario (a), Methods I-VI**

	I	**II**	**III**	**IV**	**V**	**VI**
i	$0.081 \cdot \mathbf{1}_p$	$0.081 \cdot \mathbf{1}_p$	$0.081 \cdot \mathbf{1}_p$	$0.096 \cdot \mathbf{1}_p$	$0.096 \cdot \mathbf{1}_p$	$0.096 \cdot \mathbf{1}_p$
ii	1.034	0.986	1.034	3.856	3.852	3.852
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.2: $n = 15$, $p = 50$, **Scenario (b), Methods I-VI**

	I	**II**	**III**	**IV**	**V**	**VI**
i	$0.496 \cdot \mathbf{1}_p$	$0.496 \cdot \mathbf{1}_p$	$0.496 \cdot \mathbf{1}_p$	$0.633 \cdot \mathbf{1}_p$	$0.633 \cdot \mathbf{1}_p$	$0.633 \cdot \mathbf{1}_p$
ii	1.084	1.080	1.084	3.739	3.737	3.737
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.3: $n = 15$, $p = 50$, **Scenario (c), Methods I-VI**						
	I	**II**	**III**	**IV**	**V**	**VI**
i	$(0, 0.745, \mathbf{0}_{p-2})$	$(0, 0.745, \mathbf{0}_{p-2})$	$(0, 0.745, \mathbf{0}_{p-2})$	$(0, 0.83, \mathbf{0}_{p-2})$	$(0, 0.83, \mathbf{0}_{p-2})$	$(0, 0.83, \mathbf{0}_{p-2})$
ii	1.025	1.025	1.025	4.286	4.286	4.286
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.4: $n = 15$, $p = 50$, **Scenario (d), Methods I-VI**						
	I	**II**	**III**	**IV**	**V**	**VI**
i	$(0, 4.45, \mathbf{0}_{p-2})$	$(0, 4.45, \mathbf{0}_{p-2})$	$(0, 4.45, \mathbf{0}_{p-2})$	$(0, 4.875, \mathbf{0}_{p-2})$	$(0, 4.875, \mathbf{0}_{p-2})$	$(0, 4.875, \mathbf{0}_{p-2})$
ii	1.015	0.993	1.015	3.738	3.736	3.736
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.5: $n = 15$, $p = 50$, **Scenario (e), Methods I-VI**						
	I	**II**	**III**	**IV**	**V**	**VI**
i	$0.054 \cdot \mathbf{1}_p$	$0.054 \cdot \mathbf{1}_p$	$0.054 \cdot \mathbf{1}_p$	$0.064 \cdot \mathbf{1}_p$	$0.064 \cdot \mathbf{1}_p$	$0.064 \cdot \mathbf{1}_p$
ii	1.053	1.134	1.053	3.859	3.852	3.853
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.6: $n = 15$, $p = 50$, **Scenario (f), Methods I-VI**						
	I	**II**	**III**	**IV**	**V**	**VI**
i	$0.086 \cdot \mathbf{1}_p$	$0.086 \cdot \mathbf{1}_p$	$0.086 \cdot \mathbf{1}_p$	$0.102 \cdot \mathbf{1}_p$	$0.102 \cdot \mathbf{1}_p$	$0.102 \cdot \mathbf{1}_p$
ii	1.098	1.130	1.098	3.796	3.795	3.795
iii	1.602	1.602	**1.602**	6.726	6.726	6.726
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

Table 3.7: $n = 100$, $p = 300$, **Scenarios (a)-(f), Method (III)**						
	a	**b**	**c**	**d**	**e**	**f**
i	$0.008 \cdot \mathbf{1}_p$	$0.021 \cdot \mathbf{1}_p$	$(0, 0.35, \mathbf{0}_{p-2})$	$(0, 0.632, \mathbf{0}_{p-2})$	$0.002 \cdot \mathbf{1}_p$	$0.003 \cdot \mathbf{1}_p$
ii	0.387	0.400	0.370	0.380	0.350	0.395
iii	**0.595**	**0.595**	**0.595**	**0.595**	**0.595**	**0.595**
iv	1	1	1	1	1	1
v	0.95	0.95	0.95	0.95	0.95	0.95

4 Some General Non-Asymptotic Justifications

So far our theory has focused on Gaussian regressions with homoscedastic noise. Is it possible to establish some general non-asymptotic justifications for inference in high dimensional models that involve non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients? We answer this question in this section.

4.1 Regressions with Non-Gaussian Noise

Our analysis in Section 2 exploits sharp concentration of Lipschitz functions of Gaussian variables. This analysis can be extended to regression models where the noise vector W is either bounded or has a strongly log-concave distribution. In particular, we have the following analogues of (11).

Lemma 4.1. *Suppose W has a strongly log-concave*[1] *distribution with parameter φ. Then for any $q \in [1, \infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^T W\right\|_q \geq \mathbb{E}\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + t\right\} \leq \exp\left(\frac{-n\varphi t^2}{2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{44}$$

Remarks. For a fixed design X, if $Y \sim \mathcal{N}(X\theta^*, \Sigma)$ and $\Sigma \succ 0$, φ can be set to the smallest eigenvalue of Σ^{-1}. Beyond a normal distribution, [16] discuss quite a few examples of strongly log-concave distributions.

Lemma 4.2. *Suppose W consists of independent random variables, all of which are supported on $[a, b]$. Then for any $q \in [1, \infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^T W\right\|_q \geq \mathbb{E}\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + t\right\} \leq \exp\left(\frac{-nt^2}{2(b-a)^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{45}$$

If we know the distribution of W, our analysis from Section 2 can be, in principle, extended to construct testing procedures and confidence regions for regression models where W is either bounded or has a strongly log-concave distribution. However, sometimes we might not know the distribution for W; instead, we may have more information on the distribution of Y than the distribution of W. In some applications, we might only know Y consists of entries supported on $[a, b]$. For example, [20] estimate the effect of spending on math pass rates ($Y_i \in [0, 1]$) under the assumption $\mathbb{E}(Y_i|X_i) = \Phi(X_i\theta^*)$, where $\Phi(\cdot)$ denotes the standard normal c.d.f. and

[1] A strongly log-concave distribution is a distribution with density $\mathsf{p}(z) = \exp(-\psi(z))$ such that for some $\varphi > 0$ and all $\lambda \in [0, 1]$, $z, z' \in \mathbb{R}^n$, $\lambda\psi(z) + (1-\lambda)\psi(z') - \psi(\lambda z + (1-\lambda)z') \geq \frac{\varphi}{2}\lambda(1-\lambda)\left\|z - z'\right\|_2^2$.

X_i include the spending variable as well as other covariates. Another example is the binary response model

$$\mathbb{P}\left(Y_i = 1 | X_i\right) = \Lambda\left(X_i; \theta^*\right), \qquad i = 1, ..., n, \tag{46}$$

where $Y_i \in \{0, 1\}$ and the functional form of $\Lambda\left(X_i; \theta^*\right)$ is assumed to be known; for example, Λ may be a "probit" or a "logit" in (46) and $\Lambda\left(X_i; \theta^*\right) = \Lambda\left(X_i \theta^*\right)$. Under the assumption

$$\mathbb{E}\left(Y_i | X_i\right) = \Pi\left(X_i; \theta^*\right), \tag{47}$$

both binary and bounded response models can be treated in the same framework.

4.2 Bounded Responses

In what follows, we consider (47) where $a \leq Y_i \leq b$ for all i, the functional form of $\Pi\left(X_i; \theta^*\right)$ is assumed to be known and possibly nonlinear in θ^*. Without loss of generality, we assume $a = 0$ and $b = 1$. Our test statistics now becomes

$$\Psi_q\left(\hat{\theta}_\alpha\right) := \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \hat{\theta}_\alpha\right) \right] \right\|_q, \tag{48}$$

and $\hat{\theta}_\alpha$ is a solution to

$$\begin{aligned} \left(\hat{\theta}_\alpha, \, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}^p} \|\mu_\alpha\|_{\tilde{q}} \\ \text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \theta_\alpha\right) \right] - \mu_\alpha \right\|_q \leq r_{\alpha, q}, \\ h(\theta_\alpha) = \mathbf{0}_m, \end{aligned} \tag{49}$$

or,

$$\begin{aligned} \left(\hat{\theta}_\alpha, \, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}} \mu_\alpha \\ \text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \theta_\alpha\right) \right] \right\|_q \leq r_{\alpha, q} + \mu_\alpha, \\ h(\theta_\alpha) = \mathbf{0}_m, \\ \mu_\alpha \geq 0. \end{aligned} \tag{50}$$

Throughout this section, we use $\mathbb{E}_{Y|X}\left[\cdot\right]$ to denote the expectation over the distribution of Y conditioning on X; for an i.i.d. sequence of Radamacher random variables, $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ (independent of Y and X), we use $\mathbb{E}_\varepsilon\left[\cdot\right]$ to denote the expectation over

ε only, conditioning on Y and X, and $\mathbb{E}_{\varepsilon,Y|X}[\cdot]$ to denote the expectation over the distribution of $(\varepsilon,\, Y)$ conditioning on X.

Like in the regression problem, we first establish the concentration of

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q$$

around its expectation

$$S_{\theta^*} := \mathbb{E}_{Y|X} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q \right]. \tag{51}$$

Previously we have simply replaced $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ in (11) with its Monte Carlo approximation $\frac{\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q$ and a "small" deviation. This strategy cannot be applied to the expectation S_{θ^*} directly. Instead, we first seek a reasonable upper bound which involves only $\{Y,\, X\}$ and random variables from a known distribution. These results are stated in the following proposition.

Proposition 4.1. *Assume Y consists of independent random variables. For any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q \geq S_{\theta^*} + t \right\} \leq \exp\left(\frac{-nt^2}{2 \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \right\|_q^2} \right). \tag{52}$$

Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables independent of Y and X. Under (47), we have

$$\mathbb{E}_{\varepsilon,Y|X}\left\{ \left\| \frac{1}{2n} \sum_{i=1}^{n} \varepsilon_i X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q \right\} \leq S_{\theta^*} \leq 2\mathbb{E}_{\varepsilon,Y|X}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i Y_i \right\|_q \right\}. \tag{53}$$

Remarks. Note that bound (52) holds for any fixed θ (not just the true coefficient vector, θ^*). However, (53) relies crucially on the model assumption (47).

The upper bound in (53) can be viewed as the symmetrized version of S_{θ^*}. Considering a collection of i.i.d. Radamacher random draws (independent of Y and X),

$$\{\varepsilon_{ir} :\, i = 1, ..., n,\, r = 1, ..., R\}, \tag{54}$$

we can replace S_{θ^*} with $\frac{2}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q$ (a Monte-Carlo approximation of the symmetrized version) and some "small" deviations. The complementary lower bound in (53) suggests that S_{θ^*} and its symmetrized version have the magnitude. As a consequence, our replacement strategy is not an overly conservative approach for constructing critical values.

Hypothesis Testing

To avoid repetition, we omit the discussion on the "ideal" confidence regions and directly jump to the construction of the test statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical) $r_{\alpha,q}$ and $\hat{\theta}_\alpha$. The first step is to relate S_{θ^*} with $\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$ as shown in the following proposition.

Proposition 4.2. *Assume (47) where $0 \le Y_i \le 1$ for all i and the functional form of $\Pi\left(X_i;\theta^*\right)$ is known. Given (54) which is independent of Y and X, for any $q \in [1, \infty]$, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i-\Pi\left(X_i;\theta^*\right)\right]\right\|_q \ge \frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q + t_1 + 2t_2 + 2t_3 \tag{55}$$

with probability no greater than $\alpha \in (0, 1)$, where

$$\begin{aligned}
t_1 &= \tau_{\alpha_1,q} = \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}},\\
t_2 &= \tau_{\alpha_2,q} = \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_2}},\\
t_3 &= \frac{2}{\sqrt{R}}\tau_{\alpha_3,q} = \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{8}{nR}\log\frac{1}{\alpha_3}},
\end{aligned}$$

for some chosen α_1, α_2, $\alpha_3 > 0$ such that $\sum_{k=1}^{3}\alpha_k = \alpha$.

Construction of Critical Values ($r_{\alpha,q}$) and Type I Error

Based on (55) along with the choices of t_1, t_2 and t_3 above, we set in (49) (or (50)),

$$r_{\alpha,q} = \frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q + \tau_{\alpha_1,q} + 2\tau_{\alpha_2,q} + \frac{4}{\sqrt{R}}\tau_{\alpha_3,q}. \tag{56}$$

Under H_0, $(\theta^*, \mathbf{0}_p)$ $((\theta^*, 0))$ is an optimal solution to (49) (respectively, (50)) with $r_{\alpha,q}$ specified in (56). Consequently, a (practical) optimal solution to (49) (and (50)) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha) \ge r_{\alpha,q}\right\} \le \alpha \qquad \textit{(Type I Error)}. \tag{57}$$

Separation Requirement and Type II Error

Letting $\Theta_0 := \{\theta \in \mathbb{R}^p : h(\theta) = \mathbf{0}_m\}$, we choose $\beta_1, \beta_2, \beta_3 > 0$ such that $\sum_{k=1}^3 \beta_k = \beta \in (0, 1)$, and assume

$$\inf_{\theta \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^n X_i \left[\Pi\left(X_i; \theta^*\right) - \Pi\left(X_i; \theta\right) \right] \right\|_q \geq \delta_{\alpha,\beta,q} \tag{58}$$

with

$$\begin{aligned} \delta_{\alpha,\beta,q} = \mathbb{E}_{Y|X} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\|_q \right] + \mathbb{E}_{\varepsilon,Y|X} \left[\left\| \frac{2}{n} \sum_{i=1}^n \varepsilon_i X_i Y_i \right\|_q \right] \\ + \tau_{\alpha_1,q} + 2\tau_{\alpha_2,q} + \sqrt{\frac{16}{R}} \tau_{\alpha_3,q} + 2\tau_{\beta_1,q} + \sqrt{\frac{16}{R}} \tau_{\beta_2,q} + \tau_{\beta_3,q}, \end{aligned} \tag{59}$$

for the prespecified $\alpha_1, \alpha_2, \alpha_3 > 0$ (as used in (56)) such that $\sum_{k=1}^3 \alpha_k = \alpha \in (0, 1)$. Note that the SR is imposed upon the l_q-distance between the "quasi score" vectors evaluated at θ^* and $\theta (\in \Theta_0)$, since

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n X_i \left[\Pi\left(X_i; \theta^*\right) - \Pi\left(X_i; \theta\right) \right] \right\|_q \\ = & \left\| \mathbb{E}_{Y|X} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta\right) \right] \right\} - \mathbb{E}_{Y|X} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \left[Y_i - \Pi\left(X_i; \theta^*\right) \right] \right\} \right\|_q . \end{aligned}$$

Our next result concerns the Type II error of the test based on $\Psi_q(\hat{\theta}_\alpha)$ in (48) and $r_{\alpha,q}$ defined in (56). For completeness, we also exhibit the Type I error and the practical confidence regions in this result.

Theorem 4.1. *Suppose the conditions in Propositions 4.1 and 4.2 hold. For some chosen $\alpha_1, \alpha_2, \alpha_3 > 0$ such that $\sum_{k=1}^3 \alpha_k = \alpha \in (0, 1)$, consider the statistics $\Psi_q(\hat{\theta}_\alpha)$ based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (56). For any $q \in [1, \infty]$, we have*

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \textit{(Type I Error)} \tag{60}$$

where $\mathbb{P}_0$ means under H_0. For the same $r_{\alpha,q}$ used in (60) and some $\beta_1, \beta_2, \beta_3 > 0$ such that $\sum_{k=1}^3 \beta_k = \beta \in (0, 1)$, if $h(\theta^*) \neq \mathbf{0}_m$ and (58) is satisfied, we have

$$\mathbb{P}_1 \left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \beta, \qquad \textit{(Type II Error)} \tag{61}$$

where $\mathbb{P}_1$ means under H_1 and (58).

Furthermore, an optimal solution $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ *to (49) must satisfy*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_{\tilde{q}} \;\geq\; \|\hat{\mu}_\alpha\|_{\tilde{q}}\,, \tag{62}$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right] - \hat{\mu}_\alpha\right\|_{q} \;\leq\; 2r_{\alpha,q}, \tag{63}$$

with probability at least $1-\alpha$*. Similarly, an optimal solution* $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ *to (50) must satisfy*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_{q} \;\geq\; \hat{\mu}_\alpha, \tag{64}$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i;\theta^*\right) - \mathit{\Pi}\left(X_i;\hat{\theta}_\alpha\right)\right]\right\|_{q} \;\leq\; 2r_{\alpha,q} + \hat{\mu}_\alpha, \tag{65}$$

with probability at least $1-\alpha$*.*

For deriving $r_{a,q}$ in (49) or (50), the strategy where we replace S_{θ^*} in (51) by $\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$ plus some "small" deviations only requires the correct specification of the conditional mean of Y_i, i.e., (47). This treatment delivers generic confidence regions in the form of (62)-(63) or (64)-(65).

Note that the assumptions in Theorem 4.1 allow for the possibilities of heteroscedastic "noise" $(Y_i - \mathit{\Pi}\left(X_i;\theta^*\right))$ as well as nonlinearity in θ^*, while requiring no specific knowledge on the distribution for Y (other than it is bounded). In the linear regression model $Y = X\theta^* + W$, [10] resolve the issues of heteroscedasticity and non-Gaussian responses by tailoring the Bonferroni approach to self-normalized sums. Their confidence regions involve several unknown nuisance parameters that are hard to estimate in practice. Even in the case where the noise variances are known and homoscedastic, to apply the confidence sets in [10] for testing hypotheses of the form $H_0 : \theta^*_j = 0\ \forall j \in \{1,2,...,p\}$ (for example), one would require sufficient sparsity in θ^* as well as prior knowledge on the underlying sparsity (e.g., an upper bound on the number of non-zero coefficients in θ^*).

5 A New Class of Regularized Estimators

Beyond the context of hypothesis testing, the data-driven approach proposed in Section 2 for setting $r_{\alpha,q}$ suggests a new class of regularized estimators:

$$\hat{\theta}^{new}_\alpha \in \arg\min_{\theta_\alpha\in\mathbb{R}^p} \|\theta_\alpha\|_{\tilde{q}} \;\text{ subject to }\left\|\frac{1}{n}X^T(Y - X\theta_\alpha)\right\|_q \leq r_{\alpha,q}, \tag{66}$$

where

$$r_{\alpha,q} = \frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^T Z_r\right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q}, \tag{67}$$

$$\tau_{\alpha_1,q} = \sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}}, \tag{68}$$

for some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0, 1)$.

Unlike (8) or (9), (66) has a different objective, $\min_{\theta_\alpha\in\mathbb{R}^p}\|\theta_\alpha\|_{\tilde{q}}$, and does not involve the slack vector (or variable) μ_α in $\left\|\frac{1}{n}X^T(Y - X\theta_\alpha)\right\|_q \leq r_{\alpha,q}$. Moreover, the solution to (66) is not constrained to satisfy $h(\hat{\theta}_\alpha^{new}) = \mathbf{0}_m$, whereas our estimator $\hat{\theta}_\alpha$ in Section 2 satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$. When $\tilde{q} = 1$ and $q = \infty$, we may view (66) as a variant of the Dantzig selector.

In what follows, let $\hat{\theta}_\alpha^{new}$ be a solution to the program (66) with $\tilde{q} = 1$ and $q = \infty$. We can establish an upper bound on $\left\|\hat{\theta}_\alpha^{new} - \theta^*\right\|_2$ using the l_2-sensitivity defined as follows:

$$\kappa_{J_*} := \inf_{\Delta\in\mathbb{C}_{J_*}:\|\Delta\|_2=1}\left\|\frac{1}{n}X^T X\Delta\right\|_\infty \tag{69}$$

where

$$\begin{aligned} J_* &:= \left\{j \in \{1, ..., p\} : \theta_j^* \neq 0\right\}, \\ \mathbb{C}_{J_*} &:= \left\{\Delta \in \mathbb{R}^p : \|\Delta_{J_*^c}\|_1 \leq \|\Delta_{J_*}\|_1\right\}, \end{aligned}$$

where Δ_J denotes the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on the set J and zero coordinates on the complement J^c of J. The l_2-sensitivity is introduced by [9][2] and similar to the cone invertibility factors defined in [21]. In particular, under a coherence condition introduced by [8], Proposition 4.2 in [9] shows that

$$\kappa_{J_*} \gtrsim \frac{1}{\sqrt{|J_*|}} \tag{70}$$

where $|J_*|$ denotes the cardinality of J_* and $f(n) \gtrsim g(n)$ means $f(n) \geq C_0 g(n)$ for some constant $C_0 \in (0, \infty)$.

The following result concerns the l_2-error bound for $\hat{\theta}_\alpha^{new}$.

Theorem 5.1. *Assume (1) where $W \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n)$ and is independent of X. Choosing $\tilde{q} = 1$ and $q = \infty$ in (66) and setting $r_{\alpha,q}$ according to (67) with $q = \infty$, we have*

$$\mathbb{P}\left(\left\|\hat{\theta}_\alpha^{new} - \theta^*\right\|_2 \leq \frac{2r_{\alpha,\infty}}{\kappa_{J_*}}\right) \geq 1 - \alpha \tag{71}$$

[2]In contrast to (66), the estimators in [9] and [10] rely on the Bonferroni approach tailored to the self-normalized sums.

where κ_{J_*} is defined in (69).

In view of (37), (79) and (70), we see that the rate of our $\hat{\theta}_\alpha^{new}$, i.e., $\kappa_{J_*}^{-1}\sqrt{\frac{\log p}{n}}$, is not worse than the typical rate $\sqrt{\frac{|J_*|\log p}{n}}$ for estimation (see, e.g., [3]). If $|J_*|$ is large relative to n (lack of sparsity), then $\kappa_{J_*}^{-1}$ could diverge faster than (or no slower than) $\sqrt{\frac{n}{\log p}}$.

The innovation of (66) lies in the use of (67) which can accurately approximate the term $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ in (11) via Monte-Carlo and automatically take into consideration the dependencies across coordinates. This fact makes (66) in contrast with the Bonferroni approach which would set $r_{\alpha,\infty}$ proportional to $\sqrt{\frac{1}{n}\log\frac{2p}{\alpha}}$. As we have demonstrated earlier (cf., Section 2.1.4), in the presence of strong dependencies between the columns in X, $r_{\alpha,\infty}$ for the Bonferroni approach can be substantially bigger than (67) due to the extra "$\log p$" term.

In the situation where the noise variance σ is not known *a priori*, we can always modify (66) by adopting the approach in Section 2. Alternatively, it is also possible to modify the optimization procedure in [10] with our data-driven approach for setting the constraint on $\left\|\frac{1}{n}X^T(Y-X\theta_\alpha)\right\|_\infty$.

Remarks. Note that the confidence interval in (71) cannot be computed easily as κ_{J_*} is unknown and hard to estimate. Even for testing a hypothesis such as $H_0: \theta^* = \mathbf{0}_p$, deriving a practical critical value for the statistics $\left\|\hat{\theta}_\alpha^{new}\right\|_2$ is a challenging task. For this reason, we have chosen to work with the score tests (which require no conditions on sparsity) as demonstrated in Section 2.

6 Conclusion

We have developed non-asymptotically justified methods for hypothesis testing about the coefficients ($\theta^* \in \mathbb{R}^p$) in the high dimensional (generalized) regression models where p can exceed the sample size n. Relative to existing literature, we look at broader forms of hypotheses and the impact of the number of restrictions in the null hypothesis. In particular, we consider $H_0: h(\theta^*) = \mathbf{0}_m$ against the alternative hypothesis $H_1: h(\theta^*) \neq \mathbf{0}_m$, where m can be as large as p and the function of interest $h: \mathbb{R}^p \mapsto \mathbb{R}^m$ can be nonlinear in θ^*. Our test statistics is based on the sample score vector evaluated at an estimate $\hat{\theta}_\alpha$ that satisfies $h(\hat{\theta}_\alpha) = \mathbf{0}_m$, where α is the prespecified Type I error. Our controls on the Type I and Type II errors for the score test are nonasymptotic. In addition, confidence regions are constructed in terms of the score vectors.

By exploiting the concentration phenomenon in Lipschitz functions, the key component reflecting the "dimension complexity" in our non-asymptotic thresholds uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated

around and automatically takes into account the dependencies between the coordinates. The novelty of our methods is that their validity does not rely on good behavior of $\left\|\hat{\theta}_\alpha - \theta^*\right\|_2$ or even $n^{-1/2}\left\|X\left(\hat{\theta}_\alpha - \theta^*\right)\right\|_2$ nonasymptotically or asymptotically. Most interestingly, we discover phenomena that are opposite from the existing literature: (1) More restrictions (larger m) in H_0 make our procedures more powerful; (2) whether θ^* is sparse or not, it is possible for our procedures to detect alternatives with probability at least $1 -$ Type II error when $p \geq n$ and $m > p - n$; (3) the coverage probability of our procedures is not affected by how sparse θ^* is.

The proposed procedures are evaluated with simulation studies where we consider a "small sample" setup ($n = 15$, $p = 50$) and a "larger sample" setup ($n = 100$, $p = 300$). Our designs range from highly dense θ^* to highly sparse θ^* and our null hypotheses take either the form (6) or $H_0 : A\theta^* = \mathbf{0}_m$, for some prespecified $A \in \mathbb{R}^{m\times p}$ and $m \in \{p, p-3, p-9\}$. The empirical results are promising and support our key insights.

We have also provided some general nonasymptotic justifications for inference in high dimensional models that involve non-Gaussian responses, heteroscedastic noise, and nonlinearity in the regression coefficients (including the binary response models and certain nonlinear regressions). As a secondary contribution, we have also proposed a new class of regularized estimators along with a complementary l_2-error bound, which are motivated by the data-driven feature of our concentration approach.

Acknowledgment

I thank Guang Cheng at Purdue University for discussions that improved some clarity of this work as well as pointing out several related papers. All errors are my own.

A Supplementary Materials

A.1 Preliminary

Here we include several classical results which are used in the main proofs. We first introduce a definition of sub-Gaussian variables.

Definition A.1. *A zero-mean random variable U_1 is sub-Gaussian if there is a $\nu > 0$ such that*

$$\mathbb{E}\left[\exp\left(\lambda U_1\right)\right] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right) \tag{72}$$

for all $\lambda \in \mathbb{R}$, and we refer to ν as the sub-Gaussian parameter.

Remarks.

1. Using the Chernoff bound, one can show that any zero-mean random variable U_1 obeying (72) satisfies

$$\mathbb{P}\left(U_1 \leq -t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{73}$$

$$\mathbb{P}\left(U_1 \geq t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{74}$$

   for all $t \geq 0$.

2. Let $\{U_i\}_{i=1}^R$ be independent zero-mean sub-Gaussian random variables, each with parameter at most ν. Then $R^{-1}\sum_{i=1}^R U_i$ is sub-Gaussian with parameter at $\nu/\sqrt{R}$. To see this, note that for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{R}\sum_{i=1}^R U_i\right)\right] &= \prod_{i=1}^R \mathbb{E}\left[\exp\left(\frac{\lambda U_i}{R}\right)\right] \\ &\leq \prod_{i=1}^R \exp\left(\frac{\lambda^2\nu^2}{2R^2}\right) \\ &= \exp\left(\frac{\lambda^2\nu^2}{2R}\right).\end{aligned} \tag{75}$$

The following result exhibits the type of sub-Gaussian variables that are of interest to our analysis.

Lemma A.1. *Suppose $U = \{U_i\}_{i=1}^n$ has a strongly log-concave distribution with parameter $\varphi > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$ is $L-$Lipschitz with respect to the Euclidean*

norm. Then for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left(\frac{\lambda^2 L^2}{2\varphi}\right). \tag{76}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \leq -t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2L^2}\right), \\
\mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \geq t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2L^2}\right).
\end{aligned}$$

Remarks. The proof involves the so-called "inf-convolution" argument and an application of the Brunn-Minkowski inequality; see [4] and [14].

Lemma A.2. *Assume $U = \{U_i\}_{i=1}^n$ consists of independent random variables, all of which are supported on $[a, b]$. If $f : \mathbb{R}^n \to \mathbb{R}$ is separately convex*[3] *and $L-$Lipschitz with respect to the Euclidean norm, then for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left[\frac{\lambda^2 (b-a)^2 L^2}{2}\right]. \tag{77}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \leq -t\right] &\leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right), \\
\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \geq t\right] &\leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).
\end{aligned}$$

Remarks. One proof for Lemma A.2 involves the entropy method and the so-called Herbst argument; see [5]. Talagrand and Ledoux have contributed to the result above in different papers.

A.2 Proof of Proposition 2.1

For any $q \in [1, \infty]$, $\left\|\frac{1}{n} X^T W\right\|_q$ is Lipschitz in W with respect to the Euclidean norm. To see this, note that a triangle inequality and a Cauchy-Schwarz inequality

[3] Let the function $f_j : \mathbb{R} \to \mathbb{R}$ be defined by varying only the jth co-ordinate of a function $f : \mathbb{R}^n \to \mathbb{R}$; f is *separately convex* if for each $j \in \{1, 2, ..., n\}$, f_j is a convex function of the jth coordinate.

yield

$$\begin{aligned}\left|\left\|\frac{1}{n}X^TW\right\|_q-\left\|\frac{1}{n}X^TW'\right\|_q\right| &\leq \left\|\frac{1}{n}X^T\left(W-W'\right)\right\|_q \\ &\leq \frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\left\|W-W'\right\|_2. \qquad (78)\end{aligned}$$

As a result of Lemma A.1, we have the concentration in (11).

If $h(\theta^*)=\mathbf{0}_m$, (11) then implies that $(\theta^*,\,\mathbf{0}_p)$ $((\theta^*,\,0))$ is an optimal solution to (8) (respectively, (9)). If $h(\theta^*)\neq\mathbf{0}_m$, since $\{\theta\in\mathbb{R}^p:\,h(\theta)=\mathbf{0}_m\}\neq\emptyset$, we can find some $\tilde{\theta}_\alpha$ such that $h(\tilde{\theta}_\alpha)=\mathbf{0}_m$. Letting

$$\tilde{\mu}_\alpha=\frac{1}{n}X^T(Y-X\tilde{\theta}_\alpha)-\frac{1}{n}X^T(Y-X\theta^*)=\frac{1}{n}X^T(X\theta^*-X\tilde{\theta}_\alpha),$$

(11) then implies that $\left(\tilde{\theta}_\alpha,\,\tilde{\mu}_\alpha\right)$ $(\left(\tilde{\theta}_\alpha,\,\|\tilde{\mu}_\alpha\|_q\right))$ is a feasible solution to (8) (respectively, (9)) with probability at least $1-\alpha$. In any case, an optimal solution $\left(\hat{\theta}_\alpha^*,\,\hat{\mu}_\alpha^*\right)$ to (8) must satisfy

$$\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\frac{1}{n}X^T(Y-X\theta^*)\right\|_{\tilde{q}}=\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha^*)\right\|_{\tilde{q}}\geq\|\hat{\mu}_\alpha^*\|_{\tilde{q}}$$

with probability at least $1-\alpha$. Similarly, an optimal solution $\left(\hat{\theta}_\alpha^*,\,\hat{\mu}_\alpha^*\right)$ to (9) must satisfy

$$\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\frac{1}{n}X^T(Y-X\theta^*)\right\|_q=\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha^*)\right\|_q\geq\hat{\mu}_\alpha^*$$

with probability at least $1-\alpha$. On the other hand, in terms of (8), applying the triangle inequality yields

$$\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha^*)-\hat{\mu}_\alpha^*\right\|_q\leq\left\|\frac{1}{n}X^TW\right\|_q+\left\|\frac{1}{n}X^T(Y-X\hat{\theta}_\alpha^*)-\hat{\mu}_\alpha^*\right\|_q\leq 2r_{\alpha,q}^*$$

with probability at least $1-\alpha$. In terms of (9), we simply have

$$\mathbb{P}\left(\left\|\frac{1}{n}X^TX(\theta^*-\hat{\theta}_\alpha^*)\right\|_q\leq 2r_{\alpha,q}^*+\hat{\mu}_\alpha^*\right)\geq 1-\alpha.$$

A.3 Proof of Theorem 2.1

We have already derived (32) in Section 2. To show (33), we define the event

$$\mathcal{E}=\left\{\frac{\sigma}{R}\sum_{r=1}^R\left\|\frac{1}{n}X^TZ_r\right\|_q\geq\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+\sqrt{\frac{1}{R}}\tau_{\beta_1,q}\right\}.$$

As we have argued for (18), we also have the upper deviation inequality

$$\mathbb{P}\left\{\frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^TZ_r\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+t\right\} \leq \exp\left(\frac{-nRt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right) \tag{79}$$

and consequently, $\mathbb{P}(\mathcal{E}) \leq \beta_1$. Let $\mathcal{E}^c$ denote the complement of $\mathcal{E}$. Under H_1 and (30), we have

$$\begin{aligned}
&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}\right\}\\
=&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}^c\right\}\mathbb{P}(\mathcal{E}^c)+\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}\right\}\mathbb{P}(\mathcal{E})\\
\leq&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}^c\right\}+\mathbb{P}(\mathcal{E})\\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}X^T(X\theta^*-X\hat{\theta}_\alpha)\right\|_q-\left\|\frac{1}{n}X^TW\right\|_q \leq r_{\alpha,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\mathbb{P}\left\{\delta_{\alpha,\beta,q}-\left\|\frac{1}{n}X^TW\right\|_q \leq r_{\alpha,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]+\tau_{\beta_2,q}|\mathcal{E}^c\right\}+\beta_1\\
\leq&\beta
\end{aligned}$$

where the fifth line follows from (30) and the sixth line follows from (31), the fact that we are conditioning on $\mathcal{E}^c$, as well as (11).

A.4 Additional Derivations

To show (36), we define an i.i.d. sequence of Gaussian random variables

$$\widetilde{W}_k \sim \mathcal{N}\left(0, \min_{j,l\in\{1,\ldots,p\}}\frac{1}{2n^2}\sum_{i=1}^n(X_{ij}-X_{il})^2\right)$$

for $k = 1, ..., p$. Note that we have

$$\mathbb{E}_W\left[\left(\frac{1}{n}X_j^TW-\frac{1}{n}X_l^TW\right)^2\right] \geq \mathbb{E}_{\widetilde{W}}\left[\left(\widetilde{W}_j-\widetilde{W}_l\right)^2\right].$$

By the Sudakov-Fernique Gaussian comparison result (see Corollary 3.14 in [13]), we obtain

$$\begin{aligned}\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] &\geq \mathbb{E}_W\left[\max_{j\in\{1,\ldots,p\}}\frac{1}{n}X_j^TW\right]\\ &\geq \frac{1}{2}\mathbb{E}_{\widetilde{W}}\left[\max_{j\in\{1,\ldots,p\}}\widetilde{W}_j\right]\\ &\geq \frac{1}{2}\left(1-\frac{1}{e}\right)\sqrt{\frac{\log p}{4n^2}\min_{j,l\in\{1,\ldots,p\}}\sum_{i=1}^n(X_{ij}-X_{il})^2}\end{aligned}$$

(for all $p \geq 20$), where the last line follows from a classical lower bound on the Gaussian maximum (see, e.g., [13]). The upper bound

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \leq \sqrt{\frac{2\log p}{n^2}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^n X_{ij}^2}+\sqrt{\frac{8}{n^2\log p}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^n X_{ij}^2}$$

(for all $p \geq 2$) is another classical result on the Gaussian maximum (see, e.g., [18]).

Remarks. To obtain the lower bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$, we first compare the dependent sequence $\left\{\frac{1}{n}X_j^TW\right\}_{j=1}^p$ with another independent Gaussian sequence $\widetilde{W} = \left\{\widetilde{W}_j\right\}_{j=1}^p$ and then apply a lower bound on $\mathbb{E}_{\widetilde{W}}\left[\max_{j\in\{1,\ldots,p\}}\widetilde{W}_j\right]$. In contrast, the upper bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ is obtained by applying $\sum_{j=1}^p\mathbb{P}\left(\left|\frac{1}{n}X_j^TW\right|\geq t\right)$, where independence is not needed. Moreover, the upper bound also holds when W is a sequence of sub-Gaussian variables while the lower bound requires W to be a sequence of Gaussian variables.

A.5 Proofs of Lemmas 4.1 and 4.2

As a result of Lemma A.1 and (78), we have the concentration in Lemma 4.1. Because $\left\|\frac{1}{n}X^TW\right\|_q$ is separately convex in terms of W, Lemma A.2 implies the concentration in Lemma 4.2.

A.6 Proof of Proposition 4.1

Using the argument that leads to (78), we can show $\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q$ is Lipschitz in Y with respect to the Euclidean norm for any $q \in [1, \infty]$. That is,

$$\begin{aligned}
&\left|\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q - \left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i^{'} - \Pi\left(X_i;\theta^*\right)\right]\right\|_q\right| \\
\leq &\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}\right\|_q \left\|Y - Y^{'}\right\|_2 . \qquad (80)
\end{aligned}$$

Note that $\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q$ is separately convex in terms of Y. As a result of Lemma A.2, we have the concentration in (52).

To establish (53), we exploit the convexity of l_q-norms and the fact that $\mathbb{E}\left(Y_i|X_i\right) = \Pi\left(X_i;\theta^*\right)$. Let $Y^{'} = \left\{Y_i^{'}\right\}_{i=1}^{n}$ be an independent sequence identical to but independent of Y conditioning on X, and $\varepsilon = \{\varepsilon_i\}_{i=1}^{n}$ be i.i.d. Radamacher random variables independent of Y, $Y^{'}$, and X. We obtain

$$\begin{aligned}
&\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \Pi\left(X_i;\theta^*\right)\right]\right\|_q\right\} \\
=&\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left[Y_i - \mathbb{E}_{Y_i^{'}|X_i}\left(Y_i^{'}\right)\right]\right\|_q\right\} \\
=&\mathbb{E}_{Y|X}\left\{\left\|\mathbb{E}_{Y^{'}|X}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - Y_i^{'}\right)\right]\right\|_q\right\} \\
\leq&\mathbb{E}_{Y^{'},Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\} \\
=&\mathbb{E}_{\varepsilon,Y^{'},Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i\left(Y_i - Y_i^{'}\right)\right\|_q\right\} \\
\leq&2\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i Y_i\right\|_q\right\}, \qquad (81)
\end{aligned}$$

where the second line follows since $\mathbb{E}\left(Y_i^{'}|X_i\right) = \Pi\left(X_i;\theta^*\right)$, the fourth line follows from Jensen's inequality, and the sixth line follows from the fact that $\varepsilon_i X_i\left(Y_i - Y_i^{'}\right)$ and $X_i\left(Y_i - Y_i^{'}\right)$ have the same distribution.

On the other hand, similar argument from above also yields

$$\begin{aligned}
&\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i X_i\left[Y_i-\Pi\left(X_i;\theta^*\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i X_i\left[Y_i-\mathbb{E}_{Y_i^{'}|X_i}\left(Y_i^{'}\right)\right]\right\|_q\right\}\\
\leq&\mathbb{E}_{\varepsilon,Y^{'},Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\\
=&\mathbb{E}_{Y^{'},Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}.
\end{aligned}$$

Applying the following inequality

$$\begin{aligned}
&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\\
\leq&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-\Pi\left(X_i;\theta^*\right)\right)\right\|_q+\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i^{'}-\Pi\left(X_i;\theta^*\right)\right)\right\|_q,
\end{aligned}$$

and taking expectations gives

$$\mathbb{E}_{Y^{'},Y|X}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\leq\mathbb{E}_{Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-\Pi\left(X_i;\theta^*\right)\right)\right\|_q\right\}.$$

Putting the pieces together, we obtain the result in (53).

A.7 Proof of Proposition 4.2

We first show that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i Y_i\right\|_q\right\}$ is Lipschitz in Y with respect to the Euclidean norm for any $q\in[1,\,\infty]$. That is,

$$\begin{aligned}
&\left|\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i Y_i\right\|_q\right\}-\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i Y_i^{'}\right\|_q\right\}\right|\\
\leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\mathbb{E}_\varepsilon\left[\sum_{i=1}^{n}\varepsilon_i^2\left(Y_i-Y_i^{'}\right)^2\right]}\\
\leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\left\|Y-Y^{'}\right\|_2.
\end{aligned}$$

Note that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right\}$ is separately convex in terms of Y. As a result of Lemma A.2, we have the following concentration

$$\mathbb{P}\left\{\mathbb{E}_{\varepsilon,Y|X}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right\} \geq \mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right\} + \tau_{\alpha_2,q}\right\} \leq \alpha_2. \tag{82}$$

Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables, independent of Y and X. We can again show that $\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i Y_i X_i\right\|_q$ is Lipschitz in ε with respect to the Euclidean norm for any $q \in [1, \infty]$ and the Lipschitz constant[4] is $\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i X_i)^2}\right\|_q$, which is bounded from above by $\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q$ given $0 \leq Y_i \leq 1$. Let $\{\varepsilon_{ir} : i = 1, ..., n,\ r = 1, ..., R\}$ be a collection of i.i.d. Radamacher random draws, independent of Y and X. Conditioning on Y and X, (75) and (77) imply $\frac{1}{R}\sum_{r=1}^R \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{ir} Y_i X_i\right\|_q - \mathbb{E}_\varepsilon\left(\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right)$ is sub-Gaussian with parameter at most $\frac{2}{\sqrt{nR}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{Y|X}\left[\mathbb{E}_\varepsilon\left[\exp\left(\lambda\left[\frac{1}{R}\sum_{r=1}^R\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{ir} Y_i X_i\right\|_q - \mathbb{E}_\varepsilon\left(\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right)\right]\right)\right]\right] \\
\leq &\exp\left[\lambda^2 \frac{4\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}{2nR}\right].
\end{aligned}$$

Consequently, (73) yields the following concentration

$$\mathbb{P}\left\{\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i Y_i\right\|_q\right\} \geq \frac{1}{R}\sum_{r=1}^R\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{ir} Y_i X_i\right\|_q + \frac{2}{\sqrt{R}}\tau_{\alpha_3,q}\right\} \leq \alpha_3 \tag{83}$$

Combining (52), (81), (82) and (83) yields (55).

[4]Like $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q$, we define

$$\begin{aligned}
\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i X_i)^2}\right\|_q &= \sqrt[q]{\sum_{j=1}^p\left(\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2 X_{ij}^2}\right)^q}, \qquad q \in [1, \infty) \\
\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i X_i)^2}\right\|_q &= \max_{j\in\{1,...,p\}}\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2 X_{ij}^2}, \qquad q = \infty.
\end{aligned}$$

A.8 Proof of Theorem 4.1

We have already derived (60) in Section 4. For the confidence regions in Theorem 4.1, we simply follow the same argument used in the proof for Proposition 2.1.

To show (61), let us define the event

$$\mathcal{E} = \left\{ \frac{1}{R}\sum_{r=1}^{R} \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q \geq \mathbb{E}_{\varepsilon, Y|X} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i Y_i \right\|_q \right\} + \tau_{\beta_1,q} + \frac{2}{\sqrt{R}}\tau_{\beta_2,q} \right\}$$

for some chosen β_1, $\beta_2 > 0$ such that $\beta_1 + \beta_2 \in (0,\, 1)$. As we have argued for (82) and (83), we also have the upper deviation result $\mathbb{P}\{\mathcal{E}\} \leq \beta_1 + \beta_2$. We use $\mathcal{E}^c$ to denote the complement of $\mathcal{E}$. Note that

$$\mathbb{P}\left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \mathbb{P}\left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} | \mathcal{E}^c \right\} + \mathbb{P}\left(\mathcal{E}\right).$$

Let $\beta_3 = \beta - \beta_1 - \beta_2$. Since $\mathbb{P}\left(\mathcal{E}\right) \leq \beta_1 + \beta_2$, it suffices to show that

$$\begin{aligned}
&\mathbb{P}_1\left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} | \mathcal{E}^c \right\} \\
\leq &\mathbb{P}_1\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[\mathit{\Pi}\left(X_i; \theta^*\right) - \mathit{\Pi}\left(X_i; \hat{\theta}_\alpha\right) \right] \right\|_q - \left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[Y_i - \mathit{\Pi}\left(X_i; \theta^*\right) \right] \right\|_q \leq r_{\alpha,q} | \mathcal{E}^c \right\} \\
\leq &\mathbb{P}_1\left\{ \delta_{\alpha,\beta,q} - \left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[Y_i - \mathit{\Pi}\left(X_i; \theta^*\right) \right] \right\|_q \leq r_{\alpha,q} | \mathcal{E}^c \right\} \\
\leq &\mathbb{P}_1\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} X_i \left[Y_i - \mathit{\Pi}\left(X_i; \theta^*\right) \right] \right\|_q \geq S_{\theta^*} + \tau_{\beta_3,q} | \mathcal{E}^c \right\} \\
\leq &\beta_3,
\end{aligned}$$

where the third line follows from (58) and the fourth line follows from (59), the fact that we are conditioning on $\mathcal{E}^c$, as well as (52).

A.9 Proof of Theorem 5.1

For some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$, let us define the event

$$\mathcal{E} = \left\{ \left\| \frac{1}{n} X^T W \right\|_\infty \leq r_{\alpha,\infty} \right\}$$

where $r_{\alpha,\infty}$ is defined in (67). Bound (19) implies that $\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \alpha$. We use the notation $\hat{\Delta} = \hat{\theta}_\alpha^{new} - \theta^*$ in the following. On the event $\mathcal{E}$, we obtain

$$\left\| \frac{1}{n} X^T X \hat{\Delta} \right\|_\infty \leq \left\| \frac{1}{n} X^T W \right\|_\infty + \left\| \frac{1}{n} X^T (Y - X\hat{\theta}_\alpha^{new}) \right\|_\infty \leq 2 r_{\alpha,\infty}. \tag{84}$$

Given $\mathcal{E}$, θ^* is feasible for (66) and consequently,

$$\left\|\hat{\theta}_{\alpha}^{new}\right\|_1 \leq \|\theta^*\|_1 ,$$

which implies that

$$\left\|\hat{\Delta}_{J_*^c}\right\|_1 \leq \|\theta^*_{J_*}\|_1 - \left\|\hat{\theta}^{new}_{\alpha,J_*}\right\|_1 \leq \left\|\hat{\theta}^{new}_{\alpha,J_*} - \theta^*_{J_*}\right\|_1 = \left\|\hat{\Delta}_{J_*}\right\|_1 ; \tag{85}$$

that is, $\hat{\Delta} \in \mathbb{C}_{J_*}$. Using the definition of κ_{J_*} in (69), (84) and (85) imply that

$$\left\|\hat{\theta}_{\alpha}^{new} - \theta^*\right\|_2 \leq \frac{2r_{\alpha,\infty}}{\kappa_{J_*}}$$

with probability at least $1 - \alpha$.

References

[1] Arlot, S., G. Blanchard, and E. Roquain (2010). "Some Nonasymptotic Results on Resampling in High Dimension, I: Confidence Regions." *Annals of Statistics*, 38, 51-82.

[2] Bertsimas, D. and J. Tsitsiklis (1997). *Introduction to Linear Optimization*, Athena Scientific.

[3] Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics*, 37, 1705-1732.

[4] Bobkov, S. G. and M. Ledoux (2000). "From Brunn-Minkowski to Brascamp-Lieb and to Logarithmic Sobolev Inequalities." *Geometric and Functional Analysis*. 10, 1028-1052.

[5] Boucheron, S, G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. Oxford.

[6] Chernozhukov, V., D. Chetverikov, and K. Kato (2013). "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors." *Annals of Statistics*, 41, 2786-2819.

[7] Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017). "High-Dimensional Simultaneous Inference with the Bootstrap." *Test*, 26, 685-719.

[8] Donoho, D. L., M. Elad, and V. N. Temlyakov (2006). "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise". *IEEE Transactions on Information Theory*, 52, 6-18.

[9] Gautier, E. and A. B. Tsybakov (2011). "High-Dimensional Instrumental Variables Regression and Confidence Sets." Manuscript. CREST (ENSAE).

[10] Gautier, E. and A. B. Tsybakov (2014). "High-Dimensional Instrumental Variables Regression and Confidence Sets." Manuscript. CREST (ENSAE).

[11] Horowitz, J. L. (2017). "Non-Asymptotic Inference in Instrumental Variables Estimation." Manuscript. Northwestern University.

[12] Javanmard, A. and A. Montanari (2014). "Confidence Intervals and Hypothesis Testing for High- Dimensional Regression." *Journal of Machine Learning Research*, 15, 2869-2909.

[13] Ledoux, M., and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.

[14] Maurey, B. (1991). "Some Deviation Inequalities." *Geometric and Functional Analysis*. 1, 188-197.

[15] Ning, Y and H. Liu (2017). "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models." *Annals of Statistics*, 45, 158-195.

[16] Saumard, A. and J. A. Wellner (2014). "Log-Concavity and Strong Log-Concavity: A Review." *Statistics Surveys,* 8, 45-114.

[17] van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models." *Annals of Statistics*, 42, 1166-1202.

[18] Wainwright, M. J. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. In preparation. University of California, Berkeley.

[19] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

[20] Wooldridge, J. M. and Y. Zhu (2017). "Inference in Approximately Sparse Correlated Random Effects Probit Models." Forthcoming in *Journal of Business and Economic Statistics*.

[21] Ye, F., and C.-H. Zhang (2010). "Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls". *Journal of Machine Learning Research*, 11, 3519-3540.

[22] Zhang C.-H. and S. S. Zhang (2014). "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217-242.

[23] Zhang, X. and Cheng, G. (2017). "Simultaneous Inference for High-Dimensional Linear Models." *Journal of the American Statistical Association - Theory & Methods*, 112, 757-768.

[24] Zhu, Y. and J. Bradic (2017). "Linear Hypothesis Testing in Dense High-Dimensional Linear Models." Forthcoming in *Journal of the American Statistical Association*.