All of Linear Regression

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, and Junhui Cai

University of Pennsylvania

e-mail: arunku@upenn.edu, buja.at.wharton@gmail.com

Abstract: Least squares linear regression is one of the oldest and widely used data analysis tools. Although the theoretical analysis of ordinary least squares (OLS) estimator is as old, several fundamental questions are yet to be answered. Suppose regression observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ (not necessarily independent) are available. Some of the questions we deal with are as follows: under what conditions, does the OLS estimator converge and what is the limit? What happens if the dimension is allowed to grow with n? What happens if the observations are dependent with dependence possibly strengthening with n? How to do statistical inference under these kinds of misspecification? What happens to OLS estimator under variable selection? How to do inference under misspecification and variable selection?

We answer all the questions raised above with one simple deterministic inequality which holds for any set of observations and any sample size. This implies that all our results are finite sample (non-asymptotic) in nature. At the end, one only needs to bound certain random quantities under specific settings of interest to get concrete rates and we derive these bounds for the case of independent observations. In particular the problem of inference after variable selection is studied, for the first time, when d, the number of covariates increases (almost exponentially) with sample size n. We provide comments on the "right" statistic to consider for inference under variable selection and efficient computation of quantiles.

1. Introduction

Linear regression is one of the oldest and most widely practiced data analysis method. In many real data settings least squares linear regression leads to performance in par with state-of-the-art (and often far more complicated) methods while remaining amenable to interpretation. These advantages coupled with the argument "all models are wrong" warrants a detailed study of least squares linear regression estimator in settings that are close to the practical/realistic ones. Instead of proposing assumptions that we think are practical/realistic, we start with a clean slate. We start by not assuming anything about the observations

$(X_1^\top, Y_1)^\top, \ldots, (X_n^\top, Y_n)^\top \in \mathbb{R}^d \times \mathbb{R}$ and study the OLS estimator $\hat{\beta}$ given by

$$\hat{\beta} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2,$$

where arg min represents a θ at which the minimum is attained and this $\hat{\beta}$ may not be unique, in which case any of the minimizers is set as $\hat{\beta}$. This clean slate study should be compared to the usual assumption-laden approach where one usually starts by assuming that there exists a vector $\beta_0 \in \mathbb{R}^d$ such that $Y_i = X_i^\top \beta_0 + \varepsilon_i$ for independent and identically distributed Gaussian homoscedastic errors $\varepsilon_1, \ldots, \varepsilon_n$. The classical linear regression setting (Gauss-Markov model) sometimes also assumes $X_1, \ldots, X_n$ are deterministic/non-stochastic. In this model, it is well-known that $\hat{\beta}$ has a normal distribution and is the best linear unbiased estimator (BLUE) for every sample size $n \geqslant d$.

Why is a clean slate study possible? At first glance it might seem strange how a study without assumptions is possible. For a simple explanation, set

$$\hat{\Gamma} := \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \quad \text{and} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top. \tag{1}$$

Now the vector $\hat{\beta}$ can be written as

$$\hat{\beta} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} -2\theta^\top \hat{\Gamma} + \theta^\top \hat{\Sigma} \theta, \tag{2}$$

which implies that $\hat{\beta}$ is a minimizer of a (positive semi-definite) quadratic problem. Intuition suggests that if $\hat{\Gamma} \approx \Gamma$ and $\hat{\Sigma} \approx \Sigma$ then $\hat{\beta}$ is close to β given by

$$\beta := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} -2\theta^\top \Gamma + \theta^\top \Sigma \theta. \tag{3}$$

A follow-up of this intuition suggests an explicit bound on $\|\hat{\beta} - \beta\|$ given bounds on $\|\hat{\Gamma} - \Gamma\|$ and $\|\hat{\Sigma} - \Sigma\|$, for (possibly different) norms $\|\cdot\|$. This viewpoint is usually seen in perturbation analysis of optimization problems; see Bonnans and Shapiro (2013). Note that (2) can be seen as a perturbation of (3). Implementation of this program leads to our deterministic inequality and all subsequent results follow from this result as relatively simple corollaries.

Organization of the paper. The remaining paper is organized as follows. We start, in Section 2, with a simple deterministic inequality that provides "consistency" and "asymptotic normality" of the OLS estimator $\hat{\beta}$. This will be a part

survey with full proofs since similar results appeared before. We will describe explicit corollaries of this inequality for a Berry–Esseen type result for $\hat{\beta}$ that bounds the closeness of the distribution of $\hat{\beta}$ to that of a normal distribution; this is a finite sample result. In a way, this completes the study of OLS estimator in the clean slate setting because normal approximation is the crucial ingredient in statistical inference leading to confidence intervals and hypothesis tests; this discussion is given in Section 3. The test statistics and confidence regions presented in this section are different from the ones used in the classical study. We chose to present the unconventional ones since they will be useful in the study of OLS estimator in presence of variable (or covariate) selection.

We then proceed to study OLS in presence of variable selection in Section 4. The setting here is that the analyst choses a subset of covariates (possibly depending on the data) and then consider the OLS estimator on that subset of covariates. Thanks to the deterministic inequality in Section 2, the results for this setting also follow directly. As a corollary, we also prove a Berry–Esseen type result uniformly over all subset of variables. We end Section 4 with a discussion on how to perform statistical inference under variable selection in case observations are "weakly" dependent without stressing on details (about resampling). This discussion also includes the question of the "right" statistic to consider to inference under variable selection. All the results to this point will be deterministic, finite sample (or non-asymptotic). In Section 5, we provide explicit rate bounds for remainders in the deterministic inequalities from previous sections under independence of observations. This will complete the study of inference under variable selection, at least under independence, when the number of covariates is allowed to increase. We supplement these theoretical results with some numerical evidence in Section 6 where the proposed statistics for inference under variable selection are compared to the ones in the literature. The paper ends with a discussion and some comments on computation for inference under variable selection in Section 7.

Notation. The following notation will be useful. For any vector $v \in \mathbb{R}^d$, $v^\top$ represents its transpose and $v_M \in \mathbb{R}^{|M|}$ for $M \subseteq \{1, 2, \ldots, d\}$ represents the subvector of v with entries in M. For instance $v = (4, 3, 2, 1)^\top$ and $M = \{2, 3\}$ then $v_M = (3, 2)^\top$. Similarly for a symmetric matrix $A \in \mathbb{R}^{d\times d}$, $A_M \in \mathbb{R}^{|M|\times|M|}$ represents the sub-matrix of A with entries in $M \times M$. The Euclidean norm in any dimension is given by $\|\cdot\|$. For any matrix A, let $\|A\|_{op}$ represents the operator norm of A, that is, $\|A\|_{op} = \sup_{\|\theta\|=1} \|A\theta\|$. For any vector $\mu \in \mathbb{R}^q$ and any covariance matrix

$\Omega \in \mathbb{R}^{q\times q}$, $N(\mu, \Omega)$ represents the (multivariate) normal distribution with mean μ, covariance Ω and with some abuse of notation we also use $N(\mu, \Omega)$ to denote a random vector with that Gaussian distribution. For any covariance matrix A, $A^{1/2}$ represents the matrix square root and when we write A^{-1} it is implicitly assumed that A is invertible with inverse A^{-1}. The identity matrix in dimension q is given by I_q. Further for any covariance matrix $A \in \mathbb{R}^{q\times q}$ and vector $x \in \mathbb{R}^q$, $\|x\|_A := \sqrt{x^\top A x}$.

2. Main Deterministic Inequality

Recall the quantities $\hat{\Gamma}$ and $\hat{\Sigma}$ defined in (1). The following result proves deterministic bounds on estimation error and linear representation error for the OLS estimator $\hat{\beta}$. Let $(t)_+ := \max\{0, t\}$ for $t \in \mathbb{R}$ and for any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^\Sigma := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_{op}. \tag{4}$$

Theorem 2.1 (Deterministic Inequality)**.** *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and for any vector $\beta \in \mathbb{R}^d$, we have*

$$(1+\mathcal{D}^\Sigma)^{-1}\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma \;\leqslant\; \|\hat{\beta} - \beta\|_\Sigma \;\leqslant\; (1-\mathcal{D}^\Sigma)_+^{-1}\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma. \tag{5}$$

Furthermore,

$$\|\hat{\beta} - \beta - \Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma \;\leqslant\; \mathcal{D}^\Sigma(1-\mathcal{D}^\Sigma)_+^{-1}\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma. \tag{6}$$

Proof. From the definition of $\hat{\beta}$, we have the normal equations $\hat{\Sigma}\hat{\beta} = \hat{\Gamma}$. Subtracting $\hat{\Sigma}\beta \in \mathbb{R}^d$ from both sides, we get $\hat{\Sigma}(\hat{\beta} - \beta) = \hat{\Gamma} - \hat{\Sigma}\beta$, which is equivalent to

$$(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})\Sigma^{1/2}(\hat{\beta} - \beta) \;=\; \Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta).$$

Adding and subtracting I_d from the parenthesized term with further rearrangement, we get $\Sigma^{1/2}(\hat{\beta} - \beta) - \Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta) \;=\; \left(I_d - \Sigma^{-1/2}\Sigma\Sigma^{-1/2}\right)\Sigma^{1/2}(\hat{\beta} - \beta)$. Taking Euclidean norm on both sides yields

$$\begin{aligned}\left\|\Sigma^{1/2}\left[\hat{\beta} - \beta - \Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\right]\right\| &= \|(I_d - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})\Sigma^{1/2}(\hat{\beta} - \beta)\| \\ &\leqslant \|I_d - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\|_{op}\|\Sigma^{1/2}(\hat{\beta} - \beta)\| \\ &= \mathcal{D}^\Sigma\|\hat{\beta} - \beta\|_\Sigma,\end{aligned} \tag{7}$$

where the inequality follows from the definition of the operator norm, $\|\cdot\|_{op}$. Triangle inequality shows $|\|\hat{\beta} - \beta\|_\Sigma - \|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma| \;\leqslant\; \|\hat{\beta} - \beta - \Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma$, which when combined with (7) yields

$$\|\hat{\beta} - \beta\|_\Sigma \;\leqslant\; \frac{\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma}{(1-\mathcal{D}^\Sigma)_+} \quad \text{and} \quad \|\hat{\beta} - \beta\|_\Sigma \;\geqslant\; \frac{\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma}{1+\mathcal{D}^\Sigma}.$$

These inequalities prove (5) and when combined with (7) implies (6). □

Theorem 2.1 is a very general result that holds for any set of observations (not even necessarily random). It is noteworthy that the result holds for any symmetric matrix Σ and "target" vector $\beta \in \mathbb{R}^d$. A canonical choice of Σ and β are given by

$$\Sigma := \mathbb{E}[\hat{\Sigma}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^\top\right], \quad \text{and} \quad \beta := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[(Y_i - X_i^\top \theta)^2\right]. \tag{8}$$

It is important to note here that just by taking expectations we do not necessarily require all observations to be (non-trivially) random; even fixed numbers are random with a degenerate distribution. For example, in the classical linear model X_i's are treated fixed and non-stochastic in which case $\Sigma = \hat{\Sigma}$ and hence $\mathcal{D}^\Sigma = 0$. Moreover, we neither require any specific dependence structure on the observations nor any specific scaling of dimension d with n. By a careful inspection of the proof and a slight adjustment of $\mathcal{D}^\Sigma$ in (4), it is possible to prove the result for $\|\hat{\beta} - \beta\|$, the usual Euclidean norm, instead of $\|\hat{\beta} - \beta\|_\Sigma$. The added advantage of using $\|\cdot\|_\Sigma$ is affine invariance of the result.

Flexibility in the Choice of Σ and β. For most purposes the canonical choices of Σ, β in (8) suffice but for some applications involving sub-sampling and cross-validation, the flexibility in choosing Σ, β helps. For instance, consider the OLS estimator constructed based on the first $n-1$ observations, that is,

$$\hat{\beta}_{-n} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} (n-1)^{-1} \sum_{i=1}^{n-1} (Y_i - X_i^\top \theta)^2 = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} -2\theta^\top \hat{\Gamma}_{-n} + \theta^\top \hat{\Sigma}_{-n} \theta,$$

where $\hat{\Gamma}_{-n} := (n-1)^{-1}\sum_{i=1}^{n-1} X_i Y_i$ and $\hat{\Sigma}_{-n} := (n-1)^{-1}\sum_{i=1}^{n-1} X_i X_i^\top$. It is of natural interest to compare $\hat{\beta}_{-n}$ with $\hat{\beta}$ rather than the canonical choice of β. In this case Σ is taken to be $\hat{\Sigma}$ which is much closer to $\hat{\Sigma}_{-n}$ than $\mathbb{E}[\hat{\Sigma}_{-n}]$:

$$\hat{\Sigma}_{-n} = \frac{n}{n-1}\left(\hat{\Sigma} - n^{-1} X_n X_n^\top\right) \quad \Rightarrow \quad \hat{\Sigma}^{-1/2}\hat{\Sigma}_{-n}\hat{\Sigma}^{-1/2} = \frac{n I_d}{n-1} - \frac{\hat{\Sigma}^{-1/2} X_n X_n^\top \hat{\Sigma}^{-1/2}}{n-1}.$$

Hence $\|\hat{\Sigma}^{-1/2}\hat{\Sigma}_{-n}\hat{\Sigma}^{-1/2} - I_d\|_{op} \leqslant (n-1)^{-1}[1 + \|\hat{\Sigma}^{-1/2} X_n\|^2]$.

2.1. Consistency of $\hat{\beta}$

If $\mathcal{D}^\Sigma < 1$, then inequalities in (5) provides both upper bounds and lower bounds on the estimation error $\|\hat{\beta} - \beta\|_\Sigma$ that match up to a constant multiple. This allows one to state that necessary and sufficient condition for convergence of $\|\hat{\beta} - \beta\|_\Sigma$ to zero is $\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma$ has to converge to zero. Note that with the choices in (8)

$\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)$ is a mean zero random vector obtained by averaging n random vectors and hence "weak" dependence implies convergence of covariance to zero implying convergence to zero. This implies consistency of the OLS estimator $\hat{\beta}$ to β:

Corollary 2.1 (Consistency). *If $\mathcal{D}^{\Sigma} < 1$ and $\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_{\Sigma}$ converges to zero in probability then $\|\hat{\beta} - \beta\|_{\Sigma}$ converges to zero in probability.*

Turning to inequality (6), note that if $\mathcal{D}^{\Sigma} \to 0$ (in appropriate sense) then inequality (6) provides an expansion of $\hat{\beta} - \beta$ since the remainder (the right hand side of (6)) is of smaller order than $\hat{\beta} - \beta$. Observe that $\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta) = n^{-1}\sum_{i=1}^{n} \Sigma^{-1} X_i (Y_i - X_i^{\top}\beta)$, and hence (6) shows that $\hat{\beta} - \beta$ behaves like an average (a linear functional) up to a lower order term. The claim

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma^{-1} X_i (Y_i - X_i^{\top}\beta) + o_p(1), \tag{9}$$

is usually referred to as an influence function expansion or a linear approximation result. This plays a pivotal role in statistical inference because of the following reason. Ignoring the $o_p(1)$ term, the right hand side of (9) is a mean zero (scaled) average of random vectors which, under almost all dependence settings of interest, converges to a normal distribution if the dimension d is fixed or even diverging "slow enough". This implies that $\sqrt{n}(\hat{\beta} - \beta)$ has an asymptotic normal distribution and an accessible estimator of the (asymptotic) variance implies confidence intervals/regions and hypothesis tests. This discussion is in asymptotic terms and can be made explicitly finite sample which we do in the following subsection with inference related details in the next section.

2.2. Normal Approximation: Berry–Esseen Result

In the following corollary (of Theorem 2.1), we prove a bound on closeness of distribution of $\hat{\beta} - \beta$ to a normal distribution. We need some definitions. Set

$$\Delta_n := \sup_{A \in \mathcal{C}_d} \left| \mathbb{P}(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta) \in A) - \mathbb{P}\left(N(0, K) \in A\right) \right|, \tag{10}$$

where $\mathcal{C}_d$ represents the set of all convex sets in $\mathbb{R}^d$ and $K := \mathrm{Var}(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta))$. For any matrix A, let $\|A\|_{HS}$ represent the Hilbert-Schmidt (or Frobenius) norm, that is, $\|A\|_{HS}^2 := \sum_{i,j} A^2(i,j)$. Also, for any positive semi-definite matrix A, let $\|A\|_*$ denote the nuclear norm of the matrix A.

Corollary 2.2 (Berry–Esseen bound for OLS)**.** *Fix any* $\eta \in (0,1)$*. Then there exists universal constants* $c_1, c_2 > 0$ *such that for all* $n \geqslant 1$*,*

$$\sup_{A \in \mathcal{C}_d} \left| \mathbb{P}(\hat{\beta} - \beta \in A) - \mathbb{P}\left(N(0, \Sigma^{-1/2} K \Sigma^{-1/2}) \in A\right) \right| \leqslant 4\Delta_n + 2n^{-1} + c_2 \|K_n^{-1}\|_*^{1/4} r_n \eta + \mathbb{P}\left(\mathcal{D}^\Sigma > \eta\right),$$

where recall $\mathcal{D}^\Sigma$ *from* (4) *and* $r_n := c_1^{-1} \|K^{1/2}\|_{op} \sqrt{\log n} + \|K^{1/2}\|_{HS}$*.*

The proof of the corollary can be found in Appendix A and it does not require (8). The proof of normal approximation for multivariate minimum contrast estimators in Pfanzagl (1973) is very similar to that of Corollary 2.2. Like Theorem 2.1, Corollary 2.2 is also a finite sample result that does not assume any specific dependence structure on the observations. The quantity Δ_n in (10) is a quantification of convergence of right hand side of (9) to a normal distribution and is bounded by the available multivariate Berry–Esseen bounds. Such bounds for independent (but not necessarily identically distributed) random vectors can be found in Bentkus (2004) and Raič (2018). For dependent settings, multivariate Berry–Esseen bounds are hard to find but univariate versions available (in Romano and Wolf (2000) and Hörmann (2009)) can be extended to multivariate versions by the characteristic function method and smoothing inequalities. In this respect, we note here that the proof of Corollary 2.2 can be extended to prove a normal approximation result for $\alpha^\top(\hat{\beta} - \beta)$ for any specific direction $\alpha \in \mathbb{R}^d$ and for this univariate random variable results from above references apply directly. Finally to get concrete rates from the bound in Corollary 2.2, we only need to choose $\eta \in (0,1)$ and for this we need to control the tail probability of $\mathcal{D}^\Sigma$ in (4). There are two choices for this. Firstly, assuming moment bounds for $X_1, \ldots, X_n$, it is possible to get a tail bound for $\mathcal{D}^\Sigma$ under reasonable dependence structures; see Kuchibhotla et al. (2018a) and Koltchinskii and Lounici (2017a, for independence case). Secondly, one can use a Berry–Esseen type result (Koltchinskii and Lounici, 2017b) for $\mathcal{D}^\Sigma$ which also implies an exponential tail bound up to an analogue of Δ_n term.

Glimpse of the Rates. Assuming observations $(X_i, Y_i), 1 \leqslant i \leqslant n$ are sufficiently weakly dependent and have enough moments, it can be proved that

$$\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma + \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_{op} = O_p(1)\sqrt{\frac{p}{n}}. \tag{11}$$

See Section 5. For concrete rates in normal approximation, observe that

$$\|K^{-1}\|_*^{1/4} \leqslant p^{1/4} \|K^{-1}\|_{op}^{1/4} \quad \text{and} \quad \|K^{1/2}\|_{HS} = \sqrt{\operatorname{tr}(K)} \leqslant p^{1/2} \|K\|_{op}^{1/2}.$$

This implies that

$$\|K^{-1}\|_*^{1/4} r_n \;=\; p^{1/4}\|K^{-1}\|_{op}^{1/4} \;\times\; O(\|K\|_{op}^{1/2}\sqrt{\log n} + p^{1/2}\|K\|_{op}^{1/2}).$$

Under weak enough dependence structure, $\Sigma^{1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)$ is $O_p(n^{-1/2})$ in any fixed direction and hence $\|K\|_{op} = O_p(n^{-1})$ where, recall, $K = \mathrm{Var}(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta))$. Assuming $\|K^{-1}\|_{op} \asymp \|K\|_{op}^{-1}$, we get $\|K^{-1}\|_*^{1/4} r_n = O(n^{-1/4}[p^{1/4}\sqrt{\log n} + p^{3/4}])$. In the best case scenario $\Delta_n \geqslant O(p^{7/4}n^{-1/2})$ and hence to match this rate, we can to take $\eta = O(n^{-1/4})$ which is a permissible choice under (11). Hence we can claim

$$\sup_{A \in \mathcal{C}_d} \left| \mathbb{P}(\hat{\beta} - \beta \in A) - \mathbb{P}\left(N(0, \Sigma^{-1/2} K \Sigma^{-1/2}) \in A\right) \right| = O(1)\frac{p^{7/4}}{n^{1/2}}.$$

We have intentionally left the conditions vague which will be cleared in Section 5.

The Curious Case of Fixed Covariates. In the conventional linear models theory, the covariates are treated fixed/non-stochastic. Since our results are deterministic in nature, this distinction does not matter for the validity of our results. However, in case of fixed covariates the canonical choices for Σ, β mentioned above result in simpler results. For instance, it is clear that non-stochastic covariates leads to $\Sigma = \hat{\Sigma}$, both of which are non-stochastic, and hence $\mathcal{D}^{\Sigma} = 0$. Theorem 2.1 now implies that $\|\hat{\beta} - \beta - \hat{\Sigma}^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_{\hat{\Sigma}} = 0$, or equivalently, $\hat{\beta} - \beta = \hat{\Sigma}^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)$ which is trivial from the definition of $\hat{\beta}$. Further from Corollary 2.2, we get

$$\sup_{A \in \mathcal{C}_d} |\mathbb{P}(\hat{\beta} - \beta \in A) - \mathbb{P}(N(0, \Sigma^{-1/2} K \Sigma^{-1/2}) \in A)| \leqslant 4\Delta_n + 2n^{-1},$$

since η can be taken to be zero in limit. In fact a careful modification of the proof leads to a sharper right hand side as Δ_n. These calculations hint at a previously unnoticed phenomenon: The bounds for random covariates are inherently larger than those for fixed covariates (although they are all of same order). A similar statement also holds when some of the covariates are fixed but others are random (the bounds have extra terms only for random set of covariates). This phenomenon means that, when working with finite samples, the statistical conclusions can be significantly distorted depending on whether the covariates are treated fixed or random. Here it is worth mentioning that the canonical choice of β changes depending on whether covariates are treated random or fixed. If the covariates are fixed, then the canonical choice β is $\beta = (n^{-1}\sum_{i=1}^n x_i x_i^\top)^{-1}(n^{-1}\sum_{i=1}^n x_i \mathbb{E}[Y_i])$, where we write x_i (rather than X_i) to represent fixed nature of covariates. If the covariates are random, then the canonical choice β is $\beta = (n^{-1}\sum_{i=1}^n \mathbb{E}[X_i X_i^\top])^{-1}(n^{-1}\sum_{i=1}^n \mathbb{E}[X_i Y_i])$.

3. Statistical Inference for the OLS estimator

Given that the distribution of $\hat{\beta} - \beta$ is close to a mean zero Gaussian, inference follows if the variance of the Gaussian can be estimated. The variance of the Gaussian is given by

$$\Sigma^{-1} V \Sigma^{-1} := \Sigma^{-1} \mathrm{Var}\left(n^{-1} \textstyle\sum_{i=1}^{n} X_i (Y_i - X_i^\top \beta)\right) \Sigma^{-1}, \tag{12}$$

which is, sometimes, referred to as the sandwich variance. The two ends of the variance Σ^{-1} can be estimated by $\hat{\Sigma}^{-1}$. The only troublesome part is the "meat" part which is the variance of a mean zero average. Estimation of this part requires an understanding of the dependence structure of observations. For instance if the observations are independent then we can readily write

$$V = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i (Y_i - X_i^\top \beta)) \;\preceq\; \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^\top (Y_i - X_i^\top \beta)^2\right]. \tag{13}$$

The inequality above is the matrix inequality representing the difference of matrices is positive semi-definite. A strict inequality above can hold since the observations need not satisfy $\mathbb{E}[X_i(Y_i - X_i^\top \beta)] = 0$. (The definition of β only implies $\sum_{i=1}^{n} \mathbb{E}[X_i(Y_i - X_i^\top \beta)] = 0$.) The last term on the right of (13) can be estimated by $n^{-2} \sum_{i=1}^{n} X_i X_i^\top (Y_i - X_i^\top \hat{\beta})^2$ (obtained by removing the expectation and then replacing β by $\hat{\beta}$). This leads to asymptotically conservative inference for β and it can be proved that asymptotically exact inference is impossible without further assumptions such as $\mathbb{E}[X_i(Y_i - X_i^\top \beta)] = 0$ for all i; see Bachoc et al. (2016, Proposition 3.5) for an impossibility result. Instead if the observations are not independent but m-dependent, then the first equality of (13) does not hold and a correction is needed involving the covariances of different summands; see White (2001) for details under specific dependence structures.

Once an estimator (possibly conservative) $\hat{\Sigma}^{-1} \hat{V} \hat{\Sigma}^{-1}$ of the variance is available, a (possibly conservative) $(1-\alpha)$-confidence region for $\beta \in \mathbb{R}^d$ can be obtained as

$$\hat{\mathcal{R}}_{2,\alpha} := \{\theta \in \mathbb{R}^d : (\hat{\beta} - \theta)^\top \hat{\Sigma} \hat{V}^{-1} \hat{\Sigma} (\hat{\beta} - \theta) \leqslant \chi^2_{d,\alpha}\}, \tag{14}$$

where $\chi^2_{d,\alpha}$ represents the $(1-\alpha)$-th quantile of the chi-square distribution with d degrees of freedom. If $\hat{V}$ is an asymptotically conservative estimator for V that is $\hat{V} \to \bar{V}$ (in an appropriate sense) and $\bar{V} \succeq V$, then

$$\begin{aligned} &\mathbb{P}(N(0, \Sigma^{-1} V \Sigma^{-1})^\top \hat{\Sigma} \hat{V}^{-1} \hat{\Sigma} N(0, \Sigma^{-1} V \Sigma^{-1}) \leqslant \chi^2_{d,\alpha}) \\ &\quad \to \mathbb{P}(N(0, \bar{V}^{-1/2} V \bar{V}^{-1/2})^\top N(0, \bar{V}^{-1/2} V \bar{V}^{-1/2})) \geqslant 1 - \alpha, \end{aligned}$$

where strict inequality holds if $\bar{V} \succ V$; the inequality above is true because of Anderson's lemma (Anderson, 1955, Corollary 3) and it may not be true for non-symmetric confidence regions. An alternate $(1-\alpha)$-confidence region for β is

$$\hat{\mathcal{R}}_{\infty,\alpha} := \left\{\theta \in \mathbb{R}^d : \max_{1\leqslant j\leqslant d} \left|\widehat{AV}_j^{-1/2}(\hat{\beta}_j - \theta_j)\right| \leqslant z_{\infty,\alpha}\right\}, \tag{15}$$

where $\widehat{AV}_j$ represents the j-th diagonal entry of the variance estimator $\hat{\Sigma}^{-1}\hat{V}\hat{\Sigma}^{-1}$ and $z_{\infty,\alpha}$ is the $(1-\alpha)$-th quantile of $\max_{1\leqslant j\leqslant d}|AV_j^{-1/2}N(0,AV)_j|$, with $AV \in \mathbb{R}^{d\times d}$ represents the variance matrix $\Sigma^{-1}V\Sigma^{-1}$.

Hypothesis tests for $\beta \in \mathbb{R}^d$ can also be performed based on the statistics used in (14) and (15). It is easy to verify that neither statistic uniformly beats the other in terms of power. The tests for a single coordinate β_j are easy to obtain from the statistic $(\hat{\beta}_j - \beta_j)/\widehat{AV}_j^{1/2}$ which is close to a standard normal random variable.

The advantage of $\hat{\mathcal{R}}_{\infty,\alpha}$ over $\hat{\mathcal{R}}_{2,\alpha}$ is that it leads to a rectangular region and hence easily interpretable inference for coordinates of β. The confidence region $\hat{\mathcal{R}}_{2,\alpha}$ which is elliptical makes this interpretation difficult.

Inference based on a closed form variance estimator can be thought of as a direct method and is, in general, hard to extend to general dependence structures. A safe choice and a more unified way of estimating the variance is by the use of some resampling scheme. Bootstrap and subsampling or their block versions are robust to slight changes in dependence structures and are more widely applicable. The literature along these lines is so vast to review and we refer the reader to Kunsch (1989), Liu and Singh (1992), Politis and Romano (1994), Lahiri (1999) for general block sampling techniques for variance/distribution estimation. Finite sample study of direct method is easy while such a study for resampling methods (under dependence) is yet non-existent.

4. OLS Estimator under Variable Selection

Having understood the properties of the OLS estimator obtained from the full set of covariates, we now proceed to the practically important aspect of OLS under variable selection. More often than not is the case that the set of covariates in the final reported model is not the same as the full set of covariates and more concernedly the final set of covariates is chosen based on the data at hand. For concreteness, let $\hat{M} \subseteq \{1, 2, \ldots, d\}$ represent the set of covariates selected and let $\hat{\beta}_{\hat{M}}$ represent the OLS estimator constructed based on covariate (indices) in $\hat{M}$. More generally for any set $M \subseteq \{1, 2, \ldots, d\}$, let $\hat{\beta}_M$ represent the OLS estimator

from covariates in M, that is,

$$\hat{\beta}_M := \underset{\theta \in \mathbb{R}^{|M|}}{\arg\min} \sum_{i=1}^{n} (Y_i - X_{i,M}^\top \theta)^2.$$

The aim of this section is to understand the properties of $\hat{\beta}_{\hat{M}}$ (irrespective of how $\hat{M}$ is chosen). This problem further highlights the strength of the deterministic inequality in Theorem 2.1 which applies irrespective of randomness of $\hat{M}$. Define for any $M \subseteq \{1, 2, \ldots, d\}$, the canonical "target" for OLS estimator $\hat{\beta}_M$ as

$$\beta_M := \underset{\theta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \mathbb{E}\left[(Y_i - X_{i,M}^\top \theta)^2\right].$$

Also, define $\mathcal{D}_M^\Sigma := \|\Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2} - I_{|M|}\|_{op}$. where recall Σ_M (and $\hat{\Sigma}_M$) represents the submatrix of Σ (and $\hat{\Sigma}$). Recall $(t)_+ := \max\{0, t\}$.

Corollary 4.1. *For any $\hat{M}$, we have*

$$\left\|\hat{\beta}_{\hat{M}} - \beta_{\hat{M}} - \Sigma_{\hat{M}}^{-1}(\hat{\Gamma}_{\hat{M}} - \hat{\Sigma}_{\hat{M}}\beta_{\hat{M}})\right\|_{\Sigma_{\hat{M}}} \leqslant \frac{\mathcal{D}_{\hat{M}}^\Sigma}{1 - \mathcal{D}_{\hat{M}}^\Sigma} \left\|\Sigma_{\hat{M}}^{-1}(\hat{\Gamma}_{\hat{M}} - \hat{\Sigma}_{\hat{M}}\beta_{\hat{M}})\right\|_{\Sigma_{\hat{M}}}.$$

More generally, for all $M \subseteq \{1, 2, \ldots, d\}$ (simultaneously), we have

$$\left\|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\right\|_{\Sigma_M} \leqslant \frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)_+} \left\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\right\|_{\Sigma_M}. \quad (16)$$

Corollary 4.1 follows immediately from Theorem 2.1 and for simplicity it is stated with Σ, β choices in (8) but other choices are possible. The first inequality in the corollary proves an influence function type expansion for the estimator $\hat{\beta}_{\hat{M}}$ around (a possibly random) target vector $\beta_{\hat{M}}$. In order to prove convergence of the remainder in this expansion to zero, one needs to control $\mathcal{D}_{\hat{M}}$ which can be a bit complicated to deal with directly. With some information on how "strongly" dependent $\hat{M}$ is on the data, such a direct approach can be worked out; see Russo and Zou (2016, Proposition 1), Jiao et al. (2018). If no information other than the fact that $\hat{M} \in \mathcal{M}$ for some set, $\mathcal{M}$, of subsets of covariates, then we have

$$\mathcal{D}_{\hat{M}} \leqslant U_{\hat{M}} \times \max_{M \in \mathcal{M}} \frac{\mathcal{D}_M}{U_M}, \quad (17)$$

for any set of (non-stochastic) numbers $\{U_M : M \in \mathcal{M}\}$; U_M usually converges to zero at rate $\sqrt{|M| \log(ed/|M|)/n}$; see Proposition 5.1. Some examples of $\mathcal{M}$ include $\mathcal{M}_{\leqslant k} := \{M \subseteq \{1, \ldots, d\} : 1 \leqslant |M| \leqslant k\}$, $\mathcal{M}_{=k} := \{M \subseteq \{1, \ldots, d\} : |M| = k\}$,

for some $k \geqslant 1$. Note that the maximum on the right hand side of (17) is random only through $\hat{\Sigma}$ (dissolving the randomness in $\hat{M}$ into the maximum over $\mathcal{M}$). We will take this indirect approach in our study since we do not want to make any assumption on how the model $\hat{M}$ is obtained which might as well be adversarial. Further note that (17) is tight (in that it cannot be improved) in an agnostic setting since one can take $\hat{M}$ such that $\mathcal{D}_{\hat{M}}/U_{\hat{M}} = \max_{M \in \mathcal{M}} \mathcal{D}_M/U_M$. We take the same indirect approach to bound $\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M}$ over $M \in \mathcal{M}$. These bounds prove consistency and linear representation error bounds for the OLS estimator under variable selection. Similar results can be derived for other modifications of OLS estimator such as transformations.

4.1. Consistency of $\hat{\beta}_{\hat{M}}$

From Corollary 4.1 it is easy to prove the following corollary (similar to Corollary 2.1) for consistency.

Corollary 4.2 (Consistency of $\hat{\beta}_{\hat{M}}$)**.** *If $\mathcal{D}^{\Sigma}_{\hat{M}} < 1$ and $\|\Sigma^{-1}_{\hat{M}}(\hat{\Gamma}_{\hat{M}} - \hat{\Sigma}_{\hat{M}}\beta_{\hat{M}})\| \to 0$ in probability, then $\|\hat{\beta}_{\hat{M}} - \beta_{\hat{M}}\|_{\Sigma_{\hat{M}}}$ converges to zero in probability.*

The conditions of Corollary 4.2 are reasonable and can be shown to hold under various dependence settings; see Kuchibhotla et al. (2018a). Under these conditions, we get that $\hat{\beta}_{\hat{M}}$ "converges" to $\beta_{\hat{M}}$ and hence under reasonable conditions, it is only possible to perform consistent asymptotic inference only for $\beta_{\hat{M}}$ based on $\hat{\beta}_{\hat{M}}$. In other words, if a confidence region is constructed for a parameter η centered at $\hat{\beta}_{\hat{M}}$ and that such region becomes a singleton asymptotically then $\|\eta - \beta_{\hat{M}}\|$ should converge to zero. In relation to the well-known consistent model selection literature, we can say if a claim is made about inference for β_{M_0} (for M_0 the true support) then $\|\beta_{\hat{M}} - \beta_{M_0}\|$ should converge to zero asymptotically.

4.2. Normal Approximation: Berry–Esseen result

From Corollary 4.1 (if $\mathcal{D}^{\Sigma}_{\hat{M}} \to 0$), we have

$$\hat{\beta}_{\hat{M}} - \beta_{\hat{M}} \;\approx\; \Sigma^{-1}_{\hat{M}}(\hat{\Gamma}_{\hat{M}} - \hat{\Sigma}_{\hat{M}}\beta_{\hat{M}}), \tag{18}$$

and hence inference for $\beta_{\hat{M}}$ requires understanding the asymptotic distribution of $\Sigma^{-1}_{\hat{M}}(\hat{\Gamma}_{\hat{M}} - \hat{\Sigma}_{\hat{M}}\beta_{\hat{M}})$ which is an average indexed by a random model $\hat{M}$. The impossibility results of Leeb and Pötscher (Leeb and Pötscher, 2008) imply that one cannot (uniformly) consistently estimate the asymptotic distribution of the

right hand side of (18). Hence the approach we take for inference is as follows: if we know apriori that $\hat{M}$ belongs on $\mathcal{M}$ either with probability 1 or with probability approaching 1, then by simultaneously inferring about β_M over all $M \in \mathcal{M}$ we can perform inference about $\beta_{\hat{M}}$. This is necessarily a conservative approach for any particular variable selection procedure leading to ($\hat{M}$ or) $\beta_{\hat{M}}$ but over all random models $\hat{M} \in \mathcal{M}$, this procedure is exact (or non-conservative); see Kuchibhotla et al. (2018b, Theorem 3.1). We acheive this simultaneous inference by using high-dimensional normal approximation results for averages of random vectors. Based on Corollary 4.1, we prove the following corollary (similar to Corollary 2.2).

Because of the finite sample nature (not requiring any specific structure), the result is cumbersome and requires some notation. We first briefly describe the method of proof of corollary to make the notation and result clear. We have already proved (16) for all $M \in \mathcal{M}$. Since Euclidean norm majorizes the maximum norm,

$$\max_{1 \leqslant j \leqslant |M|} |(\hat{\beta}_M - \beta_M)_j - (\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M))_j| \lesssim \frac{\mathcal{D}_M^{\Sigma} \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M}}{(1 - \mathcal{D}_M^{\Sigma})_+}.$$

Here we write $\lesssim$ since scaled Euclidean norm leads to other constant factors. We can use CLT for $(\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M))_{M \in \mathcal{M}}$ to compare $(\hat{\beta}_M - \beta_M)_{M \in \mathcal{M}}$ to a Gaussian counterpart. The CLT error term for the averages $(\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M))_{M \in \mathcal{M}}$ is defined as $\Delta_{n,\mathcal{M}}$. Here we also note that $\hat{\beta}_M - \beta_M$ is only close to the average upto an error term on the right hand side. This leads to two terms: first we need to show the right hand side term is indeed small for which we use CLT for scaled Euclidean norm (leading to $\Xi_{n,\mathcal{M}}$ below) and secondly, we need to account for closeness upto this small error which appears as probability of Gaussian process belonging in a small strip (leading to an anti-concentration term in the bound).

Now some notation. Let V_M represent the version of V in (12) for model M,

$$V_M \;:=\; \mathrm{Var}\left(n^{-1} \textstyle\sum_{i=1}^n X_{i,M}(Y_i - X_{i,M}^\top \beta_M)\right).$$

Note that $V_M = O(n^{-1})$, in general. Define the Gaussian process $(G_{M,j})_{M \in \mathcal{M}, 1 \leqslant j \leqslant |M|}$ with mean zero and the covariance operator given by: $\mathrm{Cov}(G_{M,j}, G_{M',j'})$ equals

$$\mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n \frac{(\Sigma_M^{-1} X_{i,M})_j (Y_i - X_{i,M}^\top \beta_M)}{(\Sigma_M^{-1} V_M \Sigma_M^{-1})_j^{1/2}}, \frac{1}{n}\sum_{i=1}^n \frac{(\Sigma_{M'}^{-1} X_{i,M'})_{j'} (Y_i - X_{i,M'}^\top \beta_{M'})}{(\Sigma_{M'}^{-1} V_{M'} \Sigma_{M'}^{-1})_{j'}^{1/2}}\right),$$

for all $M, M' \in \mathcal{M}$ and $1 \leqslant j \leqslant |M|, 1 \leqslant j' \leqslant |M'|$. Note $(G_{M,j})$ depends on n but the marginal variances are all 1. Let π_s for $1 \leqslant s \leqslant d$ represent the proportion

of models of size s in $\mathcal{M}$, that is, $\pi_s := \#\{M \in \mathcal{M} : |M| = s\}/|\mathcal{M}|$. Now set $D := \sum_{M\in\mathcal{M}} 5^{|M|}$ and define

$$\Xi_{n,\mathcal{M}} := \sup_{a\in\mathbb{R}_+^D} \left| \mathbb{P}\left(\left(\theta^\top V_M^{-\frac{1}{2}}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\right)_{\substack{M\in\mathcal{M},\\ \theta\in\mathcal{N}_{|M|}^{1/2}}} \preceq a \right) - \mathbb{P}\left(\left(\theta^\top \bar{G}_M\right)_{\substack{M\in\mathcal{M},\\ \theta\in\mathcal{N}_{|M|}^{1/2}}} \preceq a \right) \right|,$$

where $\preceq$ represents the vector coordinate-wise inequality, $\mathcal{N}_{|M|}^{1/2}$ represents the 1/2-net of $\{\theta \in \mathbb{R}^{|M|} : \|\theta\| \leqslant 1\}$, that is, $\min_{\theta'\in\mathcal{N}_{|M|}^{1/2}} \max_{\theta\in\mathbb{R}^{|M|}:\|\theta\|=1} \|\theta - \theta'\| \leqslant 1/2$, and $(\bar{G}_M)_{M\in\mathcal{M}}$ represents a Guassian process that has mean zero and shares the same covariance structure as $(V_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_{M\in\mathcal{M}}$. Note that $\mathrm{Var}(G_M) = I_{|M|}$ for any $M \in \mathcal{M}$. The quantity $\Xi_{n,\mathcal{M}}$ helps control one of the remainder factors, $\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}$. For the main term, define $C := \sum_{M\in\mathcal{M}} |M|$ and

$$\Delta_{n,\mathcal{M}} := \sup_{a\in\mathbb{R}_+^C} \left| \mathbb{P}\left(\left(\frac{|(\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j|}{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{1/2}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a \right) - \mathbb{P}\left((|G_{M,j}|)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a \right) \right|.$$

Corollary 4.3. *For all $M \subseteq \{1, 2, \ldots, d\}$, we have*

$$\|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M} \leqslant \frac{\mathcal{D}_M^\Sigma \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}}{(1 - \mathcal{D}_M^\Sigma)_+}.$$

Furthermore, for any $(\eta_M)_{M\in\mathcal{M}} \preceq 1/2$, we have

$$\begin{aligned}
&\sup_{a\in\mathbb{R}_+^C} \left| \mathbb{P}\left(\left(\frac{|(\hat{\beta}_M - \beta_M)_j|}{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{1/2}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a \right) - \mathbb{P}\left((|G_{M,j}|)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a \right) \right| \\
&\quad \leqslant \Delta_{n,\mathcal{M}} + 2.65\Xi_{n,M} + \mathbb{P}\left(\max_{M\in\mathcal{M}} \mathcal{D}_M^\Sigma/\eta_M \geqslant 1 \right) \\
&\qquad + \sup_{a\in\mathbb{R}_+^C} \mathbb{P}\left(\bigcup_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \left\{ ||G_{M,j}| - a_{M,j}| \leqslant 4\eta_M \sqrt{2\log\left(\frac{|\mathcal{M}|\pi_{|M|}5^{2|M|}}{\Xi_{n,M}} \right)} \right\} \right).
\end{aligned}$$

The proof of Corollary 4.3 can be found in Appendix B. The first inequality in Corollary 4.3 is slightly different from the conclusion of Corollary 4.1 but is more important for inference since the scaling in Corollary 4.3 is with respect to the "asymptotic" variance of $\hat{\beta}_M - \beta_M$. The second conclusion of Corollary 4.3 is a "randomness-free" version of finite sample Berry–Esseen type result for $(\hat{\beta}_M - \beta_M)$ simultaneously over all $M \in \mathcal{M}$. The terms each have a meaning and is explained

before the notation above. For a simpler result, consider the case of fixed (non-stochastic) covariates. In this case $\mathcal{D}_M^{\Sigma} = 0$ for all M and hence the result becomes

$$\left| \mathbb{P}\left(\left(\frac{|(\hat{\beta}_M - \beta_M)_j|}{(\Sigma_M^{-1} V_M \Sigma_M^{-1})_j^{1/2}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant |M|}} \preceq a \right) - \mathbb{P}\left((|G_{M,j}|)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant |M|}} \preceq a \right) \right| \leqslant \Delta_{n,\mathcal{M}} + 3\Xi_{n,\mathcal{M}},$$

for all $a \in \mathbb{R}_+^C$ since we can take η_M to be zero in limit. Getting back to the bound in Corollary 4.3, the quantities $\Delta_{n,\mathcal{M}}$ and $\Xi_{n,\mathcal{M}}$ can be easily controlled by using high-dimensional CLT results which only depend on the number of coordinates in the vector logarithmically. In particular for $\max\{\Delta_{n,\mathcal{M}}, \Xi_{n,\mathcal{M}}\} = o(1)$ they only require $\log(\sum_{M\in\mathcal{M}} |M|) = o(n^{\gamma})$ for some $\gamma > 0$ (Chernozhukov et al., 2017a, 2014; Zhang and Wu, 2017; Zhang and Cheng, 2014; Koike, 2019) for details. For instance, if $\mathcal{M} = \{M \subseteq \{1, \ldots, d\} : |M| \leqslant k\}$ then the requirement becomes $k \log(ed/k) = o(n^{\gamma})$. For the case of independent observations and sufficiently weakly dependent observations, we have

$$\max\{\Delta_{n,\mathcal{M}}, \Xi_{n,\mathcal{M}}\} = O(1) \left(n^{-1} \log^7 \left(\textstyle\sum_{M\in\mathcal{M}} |M| \right) \right)^{1/6}.$$

Bounds for $\mathbb{P}(\cup_{M\in\mathcal{M}} \{\mathcal{D}_M^{\Sigma} \geqslant \eta_M\})$ can be obtained using certain tail and "weak dependence" assumptions the covariates $X_1, \ldots, X_n$ (and as mentioned before one only needs to be concerned with the stochastic coordinates of covariates). This often necessitates exponential tails on the covariates if the total number of covariates d is allowed to grow almost exponentially with n (Guédon et al., 2015; Tikhomirov, 2017). Finally the control of the anti-concentration term (the last one in Corollary 4.3) only concerns a tail properties of a Gaussian process. A dimension dependent bound (that only depends logarithmically on dimension) for this probability can be found in (Nazarov, 2003; Chernozhukov et al., 2017b):

$$\mathbb{P}\left(\bigcup_{M\in\mathcal{M}, 1\leqslant j\leqslant |M|} \{||G_{M,j}| - a_{M,j}| \leqslant \varepsilon\} \right) \leqslant H\varepsilon \sqrt{\log\left(\textstyle\sum_{M\in\mathcal{M}} |M|\right)},$$

for some constant $H > 0$. Dimension-free bounds for this probability exist only for some special cases (Chernozhukov et al., 2015; Kuchibhotla et al., 2018). Regarding the constant in the anti-concentration probability, note that $\pi_{|M|}|\mathcal{M}| \leqslant (ed/|M|)^{|M|}$ for any collection $\mathcal{M}$and hence $\log(|M|\pi_{|M|} 5^{2|M|}/\Xi_{n,\mathcal{M}}) \leqslant |M| \log(25ed/\{|M|\Xi_{n,\mathcal{M}}\})$.

4.3. Inference under Variable Selection

Suppose that we can find $(\eta_M)_{M\in\mathcal{M}}$ such that $\mathbb{P}(\cup_{M\in\mathcal{M}}\{\mathcal{D}_M^\Sigma \geqslant \eta_M\})$ and the anti-concentration term goes to zero, then from Corollary 4.3 we get that

$$\mathbb{P}\left(\left((\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{-1/2}|(\hat{\beta}_M-\beta_M)_j|\right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a\right) \approx \mathbb{P}\left((|G_{M,j}|)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a\right),$$

uniformly for all $a\in\mathbb{R}^{\sum_{M\in\mathcal{M}}|M|}$. In order to perform inference (or in particular confidence regions) one can choose a vector $a=a_\alpha$ such that

$$\mathbb{P}\left((|G_{M,j}|)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \preceq a_\alpha\right) = 1-\alpha. \tag{19}$$

This implies that for any $\hat{M}\in\mathcal{M}$ chosen (possibly) randomly based on the data,

$$\mathbb{P}\left(\left((\Sigma_{\hat{M}}^{-1}V_{\hat{M}}\Sigma_{\hat{M}}^{-1})_j^{-1/2}|(\hat{\beta}_{\hat{M}}-\beta_{\hat{M}}|)_j\right)_{1\leqslant j\leqslant|\hat{M}|} \leqslant (a_\alpha)_{\hat{M}}\right) \geqslant 1-\alpha+o(1),$$

asymptotically. This means that with (asymptotic) probability of at least $1-\alpha$, $\beta_{\hat{M},j}$ belongs in the interval $[\hat{\beta}_{\hat{M},j}\pm(a_\alpha)_{\hat{M},j}(\Sigma_{\hat{M}}^{-1}V_{\hat{M}}\Sigma_{\hat{M}}^{-1})_j^{1/2}]$ simultaneously for all $1\leqslant j\leqslant|\hat{M}|$. If no variable selection is involved and no simultaneity over $1\leqslant j\leqslant|\hat{M}|$ is required, then $(a_\alpha)_{\hat{M},j}$ would just be $z_{\alpha/2}$ (the usual normal quantile for a $(1-\alpha)$-confidence interval). This is the essential point of post-selection inference wherein we enlarge the usual confidence intervals to make them simultaneous.

The above discussion completes inference for the OLS estimator under variable selection for all types of observations (that allow for a CLT: $\Delta_{n,\mathcal{M}}\asymp\Xi_{n,\mathcal{M}}\asymp 0$) except for two important points: firstly, we have proved the CLT result with the true "asymptotic" variance $\Sigma_M^{-1}V_M\Sigma_M^{-1}$ which is unknown in general; it is, however, easy to estimate this variance using the techniques described in Section 3. Secondly and more importantly, there are infinitely many different choices of a_α satisfying (19); what is the right choice? The first problem is easy to rectify in that if a variance estimator $\hat{\sigma}_{M,j}$ (for $(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{1/2}$) has a good enough rate of convergence with respect to the metric $|\hat{\sigma}_{M,j}/(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{1/2}-1|$ uniformly over all $M\in\mathcal{M}, 1\leqslant j\leqslant|M|$ then it is easy to prove a version of Corollary 4.3 with the unknown variance replaced by the estimator in the first probability.

Related to the choice of $(a_\alpha)_{M\in\mathcal{M},1\leqslant j\leqslant|M|}$, in the path-breaking work Berk et al. (2013), the authors have used $(a_\alpha)=a\mathbf{1}$ for some constant a, which means that the simultaneous inference is based on quantiles of the maximum statistic:

$$\max_{M\in\mathcal{M}}\max_{1\leqslant j\leqslant|M|}(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{-1/2}|(\hat{\beta}_M-\beta_M)_j|. \tag{20}$$

Berk et al. (2013) assumed non-stochastic covariates and an independent homoscedsatic Gaussian model for the response. This statistic was also adopted in Bachoc et al. (2016) where the framework was generalized to the case of non-Gaussian responses (but with non-stochastic covariates); further both works require the total number of covariates to be fixed and not change with n. The analysis above does not require either of these conditions since our results are *deterministic*. Hence

$$\mathbb{P}\left(\max_{M\in\mathcal{M},1\leqslant j\leqslant|M|}|G_{M,j}|\leqslant K(\alpha)\right)=1-\alpha, \tag{21}$$

implies for any $\hat{M}$ such that $\mathbb{P}(\hat{M}\in\mathcal{M})=1$, we have asymptotically

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant|\hat{M}|}\left|(\Sigma_{\hat{M}}^{-1}V_{\hat{M}}\Sigma_{\hat{M}}^{-1})_j^{-1/2}(\hat{\beta}_{\hat{M}}-\beta_{\hat{M}})_j\right|\leqslant K(\alpha)\right)\geqslant 1-\alpha.$$

The quantile $K(\alpha)$ in (21) can be computed by bootstrapping the maximum statistic using the linear representation result; see Belloni et al. (2018), Deng and Zhang (2017) and Zhang and Cheng (2014) for details on bootstrap for independent/dependent summands in averages.

The maximum statistic in (20) (used in Berk et al. (2013) and Bachoc et al. (2016)) is only one of the many different ways of performing valid post-selection inference. It is clear that if for some $\alpha\in[0,1]$ and numbers $\{K_M(\alpha):M\in\mathcal{M}\}$,

$$\mathbb{P}\left(\bigcap_{M\in\mathcal{M}}\left\{\max_{1\leqslant j\leqslant|M|}|G_{M,j}|\leqslant K_M(\alpha)\right\}\right)=1-\alpha, \tag{22}$$

then we have

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant|\hat{M}|}\left|(\Sigma_{\hat{M}}^{-1}V_{\hat{M}}\Sigma_{\hat{M}}^{-1})_j^{-1/2}(\hat{\beta}_{\hat{M}}-\beta_{\hat{M}})_j\right|\leqslant K_{\hat{M}}(\alpha)\right)\geqslant 1-\alpha+o(1), \tag{23}$$

for any $\hat{M}$ (possibly random) such that $\mathbb{P}(\hat{M}\in\mathcal{M})=1$ (this equality can be relaxed to convergence to 1). Inequality (23) readily implies (asymptotically valid) post-selection confidence region for $\beta_{\hat{M}}$ as

$$\hat{\mathcal{R}}_{\infty,\hat{M}}:=\left\{\theta\in\mathbb{R}^{|\hat{M}|}:\max_{1\leqslant j\leqslant|\hat{M}|}\left|(\Sigma_{\hat{M}}^{-1}V_{\hat{M}}\Sigma_{\hat{M}}^{-1})_j^{-1/2}(\hat{\beta}_{\hat{M},j}-\theta_j)\right|\leqslant K_{\hat{M}}(\alpha)\right\}.$$

Note that the confidence regions or more generally inference obtained from the maximum statistic corresponds to taking $(K_M(\alpha))_{M\in\mathcal{M}}$ in (22) to be a constant multiple of $(1)_{M\in\mathcal{M}}$ (all 1's vector). Further note that the event in (22) represents a specific choice of vector a_α in (19) for which Corollary 4.3 applies. Before we discuss how to choose $(K_M(\alpha))_{M\in\mathcal{M}}$, we list out some of the disadvantages of using the maximum statistic (20).

Disadvantages of the maximum statistic. The maximum statistic is a natural generalization of inference for a single model to simultaneous inference over a collection of models. The maximum statistic would be the right thing to do if we are concerned with simultaneous inference for p parameters (all of which are of same order) but this is not the case with OLS under variable selection. It is intuitively expected that models with more number of covariates would have larger width intervals. For this reason by taking the maximum over the collection $\mathcal{M}$ of models, one is ignoring the smaller models and the fact that small models have smaller width confidence intervals. To be concrete, if $\mathcal{M}$ is $\mathcal{M}_{\leqslant k}$ it follows from the results of Berk et al. (2013); Zhang (2017) that

$$\max_{M\in\mathcal{M}} \max_{1\leqslant j\leqslant |M|} |G_{M,j}| = O_p(\sqrt{k\log(ed/k)}), \tag{24}$$

and in the worst case this rate can be attained. But if $k = 40$ (for example) but the selected model $\hat{M}$ happened to have only two covariates, then the confidence interval is (unnecessarily) wider by a factor of $\sqrt{20}$. By allowing model dependent quantile $K_M(\alpha)$ as in (22) we can tighten confidence intervals appropriately. For this particular disadvantage, it is enough to have $K_M(\alpha)$ depend on M only through $|M|$, its size. There is a second disadvantage of the maximum statistic that requires dependence of $K_M(\alpha)$ on the covariates in M.

To describe the second disadvantage we look at the conditions under which worst case rate in (24) is attained when $k = d$. Berk et al. (2013, Section 6.2) shows that if the covariates are non-stochastic, and

$$\hat{\Sigma} := \begin{bmatrix} I_{d-1} & c\mathbf{1}_{d-1} \\ \mathbf{0}_{d-1}^\top & \sqrt{1-(d-1)c^2} \end{bmatrix}, \text{ for some } c^2 < 1/(d-1),$$

then there exists a constant $\mathfrak{C} > 0$, such that

$$\max_{M\in\mathcal{M}_{\leqslant d}} \max_{1\leqslant j\leqslant |M|} |G_{M,j}| \geqslant \mathfrak{C}\sqrt{d}. \tag{25}$$

Now define $\mathcal{M} = \{M \subseteq \{1,\ldots,d\} : M \subseteq \{1,\ldots,d-1\}\}$, that is, $\mathcal{M}$ is the collection of models that only contain the first $d-1$ covariates. It now follows from (Berk et al., 2013, Section 6.1) that

$$\max_{M\in\mathcal{M}} \max_{1\leqslant j\leqslant |M|} |G_{M,j}| \asymp \sqrt{\log(ed)}. \tag{26}$$

Comparing (25) and (26), it is clear that the inclusion of the last covariate increases the order of the maximum statistic from $\sqrt{\log(ed)}$ to $\sqrt{d}$; this shift is because of

increased collinearity. This means that if in the selection procedure we allow all models but end up choosing the model that only contains the first $d-1$ covariates, we pay of lot more price than necessary. Note that if d increases with n, this increase (in rate) could hurt more. Once again allowing for $K_M(\alpha)$ a model dependent quantile for maximum (over j) in that model resolves this disadvantage.

How to choose $K_M(\alpha)$? Now that we have understood the need for model M dependent quantiles $K_M(\alpha)$, it remains to decide how to find these quantiles. But first note that these are not uniquely defined because multivariate quantiles are not unique. We do not yet know of an "optimal" construction of $K_M(\alpha)$ and we describe a few choices below motivated by multi-scale testing literature (Dumbgen and Spokoiny, 2001; Datta and Sen, 2018). Before we proceed to this, we note an impossibility on uniform improvement over the maximum statistic. Suppose we select a (random) model $\hat{M}$ such that

$$\max_{1\leqslant j\leqslant|\hat{M}|}\left|(\hat{\beta}_{\hat{M},j}-\beta_{\hat{M},j})/\sigma_{\hat{M},j}\right| \;=\; \max_{M\in\mathcal{M}}\max_{1\leqslant j\leqslant|M|}\left|(\hat{\beta}_{M,j}-\beta_{M,j})/\sigma_{M,j}\right|,$$

where $\sigma_{M,j}$ represents the standard deviation, $(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j^{1/2}$, of $\hat{\beta}_{M,j}-\beta_{M,j}$. For this random model $\hat{M}$, $K(\alpha)$ the quantile of the maximum statistic in (21) leads to the smallest possible rectangular confidence region for $\beta_{\hat{M}}$. This implies that $K_{\hat{M}}(\alpha)\geqslant K(\alpha)$ for any $\alpha\in[0,1]$ and any sequence $(K_M(\alpha))_{M\in\mathcal{M}}$. Therefore no sequence of quantiles $(K_M(\alpha))_{M\in\mathcal{M}}$ satisfying (22) can improve on $K(\alpha)$ uniformly over $M\in\mathcal{M}$; any gain for some model is paid for by a loss for some other model. The hope is that the gain outweighs the loss and we see this in our simulations.

Getting back to the construction of $K_M(\alpha)$, let the maximum for model M be

$$T_M := \max_{1\leqslant j\leqslant|M|}\left|(\hat{\beta}_{M,j}-\beta_{M,j})/\hat{\sigma}_{M,j}\right|,$$

for an estimator $\hat{\sigma}_{M,j}$ of the standard deviation $\sigma_{M,j}$; recall $\sigma_{M,j}$ involves V_M that converges to zero. Recall that the maximum statistic (20) is given by $\max_{M\in\mathcal{M}}T_M$. We now present three choices that will lead to three different quantiles $K_M(\alpha)$.

1. In order to take into account the set of covariates in M, we center T_M by its median before taking the maximum:

$$\max_{M\in\mathcal{M}}\left\{T_M-\texttt{med}(T_M)\right\}, \tag{27}$$

   where $\texttt{med}(\cdot)$ represents the median. One can center by the mean of T_M but estimation of mean of a maximum using bootstrap is not yet clear. Higher

collinearity between the covariates in M could increase the order of T_M, the effect of which we avoid spilling into other models by centering by the median. Also, it is clear that the median of T_M has order depending only on M not the maximum model size in collection $\mathcal{M}$. Further it is well-known that the maximum of Gaussians exhibit a super-concentration phenomenon in that their variance decreases to zero as the number of entries in the maximum goes to infinity. For this reason, it may not be of importance to scale by the standard deviation of T_M. If $K^{(1)}_{\mathcal{M}}(\alpha)$ represents the quantile of the statistic (27), then the post-selection confidence intervals are given by

$$\hat{\mathcal{R}}^{(1)}_M := \left\{ \theta \in \mathbb{R}^{|M|} : \max_{1 \leqslant j \leqslant |M|} |(\hat{\beta}_{M,j} - \theta_j)/\hat{\sigma}_{M,j}| \leqslant \widehat{\texttt{med}}(T_M) + K^{(1)}_{\mathcal{M}}(\alpha) \right\}.$$

2. The super-concentration of the maximum of Gaussians holds only under certain "strong uncorrelatedness" assumption. Following the previous suggestion, we can normalize the centered T_M by its median absolute deviation (MAD) to account for the variance:

$$\max_{M \in \mathcal{M}} \{T_M - \texttt{med}(T_M)\}/\texttt{MAD}(T_M), \tag{28}$$

where $\texttt{MAD}(T_M) := \texttt{med}(|T_M - \texttt{med}(T_M)|)$. If $K^{(2)}_{\mathcal{M}}(\alpha)$ represents the quantile of the statistic (28), then the post-selection confidence intervals are given by

$$\hat{\mathcal{R}}^{(2)}_M := \left\{ \theta \in \mathbb{R}^{|M|} : \max_{1 \leqslant j \leqslant |M|} |(\hat{\beta}_{M,j} - \theta_j)/\hat{\sigma}_{M,j}| \leqslant \widehat{\texttt{med}}(T_M) + \widehat{\texttt{MAD}}(T_M) K^{(2)}_{\mathcal{M}}(\alpha) \right\}.$$

3. Now that we have centered and scaled T_M with its median and MAD, it is expected that even for models of different sizes, $(T_M - \texttt{med}(T_M))/\texttt{MAD}(T_M)$ are of the same order. However, when we take the maximum over all models of same size they may not be. The reason for this is the maximum over models of size 1 involves d terms and the maximum over models of size 2 involves $d(d-1)/2$ terms. Hence naturally the maximum over models of size 2 is expected to be bigger. To account for this discrepancy define the centered and scaled maximum statistic for model size s as

$$\mathfrak{T}_s := \max_{|M|=s} \{T_M - \texttt{med}(T_M)\}/\texttt{MAD}(T_M),$$

and take quantile of

$$\max_{1 \leqslant s \leqslant k} \{\mathfrak{T}_s - \texttt{med}(\mathfrak{T}_s)\}. \tag{29}$$

If $K^{(3)}_{\mathcal{M}}(\alpha)$ represents the quantile of the statistic (29), then the post-selection confidence intervals are given by

$$\hat{\mathcal{R}}^{(3)}_M := \left\{ \theta : \max_{1 \leqslant j \leqslant |M|} \left| \frac{\hat{\beta}_{M,j} - \theta_j}{\hat{\sigma}_{M,j}} \right| \leqslant \widehat{\texttt{med}}(T_M) + \widehat{\texttt{MAD}}(T_M)[K^{(3)}_{\mathcal{M}}(\alpha) + \widehat{\texttt{med}}(\mathfrak{T}_{|M|})] \right\}.$$

We emphasize once again that even though these choices improve the width of confidence intervals for some models, they will deteriorate the width for other models. We will see from the simulations in Section 6 that the gain (for some models) outweighs the loss (for other models) in width. All the choices above involve $\texttt{med}(T_M)$, $\texttt{MAD}(T_M)$ which are simple functions of quantiles and can be computed readily from bootstrap procedures mentioned above.

5. Rates under Independence

All the theoretical analysis in previous sections is deterministic and the complete study in any specific setting requires bounding the remainder terms in the deterministic inequalities above. In this section, we complete the program by bounding the remainder terms in case of independent observations. The two main quantities that need bounding for Theorem 2.1 are

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_{op} \quad \text{and} \quad \|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_{\Sigma} = \|\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)\|.$$

The concentration of the sample covariance matrix to its expectation has been the study for decades documented in the works of Vershynin (2012, 2018), Rudelson and Zhou (2013), Guédon et al. (2015), Tikhomirov (2017). We state here the result from Tikhomirov (2017) with minimal tail assumptions that we know of.

Theorem 5.1 (Theorem 1.1 of Tikhomirov (2017))**.** *Fix $n \geqslant 2d$ and $p \geqslant 2$. If $X_1, \ldots, X_n$ are centered iid random vectors satisfying: for some $B \geqslant 1$,*

$$\mathbb{E}|a^{\top}\Sigma^{-1/2}X|^p \leqslant B^p \quad \textit{for all} \quad a \in \mathbb{R}^d, \ \textit{with} \ \|a\| = 1. \tag{30}$$

Then there exists a constant $K_p > 0$ with probability at least $1 - 1/n$,

$$\mathcal{D}^{\Sigma} \leqslant \frac{K_p}{n} \max_{1 \leqslant i \leqslant n} \|\Sigma^{-1/2}X_i\|^2 + K_p B^2 \left(\frac{d}{n}\right)^{1-2/p} \log^4\left(\frac{n}{d}\right) + K_p B^2 \left(\frac{d}{n}\right)^{1-2/\min\{p,4\}}.$$

The random quantity on the right hand side can be bounded using appropriate bounds on $\mathbb{E}[\|\Sigma^{-1/2}X_i\|^{2q}/d^q]$ for some $q \geqslant 1$. Assuming the first term can be ignored compared to the others, we get that $\mathcal{D}^{\Sigma}$ converges to zero as long as

$d = o(n)$ when the covariates have at least $(2+\delta)$-moments. Further if $p \geqslant 4$, then $\mathcal{D}^{\Sigma} = O_p(\sqrt{d/n})$. Regarding the term $\|\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)\|$, we have

$$\mathbb{E}\|\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)\| \leqslant \sqrt{\operatorname{tr}(\operatorname{Var}(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)))} = \frac{(\mathbb{E}[\|\Sigma^{-1/2}X\|^2(Y - X^{\top}\beta)^2])^{1/2}}{\sqrt{n}}.$$

Hence if $\mathbb{E}[\|\Sigma^{1/2}X\|^2(Y - X^{\top}\beta)^2] = O(d)$, then we get $\|\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta)\| = O_p(\sqrt{d/n})$. Combining these calculations with Theorem 2.1, we get

$$\|\hat{\beta} - \beta\|_{\Sigma} = O_p(1)\sqrt{\frac{d}{n}} \quad \text{and} \quad \|\hat{\beta} - \beta - \Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_{\Sigma} = O_p(1)\frac{d}{n},$$

allowing for d growing with the sample size n; consistency holds when $d = o(n)$ and asymptotic normality holds when $d = o(\sqrt{n})$. Asymptotic analysis for $d/n \to \kappa \in [0, 1)$ can be done with more stringent conditions on the observations.

Regarding Corollary 4.1, we need to control *simultaneously* over $M \in \mathcal{M}$,

$$\mathcal{D}^{\Sigma}_M = \|\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2} - I_{|M|}\|_{op} \quad \text{and} \quad \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|. \tag{31}$$

This simultaneous control often necessitates exponential tails for covariates if one needs to allow d to grow (almost exponentially) with n. Guédon et al. (2015) provide sharp results for $\sup_{|M| \leqslant k} \|\hat{\Sigma}_M - \Sigma_M\|_{op}$ for both polynomial and exponential tails on covariates. We do not know such sharp results for $\sup_{M \in \mathcal{M}} \mathcal{D}^{\Sigma}_M$. By a simple union bound the following result can be proved for both quantities in (31). For this we assume the following extension of (30): For all $1 \leqslant i \leqslant n$,

$$\mathbb{E}\left[\exp\left(\frac{|a^{\top}X_i|^{\beta}}{\mathfrak{K}_{\beta}^{\beta}\|a\|_{\Sigma}^{\beta}}\right)\right] \leqslant 2, \quad \text{for some } \beta > 0,\ 0 < \mathfrak{K}_{\beta} < \infty \text{ and for all } a \in \mathbb{R}^d. \tag{32}$$

Condition (32) is same as sub-Gaussianity if $\beta = 2$ and is same as sub-exponentiality if $\beta = 1$. With $\beta = \infty$, it becomes a boundedness condition. If X_i's satisfy condition (32) with $\beta < 1$ then their moment generating function may not exist but they still exhibit "weak" exponential tails. Additionally note that (32) does not require Σ to be invertible and it implies that for all $M \subseteq \{1, 2, \ldots, d\}$,

$$\mathbb{E}\left[\exp\left(\mathfrak{K}_{\beta}^{-\beta}|a^{\top}\Sigma_M^{-1/2}X_{i,M}|^{\beta}\right)\right] \leqslant 2 \quad \text{for all} \quad a \in \mathbb{R}^{|M|} \text{ such that } \|a\| = 1. \tag{33}$$

Define the kurtosis and "regression variance" for model M as

$$\kappa^{\Sigma}_M := \max_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n}\sum_{i=1}^{n} \frac{\operatorname{Var}((X_{i,M}^{\top}\theta)^2)}{\|\Sigma_M^{1/2}\theta\|^4} \quad \text{and} \quad \mathfrak{V}_M := \max_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n}\sum_{i=1}^{n} \operatorname{Var}(\theta^{\top}\Sigma_M^{-1/2}X_{i,M}Y_i).$$

Assume the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are just independent.

Proposition 5.1. *Fix any $t \geqslant 0$. Under (33), we have with probability at least $1 - 3e^{-t}$, simultaneously for any $1 \leqslant s \leqslant d$, for any $M \subseteq \{1, \ldots, d\}$ with $|M| = s$,*

$$\mathcal{D}_M^{\Sigma} \leqslant 14\sqrt{\frac{\kappa_M^{\Sigma}(t + s\log(9e^2 d/s))}{n}} + \frac{C_\beta \mathfrak{K}_\beta^2 (\log(2n))^{2/\beta}(t + s\log(9e^2 d/s))^{\max\{1,2/\beta\}}}{n}. \tag{34}$$

If (33) and $\mathbb{E}[Y_i^r] \leqslant K_{n,r}^r$ for some $r \geqslant 2$ hold true, then with probability at least $1 - 3e^{-t} - t^{-r+1}$, for any $1 \leqslant s \leqslant d$, for any model $M \subseteq \{1, \ldots, d\}$ with $|M| = s$,

$$\begin{aligned} \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\| \leqslant 14\sqrt{\frac{\mathfrak{V}_M(t + s\log(5e^2 d/s))}{n}} + \mathcal{D}_M^{\Sigma}(\textstyle\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2} \\ + \frac{C_\beta K_{n,r} \mathfrak{K}_\beta (\log(2n))^{1/\beta}(t + s\log(5e^2 d/s))^{\max\{1,1/\beta\}}}{n^{1-1/r}} \\ + \frac{t C_{\beta,r} K_{n,r} \mathfrak{K}_\beta (s\log(5e^2 d/s) + \log n)^{1/\beta}}{n^{1-1/r}}, \end{aligned} \tag{35}$$

for some constants $C_\beta, C_{\beta,r} > 0$ depending only on β and (β, r), respectively.

The proof of Proposition 5.1 can be found in Appendix C. Note that the rates for $\mathcal{D}_M^{\Sigma}$ and for $\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|$ scale with $|M|$ (and only logarithmically on the total number of covariates d) and we did not just bound $\max_{|M| \leqslant k} \mathcal{D}_M^{\Sigma}$. This is what we tried to replicate in a data-driven way from the post-selection confidence regions in Page 19 by centering with quantities depending on M. Ignoring the lower order terms, we have uniformly over all $M \subseteq \{1, 2, \ldots, d\}$,

$$\max\{\mathcal{D}_M^{\Sigma},\ \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|\} = O_p(1)\sqrt{\frac{|M|\log(ed/|M|)}{n}},$$

which also provides η_M for an application of Corollary 4.3. It is noteworthy that we only require finite number of moments on the response.

6. Simulation Results

We consider three different settings and compare different ways of post-selection inference as described in Page 19. We only consider the case of fixed design under the well-specified linear model (that unfortunately goes against the philosophy of the paper) which we do since the lower bound and worst case results in post-selection inference are only available for fixed design case (Berk et al., 2013). The fixed design for each of the cases are as follows:

(a) **Orthogonal design.** We take $x_1, \ldots, x_n$ such that $\hat{\Sigma} = n^{-1}\sum_{i=1}^n x_i x_i^\top = I_d$. We find the x_i by first taking a matrix $\mathcal{X} \in \mathbb{R}^{n \times d}$ satisfying $\mathcal{X}^\top \mathcal{X} = I_d$ and then multiply this matrix with $\hat{\Sigma}^{1/2}$ (which in this setting is I_d).

(b) **Exchangeable design.** We take $x_1, \ldots, x_n$ such that $\hat{\Sigma} = I_d + \alpha \mathbf{1}_d \mathbf{1}_d^\top$ with $\alpha = -1/(d+2)$. Here $\mathbf{1}_d$ is the all 1's vector of dimension d.

(c) **Worst-case design.** We take $x_1, \ldots, x_n$ such that

$$\hat{\Sigma} := \begin{bmatrix} I_{d-1} & c\mathbf{1}_{d-1} \\ \mathbf{0}_{d-1}^\top & \sqrt{1-(d-1)c^2} \end{bmatrix}, \text{ with } c^2 = \frac{1}{2(d-1)},$$

For the first two settings, it is known that the maximum statistic (20) is of order $\sqrt{\log(d)}$ and for the last setting it is known that it is of order $\sqrt{d}$ (where we hope the other ways of PoSI would help improve the confidence intervals). See Berk et al. (2013, Section 6) for details. For each setting, the model is $Y_i = x_i^\top \beta_0 + \varepsilon_i$, with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \sigma = 1$ and β_0 is randomly generated as a vector with each coordinate being a $\texttt{Unif}(-1, 1)$ independently. We consider $d = 20, \mathcal{M} = \mathcal{M}_{\leqslant 10}, \alpha = 0.05$ (confidence level is 0.95). Even though the variance of $\hat{\beta}_M - \beta_M$ is $\sigma^2 \hat{\Sigma}_M^{-1}$, we estimate it by using (13) ignoring the Gaussian response knowledge.

We report the simulations in the following way: For all designs, we split all models in $\mathcal{M}$ into models of different sizes. We compute the (average over 500 simulations) coverage for all models of a given size and minimum, median as well as maximum (average) confidence interval length for that model size, that is,

$$\mathbb{P}\left(\bigcap_{M \in \mathcal{M}, |M| = s} \{\beta_M \in \hat{\mathcal{R}}_M\} \right), \quad \left\{ \min_{|M|=s}, \texttt{med}_{|M|=s}, \max_{|M|=s} \right\} \mathbf{m}(\hat{\mathcal{R}}_M), \tag{36}$$

are reported with $\hat{\mathcal{R}}_M$ replaced by $\hat{\mathcal{R}}_M^{(j)}, 1 \leqslant j \leqslant 3$ given in Page 19, where $\mathbf{m}(\hat{\mathcal{R}}_M)$ represents the threshold of the confidence region for model M (e.g., for $\hat{\mathcal{R}}_M^{(1)}$ it is $\widehat{\texttt{med}}(T_M) + K_{\mathcal{M}}^{(1)}(\alpha)$). Note that this threshold is a proxy for the volume of the confidence region. Additionally we consider $\hat{\mathcal{R}}_M^{(0)}$ given by

$$\left\{ \theta \in \mathbb{R}^{|M|} : \max_j \left| \frac{\hat{\beta}_{M,j} - \theta_j}{\hat{\sigma}_{M,j}} \right| \leqslant K_{\mathcal{M}}^{(0)}(\alpha) \right\} \text{ with } K_{\mathcal{M}}^{(0)}(\alpha) := (1-\alpha)\text{-quantile} \left(\max_{M \in \mathcal{M}} T_M \right).$$

Finally we also report $\mathbb{P}(\cap_{M \in \mathcal{M}} \{\beta_M \in \hat{\mathcal{R}}_M\})$. Note that by construction this probability has to be about 0.95 and by noting the first quantity in (36), we see if the constructed confidence regions are too conservative for models of smaller sizes. Table 1 shows the average coverage from all methods in all settings confirming that these are valid post-selection confidence regions.

Figures 1, 2, and 3 show the results (for settins (a), (b) and (c), respectively) from 500 simulations within each 200 bootstrap samples were used. In all the settings, the coverage from the proposed methods ($\hat{\mathcal{R}}_M^{(j)}, j = 1, 2, 3$) is closer to 0.95 and for many models the proposed intervals are shorter than the ones from $\hat{\mathcal{R}}_M^{(0)}$.

method	Setting (a)	Setting (b)	Setting (c)
method 0	0.986	0.976	0.972
method 1	0.964	0.960	0.964
method 2	0.964	0.956	0.958
method 3	0.964	0.956	0.958

TABLE 1

The numbers in table represent the average simultaneous coverage of all confidence regions for all settings estimate of $\mathbb{P}(\cap_{M\in\mathcal{M}}\{\beta_M \in \hat{\mathcal{R}}_M\})$ *based on 500 replications.*

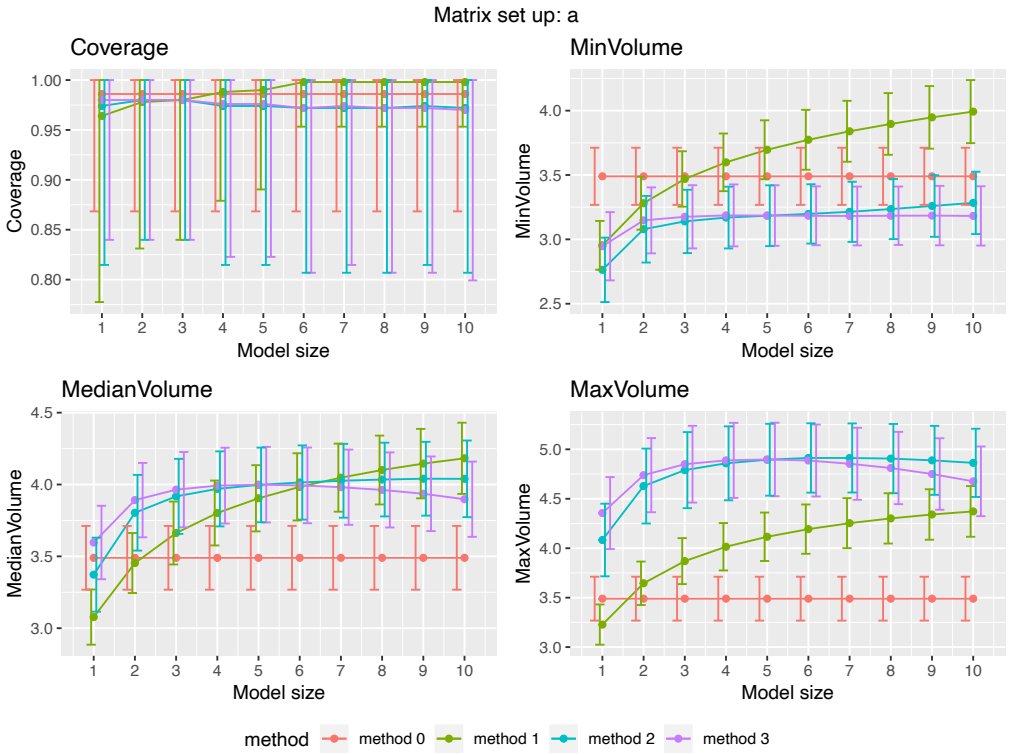

Fig 1: The results for setting (a) with orthogonal design. The lines represents average (of quantities in (36)) over 500 replications and error bars are ± 1 SD over replications. Method j in legend refers to confidence regions $\hat{\mathcal{R}}_M^{(j)}$ for $j = 0, 1, 2, 3$. Volume in the plots refers to the threshold $\mathbf{m}(\hat{\mathcal{R}}_M)$.

7. Summary and Final Word

We have provided a completely deterministic study of ordinary least squares linear regression setting which implies asymptotic normality, inference, inference under variable selection and much more without requiring any of the classical model assumptions. This study brings out two important quantities that needs to be controlled for a complete study of the OLS estimator. We control these quantities

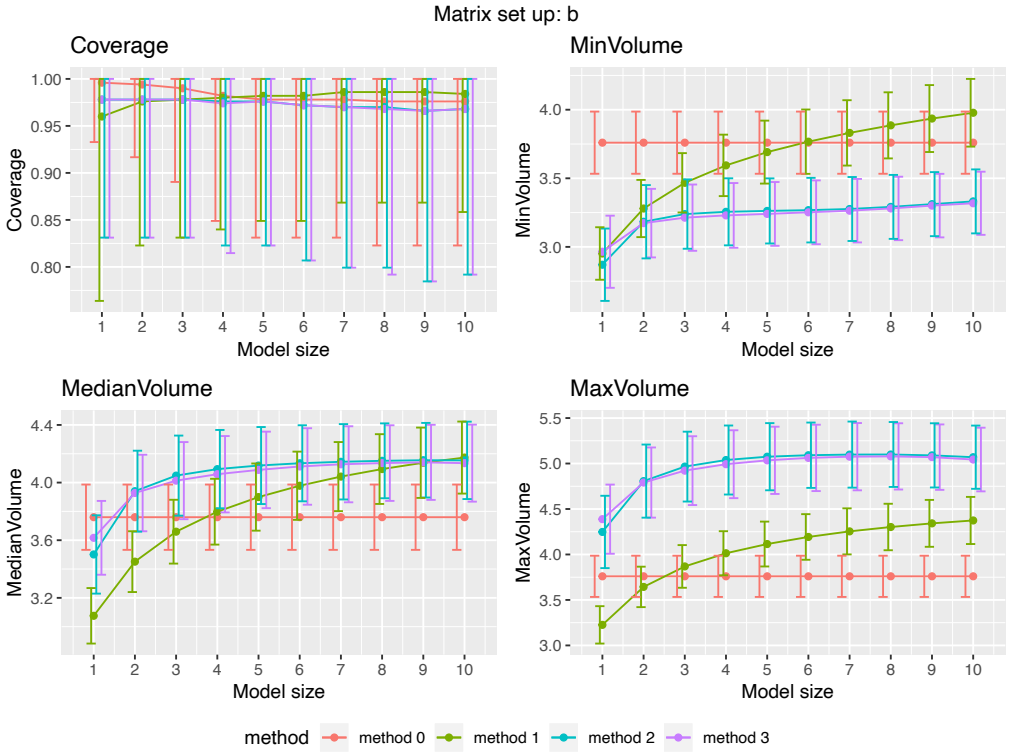

Fig 2: The results for setting (b) with exhangeable design.

in case of independent observations allowing for the total number of covariates to diverge with the sample size (almost exponentially).

We have shown through our results here that the study of an estimator can be split into two parts. One that leads to (deterministic) inequalities that hold for any set of observations and one that requires assumptions on data generating process to control the remainder terms in the deterministic inequalities or Berry–Esseen type results or (more importantly) for inference. We have extensively studied the first part in this paper and the second part (inferential part) needs to be understood more carefully when the observations are dependent; the references mentioned about block bootstrap/resampling techniques would be a starting point but rates in finite samples with increasing dimensions needs to be understood.

In the later part of the paper, we have focused on OLS under variable selection. From the derivation it should be clear that variable selection is just a choice we made and one can easily study OLS under transformations of response and/or covariates using the deterministic inequality.

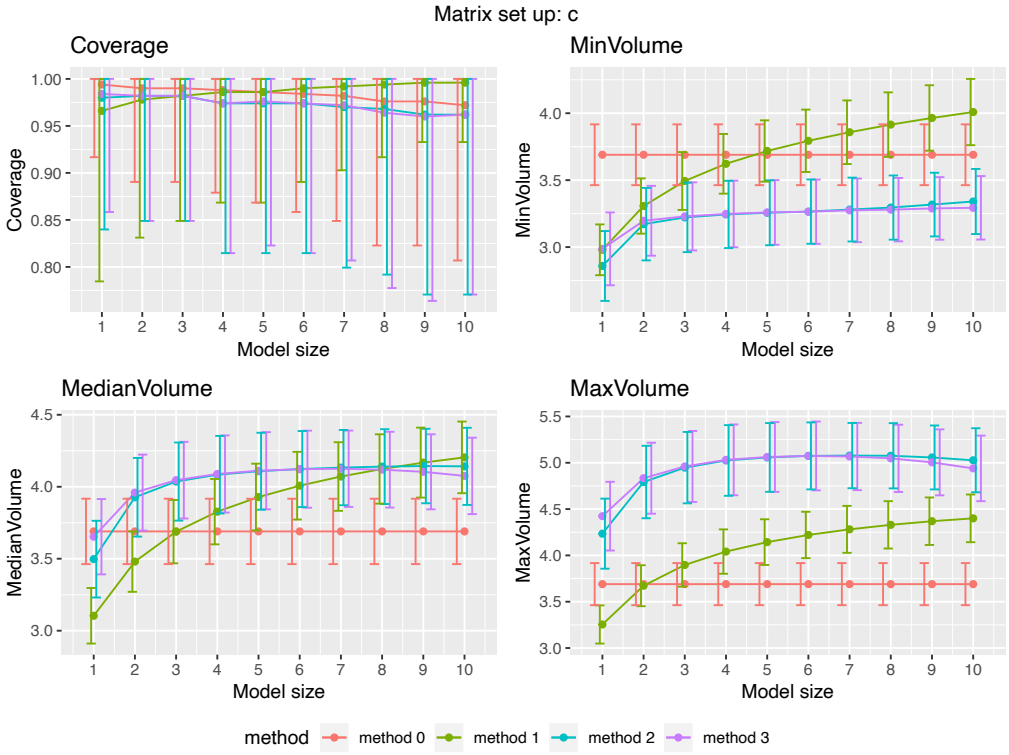

Fig 3: The results for setting (c) with worst case design.

We have chosen to study the OLS linear regression estimator because of its simplicity; even in this case some calculations get messy. An almost parallel set of results can be derived for other regression estimators including GLMs, Cox proportional hazards model and so on; see Kuchibhotla (2018) for details.

Finally we close with some comments on computation. The methods of inference after variable selection mentioned in Section 4.3 involve computing maximum over all models $|M| = s$ and there are $\binom{d}{s}$ many such. This can be prohibitive if d or s is large. Allowing for slightly enlarged confidence regions (conservative inference), one can try to approximate these maximums from above without exact computation. We now briefly discuss one way of doing this and details are left for a future work. Suppose we want to find the maximum norm of $w = (w_1, \ldots, w_m) \in \mathbb{R}^m_+$. Further suppose we know an upper bound, B, on $\|w\|_\infty$. Note the trivial inequality

$$\left(m^{-1}\sum_{j=1}^m w_i^q\right)^{1/q} \leqslant \|w\|_\infty \leqslant m^{1/q}\left(m^{-1}\sum_{j=1}^m w_i^q\right)^{1/q},$$

for any $q \geqslant 1$. If $q = \log(m)/\varepsilon$, then $\|w\|_\infty$ is $(m^{-1}\sum_{j=1}^m w_i^q)^{1/q}$ up to a factor of

e^{ε}. Observe now that $(m^{-1}\sum_{j=1}^{m} w_i^q) = \mathbb{E}_J[w_J^q]$ (for $J \sim \texttt{Unif}\{1,\ldots,m\}$) is an expectation which can be estimated by $k^{-1}\sum_{\ell=1}^{k} w_{j_\ell}^q$ for $j_1,\ldots,j_k \overset{iid}{\sim} \texttt{Unif}\{1,\ldots,m\}$. This is only an estimator of the expectation but using the apriori upper bound B, one can use any of the existing concentration inequalities to get a finite sample confidence interval for $(m^{-1}\sum_{j=1}^{m} w_i^q)^{1/q}$ which leads to an upper estimate of $\|w\|_\infty$. The details such as "which concentration inequality is good?, how good the upper bound is?" will be given elsewhere.

Bibliography

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.*, 6:170–176.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016). Uniformly valid confidence intervals post-model-selection. *ArXiv e-prints*.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and generalized gmm. *arXiv preprint arXiv:1806.01888*.

Bentkus, V. (2003). On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402.

Bentkus, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.*, 49(2):400–410.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Bonnans, J. F. and Shapiro, A. (2013). *Perturbation analysis of optimization problems*. Springer Science & Business Media.

Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields*, 162(1-2):47–70.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017a). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017b). Detailed proof of Nazarov's inequality. *arXiv preprint arXiv:1711.10696*.

Datta, P. and Sen, B. (2018). Optimal inference with a multidimensional multiscale statistic. *arXiv preprint arXiv:1806.02194*.

Deng, H. and Zhang, C.-H. (2017). Beyond gaussian approximation: bootstrap for maxima of sums of independent random vectors. *arXiv preprint arXiv:1705.09528*.

Dumbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, pages 124–152.

Guédon, O., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2015). On the interval of fluctuation of the singular values of random matrices. *arXiv preprint arXiv:1509.02322*.

Hörmann, S. (2009). Berry-Esseen bounds for econometric time series. *ALEA Lat. Am. J. Probab. Math. Stat.*, 6:377–397.

Jiao, J., Han, Y., and Weissman, T. (2018). Generalizations of maximal inequalities to arbitrary selection rules. *Statistics and Probability Letters*, 137:19–25.

Koike, Y. (2019). High-dimensional central limit theorems for homogeneous sums. *arXiv preprint arXiv:1902.03809*.

Koltchinskii, V. and Lounici, K. (2017a). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133.

Koltchinskii, V. and Lounici, K. (2017b). Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.*, 45(1):121–157.

Kuchibhotla, A. K. (2018). Deterministic Inequalities for Smooth M-estimators. *ArXiv e-prints:1809.05172*.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018a). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. *arXiv preprint arXiv:1802.05801*.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018b). Valid post-selection inference in assumption-lean linear regression. *arXiv preprint arXiv:1806.04119*.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving Beyond Sub-Gaussianity in High-Dimensional Statistics: Applications in Covariance Estimation and Linear Regression. *ArXiv e-prints:1804.02605*.

Kuchibhotla, A. K., Mukherjee, S., and Banerjee, D. (2018). High-dimensional CLT: Improvements, Non-uniform Extensions and Large Deviations. *arXiv:1806.06153*, page arXiv:1806.06153.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, pages 386–404.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, 225:248.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*, pages 169–187. Springer.

Pfanzagl, J. (1973). The accuracy of the normal approximation for estimates of vector parameters. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 25:171–198.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.

Raič, M. (2018). A multivariate berry–esseen theorem with explicit constants. *arXiv preprint arXiv:1802.06475*.

Rigollet, P. and Hütter, J.-C. (2015). High dimensional statistics. *Lecture notes for course 18S997*.

Romano, J. P. and Wolf, M. (2000). A more general central limit theorem for m-dependent random variables with unbounded m. *Statist. Probab. Lett.*, 47(2):115–124.

Rudelson, M., Vershynin, R., et al. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59(6):3434–3447.

Russo, D. and Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain. PMLR.

Tikhomirov, K. (2017). Sample covariance matrices of heavy-tailed distributions. *International Mathematics Research Notices*, 2018(20):6254–6289.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *J. Theoret. Probab.*, 25(3):655–686.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*.

White, H. (2001). *Asymptotic Theory for Econometricians*. Economic theory, econometrics, and mathematical economics. Academic Press.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919.

Zhang, K. (2017). Spherical cap packing asymptotics and rank-extreme detection. *IEEE Transactions on Information Theory*, 63(7):4572–4584.

Zhang, X. and Cheng, G. (2014). Bootstrapping High Dimensional Time Series. *ArXiv e-prints*.

Appendix A: Proof of Corollary 2.2

Proof. From the definition of Δ_n and Theorem 2.1 of Rudelson et al. (2013), we get for all $r > 0$,

$$\begin{aligned}\mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma > r + \|K^{1/2}\|_{HS}\right) &\leqslant \mathbb{P}\left(\|K^{1/2}N(0, I_d)\|_2 > r + \|K^{1/2}\|_{HS}\right) + \Delta_n \\ &\leqslant 2\exp\left(-\frac{c_1^2 r^2}{\|K^{1/2}\|_{op}^2}\right) + \Delta_n,\end{aligned}$$

for some constant $c_1 > 0$ (independent of p and n). Thus, we get for all $n \geqslant 1$,

$$\mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma > c_1^{-1}\|K^{1/2}\|_{op}\sqrt{\log n} + \|K^{1/2}\|_{HS}\right) \leqslant 2n^{-1} + \Delta_n. \tag{37}$$

For any set $A \subseteq \mathbb{R}^p$ and $\epsilon > 0$, let A^ϵ denote the ϵ-inflation of the set A with respect to the norm $\|\cdot\|_\Sigma$, that is, $A^\epsilon := \{y \in \mathbb{R}^p : \|y - x\|_\Sigma \leqslant \epsilon \text{ for some } x \in A\}$. Using Theorem 2.1, we get with $\mathcal{D}^\Sigma$ as in (4), for any set $A \subseteq \mathbb{R}^p$,

$$\begin{aligned}\mathbb{P}\left(\Sigma^{1/2}(\hat{\beta} - \beta) \in A\right) &\leqslant \mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta) \in A^{r_n\eta}\right) \\ &\quad + \mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma > r_n\right) + \mathbb{P}\left(\mathcal{D}^\Sigma > \eta\right), \\ \mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma} - \hat{\Sigma}\beta) \in A\right) &\leqslant \mathbb{P}\left(\Sigma^{1/2}(\hat{\beta} - \beta) \in A^{r_n\eta}\right) \\ &\quad + \mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma} - \hat{\Sigma}\beta)\|_\Sigma > r_n\right) + \mathbb{P}\left(\mathcal{D}^\Sigma > \eta\right).\end{aligned}$$

Therefore, we get

$$\begin{aligned}&\left|\mathbb{P}\left(\Sigma^{1/2}(\hat{\beta}-\beta)\in A\right)-\mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma}-\hat{\Sigma}\beta)\in A\right)\right|\\&\quad\leqslant\mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma}-\hat{\Sigma}\beta)\in A^{r_n\eta}\backslash A\right)+\mathbb{P}\left(\mathcal{D}^{\Sigma}>\eta\right)+\mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma}-\hat{\Sigma}\beta)\|_{\Sigma}>r_n\right).\end{aligned}$$

Additionally from the definition of Δ_n, we get for any convex set $A\subseteq\mathbb{R}^p$,

$$\begin{aligned}&\left|\mathbb{P}\left(\Sigma^{1/2}(\hat{\beta}-\beta)\in A\right)-\mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma}-\hat{\Sigma}\beta)\in A\right)\right|\\&\quad\leqslant\mathbb{P}\left(K^{1/2}N(0,I_d)\in A^{r_n\eta}\backslash A\right)+2\Delta_n+\mathbb{P}\left(\mathcal{D}^{\Sigma}>\eta\right)+\mathbb{P}\left(\|\Sigma^{-1}(\hat{\Gamma}-\hat{\Sigma}\beta)\|_{\Sigma}>r_n\right).\end{aligned}$$

Recall that $N(0,I_d)$ represents a standard normal random vector. Now we get, from Lemma 2.6 of Bentkus (2003) and the discussion following, that there exists a constant $c_2>0$ such that $\sup_{A\in\mathcal{C}_d}\mathbb{P}\left(K^{1/2}N(0,I_d)\in A^{r_n\eta}\backslash A\right)\leqslant c_2\|K^{-1}\|_*^{1/4}r_n\eta$, where $\|M\|_*$ for a matrix $M\in\mathbb{R}^{p\times p}$ denotes the nuclear norm of M. Hence

$$\begin{aligned}&\sup_{A\in\mathcal{C}_d}\left|\mathbb{P}\left(\Sigma^{1/2}(\hat{\beta}-\beta)\in A\right)-\mathbb{P}\left(\Sigma^{-1/2}(\hat{\Gamma}-\hat{\Sigma}\beta)\in A\right)\right|\\&\quad\leqslant c_2\|K^{-1}\|_*^{1/4}r_n\eta+2n^{-1}+3\Delta_n+\mathbb{P}\left(\mathcal{D}^{\Sigma}>\eta\right).\end{aligned}$$

Here we have used inequality (37). Finally, from the definition of Δ_n, we get

$$\begin{aligned}&\sup_{A\in\mathcal{C}_d}\left|\mathbb{P}\left(\Sigma^{1/2}(\hat{\beta}-\beta)\in A\right)-\mathbb{P}\left(K^{1/2}N(0,I_d)\in A\right)\right|\\&\quad\leqslant c_2\|K^{-1}\|_*^{1/4}r_n\eta+2n^{-1}+4\Delta_n+\mathbb{P}\left(\mathcal{D}^{\Sigma}>\eta\right).\end{aligned}$$

Since $\mathcal{C}_d$ is invariant under linear transformations, the result follows. □

Appendix B: Proof of Corollary 4.3

We first prove a version of Corollary 4.1 for the purpose of normal approximation with $\|\cdot\|_{\Sigma_M}$ replaced by $\|\cdot\|_{\Sigma_M V_M^{-1}\Sigma_M}$. We start with equality before (7) in the proof of Theorem 2.1 for model M:

$$\Sigma_M^{1/2}\left[\hat{\beta}_M-\beta_M-\Sigma_M^{-1}(\hat{\Gamma}_M-\hat{\Sigma}_M\beta_M)\right]\;=\;(I_{|M|}-\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2})\Sigma_M^{1/2}(\hat{\beta}_M-\beta_M).$$

Multiplying both sides by $V_M^{-1/2}$ and applying Euclidean norm, we get

$$\begin{aligned}&\|\hat{\beta}_M-\beta_M-\Sigma_M^{-1}(\hat{\Gamma}_M-\hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}\\&\quad\leqslant\|V_M^{-1/2}(I_{|M|}-\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2})V_M^{1/2}V_M^{-1/2}\Sigma_M^{1/2}(\hat{\beta}_M-\beta_M)\|\\&\quad\leqslant\|V_M^{-1/2}(I_{|M|}-\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2})V_M^{1/2}\|_{op}\|\hat{\beta}_M-\beta_M\|_{\Sigma_M V_M^{-1}\Sigma_M}\\&\quad=\mathcal{D}_M^{\Sigma}\|\hat{\beta}_M-\beta_M\|_{\Sigma_M V_M^{-1}\Sigma_M}.\end{aligned}$$

The last equality above follows from the fact that $\|AB\|_{op} = \|BA\|_{op}$. This implies

$$\|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M} \leqslant \frac{\mathcal{D}_M^{\Sigma}}{(1-\mathcal{D}_M^{\Sigma})_+}\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}. \tag{38}$$

Observe now that for any $x \in \mathbb{R}^{|M|}$ and any invertible matrix A,

$$\|x\|_A = \|A^{1/2}x\| = \max_{\theta\in\mathbb{R}^{|M|}} \frac{\theta^\top x}{\sqrt{\theta^\top A^{-1}\theta}} \geqslant \max_{\substack{\theta=\pm e_j,\\ 1\leqslant j\leqslant |M|}} \frac{|\theta^\top x|}{\sqrt{\theta^\top A^{-1}\theta}} = \max_{1\leqslant j\leqslant |M|} \frac{|x_j|}{\sqrt{(A^{-1})_j}} \tag{39}$$

Therefore, combining (38) and (39), we get for all $M \in \mathcal{M}$,

$$\max_{1\leqslant j\leqslant |M|} \frac{|(\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j|}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \leqslant \frac{\mathcal{D}_M^{\Sigma}\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}}{(1-\mathcal{D}_M^{\Sigma})_+}.$$

From the definition of the 1/2-net, it follows that

$$\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M} \leqslant 2\max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \theta^\top V_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M).$$

See, e.g., Rigollet and Hütter (2015, Theorem 1.19). Therefore, for all $M \in \mathcal{M}$,

$$\max_{1\leqslant j\leqslant |M|} \frac{|(\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j|}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \leqslant \frac{2\mathcal{D}_M^{\Sigma}\max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \theta^\top V_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)}{(1-\mathcal{D}_M^{\Sigma})_+}.$$

Using the definition of $\Xi_{n,\mathcal{M}}$, we can control $\max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \theta^\top V_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)$. Observe first that

$$\begin{aligned}
&\mathbb{P}\left(\max_{M\in\mathcal{M}}\max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \frac{\theta^\top \bar{G}_M}{\sqrt{2\log(|\mathcal{M}|5^{|M|}\pi_{|M|}) + 2\log(|M|^2/\Xi_{n,\mathcal{M}})}} \geqslant 1\right) \\
&\qquad \leqslant \sum_{s=1}^{d} \mathbb{P}\left(\max_{M\in\mathcal{M},|M|=s}\max_{\theta\in\mathcal{N}_s^{1/2}} \theta^\top \bar{G}_M \geqslant \sqrt{2\log(|\mathcal{M}|5^s\pi_s) + 2\log(s^2/\Xi_{n,\mathcal{M}})}\right).
\end{aligned} \tag{40}$$

Since $\bar{G}_M$ is a standard normal random vector for each $M \in \mathcal{M}$, $\theta^\top \bar{G}_M$ is a standard Gaussian random variable and it follows from Rigollet and Hütter (2015, Theorem 1.14) that for all $t \geqslant 0$,

$$\mathbb{P}\left(\max_{M\in\mathcal{M},|M|=s}\max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \theta^\top \bar{G}_M \geqslant \sqrt{2\log(|\mathcal{M}|5^s\pi_s) + 2t}\right) \leqslant \exp(-t),$$

Taking $t = \log(s^2/\Delta_{n,M})$ yields

$$\mathbb{P}\left(\max_{M\in\mathcal{M},|M|=s} \max_{\theta\in\mathcal{N}_s^{1/2}} \theta^\top \bar{G}_M \geqslant \sqrt{2\log(|\mathcal{M}|5^s\pi_s) + 2\log(s^2/\Xi_{n,\mathcal{M}})} \right) \leqslant \frac{\Xi_{n,\mathcal{M}}}{s^2}.$$

Combining this with (40) and using $\sum_{s=1}^d s^{-2} \leqslant \pi^2/6 < 1.65$, we get

$$\mathbb{P}\left(\max_{M\in\mathcal{M}} \max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \frac{\theta^\top \bar{G}_M}{\sqrt{2\log(|\mathcal{M}|5^{|M|}\pi_{|M|}) + 2\log(|M|^2/\Xi_{n,\mathcal{M}})}} \geqslant 1 \right) \leqslant 1.65\Xi_{n,\mathcal{M}}.$$

From the definition of $\Xi_{n,M}$, it follows that

$$\mathbb{P}\left(\max_{M\in\mathcal{M}} \max_{\theta\in\mathcal{N}_{|M|}^{1/2}} \frac{\theta^\top V_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)}{\sqrt{2\log(|\mathcal{M}|5^{|M|}\pi_{|M|}) + 2\log(|M^2|/\Xi_{n,\mathcal{M}})}} > 1 \right) \leqslant 2.65\Xi_{n,\mathcal{M}}.$$

Hence for any $(\eta_M)_{M\in\mathcal{M}}(\leqslant 1/2)$, on an event with probability at least $1-2.65\Xi_{n,M} - \mathbb{P}(\cup_{M\in\mathcal{M}}\{\mathcal{D}_M^\Sigma \geqslant \eta_M\})$, we get

$$\max_{1\leqslant j\leqslant|M|} \frac{|(\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j|}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \leqslant 4\eta_M\sqrt{2\log(|\mathcal{M}|5^{|M|}|M|^2\pi_{|M|}/\Xi_{n,M})}. \tag{41}$$

Define a vector $\varepsilon \in \mathbb{R}^{\sum_{M\in\mathcal{M}}|M|}$ indexed by $M \in \mathcal{M}, 1 \leqslant j \leqslant |M|$ such that

$$\varepsilon_{M,j} := 4\eta_M\sqrt{2\log(|\mathcal{M}|5^{|M|}|M|^2\pi_{|M|}/\Xi_{n,M})}.$$

Fix any set $A \in \mathcal{A}^{sre}$. Then from (41), we get

$$\mathbb{P}\left(\left(\frac{(\hat{\beta}_M - \beta_M)_j}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \in A \right) \leqslant \mathbb{P}\left(\left(\frac{(\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \in A + \varepsilon \right) + 2.65\Xi_{n,M} + \mathbb{P}\left(\bigcup_{M\in\mathcal{M}} \{\mathcal{D}_M^\Sigma \geqslant \eta_M\} \right),$$

and

$$\mathbb{P}\left(\left(\frac{(\hat{\beta}_M - \beta_M)_j}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \in A \right) \geqslant \mathbb{P}\left(\left(\frac{(\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M))_j}{\sqrt{(\Sigma_M^{-1}V_M\Sigma_M^{-1})_j}} \right)_{\substack{M\in\mathcal{M},\\ 1\leqslant j\leqslant|M|}} \in A - \varepsilon \right) - 2.65\Xi_{n,M} - \mathbb{P}\left(\bigcup_{M\in\mathcal{M}} \{\mathcal{D}_M^\Sigma \geqslant \eta_M\} \right),$$

Hence the result follows from the definition of $\Delta_{n,\mathcal{M}}$.

Appendix C: Proof of Proposition 5.1

Observe that

$$\mathcal{D}_M^{\Sigma} = \|\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2} - I_{|M|}\|_{op} \leqslant 2 \sup_{\nu\in\mathcal{N}_{|M|}^{1/4}} \left|\frac{1}{n}\sum_{i=1}^{n}(\nu^\top\Sigma_M^{-1/2}X_{i,M})^2 - 1\right|, \tag{42}$$

where $\mathcal{N}_{|M|}^{1/4}$ represents the 1/4-net of $\{\theta \in \mathbb{R}^{|M|} : \|\theta\| = 1\}$; see Lemma 2.2 of Vershynin (2012). Note that $|\mathcal{N}_{|M|}^{1/4}| \leqslant 9^{|M|}$. Therefore the right hand side of (42) is a maximum over a finite number of mean zero averages with summands satisfying

$$\mathbb{E}\left[\exp\left(\mathfrak{K}_\beta^{-\beta}|\nu^\top\Sigma_M^{-1/2}X_{i,M}|^\beta\right)\right] \leqslant 2, \text{ for all } \nu \in \mathcal{N}_{|M|}^{1/4} \text{ and } M \subseteq \{1,2,\ldots,d\}.$$

Applying Theorem 3.4 of Kuchibhotla and Chakrabortty (2018), we get for any $t \geqslant 0$ that with probability $1 - 3e^{-t}$,

$$\mathcal{D}_M^{\Sigma} \leqslant 14\sqrt{\frac{\kappa_M^{\Sigma}(t + |M|\log(9))}{n}} + \frac{C_\beta\mathfrak{K}_\beta^2(\log(2n))^{2/\beta}(t + |M|\log(9))^{\max\{1,2/\beta\}}}{n},$$

for some constant $C_\beta > 0$ depending only β. Since there are $\binom{d}{s} \leqslant (ed/s)^s$ models of size s, taking $t = s\log(ed/s) + u$ (for any $u \geqslant 0$) and applying union bound over all models of size s, we get that with probability $1 - 3e^{-u}$, simultaneously for all $M \subseteq \{1,2,\ldots,d\}$ with $|M| = s$,

$$\mathcal{D}_M^{\Sigma} \leqslant 14\sqrt{\frac{\kappa_M^{\Sigma}(u + s\log(9ed/s))}{n}} + \frac{C_\beta\mathfrak{K}_\beta^2(\log(2n))^{2/\beta}(u + s\log(9ed/s))^{\max\{1,2/\beta\}}}{n}.$$

To prove the result simultaneously over all $1 \leqslant s \leqslant d$, take $u = v + \log(\pi^2 s^2/6)$ and apply union bound over $1 \leqslant s \leqslant d$ to get with probability $1 - 3e^{-v}$ simulataneously over all $M \subseteq \{1,2,\ldots,d\}$ with $|M| = s$ for some $1 \leqslant s \leqslant d$,

$$\mathcal{D}_M^{\Sigma} \leqslant 14\sqrt{\frac{\kappa_M^{\Sigma}(v + \log(\pi^2 s^2/6) + s\log(9ed/s))}{n}} + \frac{C_\beta\mathfrak{K}_\beta^2(\log(2n))^{2/\beta}(v + \log(\pi^2 s^2/6) + s\log(9ed/s))^{\max\{1,2/\beta\}}}{n}.$$

Since $s^{-1}\log(\pi^2 s^2/6) \leqslant (2\pi/\sqrt{6})\sup_{x\geqslant\pi/\sqrt{6}}\exp(-x)x \leqslant 1$, we get with probability $1 - 3e^{-v}$ simultaneously for any $1 \leqslant s \leqslant d$ and for any model $M \subseteq \{1,2,\ldots,d\}$ with $|M| = s$,

$$\mathcal{D}_M^{\Sigma} \leqslant 14\sqrt{\frac{\kappa_M^{\Sigma}(v + s\log(9e^2 d/s))}{n}} + \frac{C_\beta\mathfrak{K}_\beta^2(\log(2n))^{2/\beta}(v + s\log(9e^2 d/s))^{\max\{1,2/\beta\}}}{n}.$$

This completes the proof of (34).

We now bound $\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|$ simultaneously over all M. Observe from the definition of β_M that

$$0 \leqslant \sum_{i=1}^n \mathbb{E}\left[(Y_i - X_{i,M}^\top \beta_M)^2\right] = \sum_{i=1}^n \mathbb{E}[Y_i^2] - \sum_{i=1}^n \mathbb{E}[(X_{i,M}^\top \beta_M)^2],$$

and hence $\|\tilde{\beta}_M\| = \|\Sigma_M^{1/2}\beta_M\| \leqslant (\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2}$. Now note that since $\mathbb{E}[\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M] = 0$ (from the definition of β_M), we have

$$\begin{aligned}
\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\| &= \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M) - \Sigma_M^{-1/2}(\hat{\Sigma}_M - \Sigma_M)\beta_M\| \\
&\leqslant \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| + \|\Sigma_M^{-1/2}(\hat{\Sigma}_M - \Sigma_M)\Sigma_M^{-1/2}\|_{op}\|\Sigma_M^{1/2}\beta_M\| \\
&\leqslant \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| + \mathcal{D}_M^\Sigma(\textstyle\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2}.
\end{aligned}$$

We have already controlled $\mathcal{D}_M^\Sigma$ uniformly over all models $M \subseteq \{1, 2, \ldots, d\}$ and hence it is enough to control $\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\|$. As before, observe that

$$\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| \leqslant 2 \max_{\nu \in \mathcal{N}_{|M|}^{1/2}} \left|\frac{1}{n}\sum_{i=1}^n \left\{\nu^\top \tilde{X}_{i,M} Y_i - \mathbb{E}[\nu^\top \tilde{X}_{i,M} Y_i]\right\}\right| =: 2\mathcal{E}_M,$$

where $\tilde{X}_{i,M} := \Sigma_M^{-1/2} X_{i,M}$. To control $\mathcal{E}_M$ we split Y_i in to two parts depending on whether $\{|Y_i| \leqslant B\}$ or $\{|Y_i| > B\}$ (for a B to be chosen later). Define $Y_{i,1} = Y_i \mathbb{1}\{|Y_i| \leqslant B\}$, $Y_{i,2} = Y_i - Y_{i,1}$ and for $\ell = 1, 2$,

$$\mathcal{E}_{M,\ell} := \max_{\nu \in \mathcal{N}_{|M|}^{1/2}} \left|\frac{1}{n}\sum_{i=1}^n \left\{\nu^\top \tilde{X}_{i,M} Y_{i,\ell} - \mathbb{E}[\nu^\top \tilde{X}_{i,M} Y_{i,\ell}]\right\}\right|.$$

Since $|Y_{i,1}| \leqslant B$, we have for any $\nu \in \mathcal{N}_{|M|}^{1/2}$ and $M \subseteq \{1, 2, \ldots, d\}$ that

$$\mathbb{E}\left[\exp\left(\frac{|\nu^\top \tilde{X}_{i,M} Y_{i,1}|^\beta}{(B\mathfrak{K}_\beta)^\beta}\right)\right] \leqslant 2.$$

Hence we get by Theorem 3.4 of Kuchibhotla and Chakrabortty (2018) that for any $t \geqslant 0$, with probability $1 - 3e^{-t}$

$$\mathcal{E}_{M,1} \leqslant 7\sqrt{\frac{\mathfrak{V}_M(t + |M|\log(5))}{n}} + \frac{C_\beta B\mathfrak{K}_\beta(\log(2n))^{1/\beta}(t + |M|\log(5))^{\max\{1,1/\beta\}}}{n}.$$

Now following same approach as used for $\mathcal{D}_M^\Sigma$, we get with probability $1 - 3e^{-u}$, for any $1 \leqslant s \leqslant d$, for any model $M \subseteq \{1, 2, \ldots, d\}$ such that $|M| = s$,

$$\mathcal{E}_{M,1} \leqslant 7\sqrt{\frac{\mathfrak{V}_M(v + s\log(5e^2 d/s))}{n}} + \frac{C_\beta B\mathfrak{K}_\beta(\log(2n))^{1/\beta}(v + s\log(5e^2 d/s))^{\max\{1,1/\beta\}}}{n}. \tag{43}$$

To bound $\mathcal{E}_{M,2}$ simultaneously over all M, we take

$$B := 8\mathbb{E}\left[\max_{1\leqslant i\leqslant n}|Y_i|\right] \leqslant 8n^{1/r}\max_{1\leqslant i\leqslant n}\left(\mathbb{E}[|Y_i|^r]\right)^{1/r} = 8n^{1/r}K_{n,r},$$

which is motivated by Proposition 6.8 of Ledoux and Talagrand (1991). Now consider the normalized process

$$\mathcal{E}_{2,\texttt{Norm}} := \max_{1\leqslant s\leqslant d}\max_{|M|=s}\frac{n^{1/2}\mathcal{E}_{M,2}}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}.$$

Observe first that $\mathcal{E}_{2,\texttt{Norm}} \leqslant \mathcal{E}^{(1)} + \mathbb{E}[\mathcal{E}^{(1)}]$, where

$$\mathcal{E}^{(1)} = \frac{1}{n}\sum_{i=1}^n \max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{1/2}|\nu^\top\tilde{X}_{i,M}Y_{i,2}|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}.$$

Note that $\mathcal{E}^{(1)}$ is an average of non-negative random variables and hence by the choice of B above and Proposition 6.8 of Ledoux and Talagrand (1991), we get

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}^{(1)}\right] &\leqslant 8\mathbb{E}\left[\frac{1}{n}\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{1/2}|\nu^\top\tilde{X}_{i,M}Y_{i,2}|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right]\\
&\leqslant 8\mathbb{E}\left[\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{-1/2}|\nu^\top\tilde{X}_{i,M}Y_i|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right] \quad (44)\\
&\leqslant 8\left\|\max_{1\leqslant i\leqslant n}\frac{|Y_i|}{K_{n,r}n^{1/r}}\right\|_2\left\|\max_{1\leqslant s\leqslant d}\max_{\substack{1\leqslant i\leqslant n,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{|\nu^\top\tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_2.
\end{aligned}$$

Here we use $\|W\|_2$ for a random variable W to denote $(\mathbb{E}[W^2])^{1/2}$. In the second factor, the number of items in the maximum for any fixed s is given by $n\binom{d}{s}5^s \leqslant n(5ed/s)^s$ and hence from (33), we get

$$\mathbb{P}\left(\max_{\substack{1\leqslant i\leqslant n,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}|\nu^\top\tilde{X}_{i,M}| \geqslant \mathfrak{K}_\beta(t+s\log(5ed/s)+\log(n))^{1/\beta}\right) \leqslant 2e^{-t},$$

and an application of union bound over $1\leqslant s\leqslant d$ yields

$$\mathbb{P}\left(\bigcup_{1\leqslant s\leqslant d}\left\{\max_{\substack{1\leqslant i\leqslant n,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}|\nu^\top\tilde{X}_{i,M}| \geqslant \mathfrak{K}_\beta(t+\log(\pi^2s^2/6)+s\log(5ed/s)+\log(n))^{1/\beta}\right\}\right) \leqslant 2e^{-t},$$

which implies

$$\mathbb{P}\left(\bigcup_{1\leqslant s\leqslant d}\left\{\max_{\substack{1\leqslant i\leqslant n,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}|\nu^\top\tilde{X}_{i,M}|\geqslant\mathfrak{K}_\beta(t+s\log(5e^2d/s)+\log(n))^{1/\beta}\right\}\right)\leqslant 2e^{-t}. \tag{45}$$

Hence for a constant $C_\beta>0$ (depending only on β),

$$\left\|\max_{1\leqslant s\leqslant d}\max_{\substack{1\leqslant i\leqslant n,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{|\nu^\top\tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_2\leqslant C_\beta. \tag{46}$$

For the first factor in (44), note that (since $r\geqslant 2$)

$$\left\|\max_{1\leqslant i\leqslant n}\frac{|Y_i|}{K_{n,r}n^{1/r}}\right\|_2\leqslant\left\|\max_{1\leqslant i\leqslant n}\frac{|Y_i|}{K_{n,r}n^{1/r}}\right\|_r\leqslant\left(\sum_{i=1}^n\mathbb{E}\left[\frac{|Y_i|^r}{K_{n,r}^r n}\right]\right)^{1/r}\leqslant 1. \tag{47}$$

Substituting the bounds (47) and (46) in (44) yields

$$\mathbb{E}[\mathcal{E}_{2,\texttt{Norm}}]\leqslant 2\mathbb{E}[\mathcal{E}^{(1)}]\leqslant C_\beta, \tag{48}$$

for a constant $C_\beta>0$ (which is different from the one in (46)). Applying Theorem 8 of Boucheron et al. (2005) now yields for every $q\geqslant 1$

$$\|\mathcal{E}^{(1)}\|_q\leqslant 2\mathbb{E}[\mathcal{E}^{(1)}]+Cq\left\|\frac{1}{n}\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{1/2}|\nu^\top\tilde{X}_{i,M}Y_{i,2}|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_q,$$

for some (other) absolute constant $C>0$. This implies (using (48)) that

$$\|\mathcal{E}_{2,\texttt{Norm}}\|_q\leqslant 3C_\beta+Cq\left\|\frac{1}{n}\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{1/2}|\nu^\top\tilde{X}_{i,M}Y_{i,2}|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_q.$$

As before, we have

$$\begin{aligned}
&\left\|\frac{1}{n}\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{n^{1/2}|\nu^\top\tilde{X}_{i,M}Y_{i,2}|}{n^{-1/2+1/r}K_{n,r}\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_q\\
&\quad\leqslant\left\|\max_{1\leqslant i\leqslant n}\frac{|Y_i|}{K_{n,r}n^{1/r}}\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{|\nu^\top\tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_q\\
&\quad\leqslant\left\|\max_{1\leqslant i\leqslant n}\frac{|Y_i|}{K_{n,r}n^{1/r}}\right\|_r\left\|\max_{1\leqslant i\leqslant n}\max_{\substack{1\leqslant s\leqslant d,\\ |M|=s}}\max_{\nu\in\mathcal{N}_s^{1/2}}\frac{|\nu^\top\tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s\log(5e^2d/s)+\log n)^{1/\beta}}\right\|_{rq/(r-q)}.
\end{aligned}$$

where the last inequality holds for any $q < r$ by Hölder's inequality. We already have that the first factor is bounded be 1. From (45), we have

$$\left\| \max_{1 \leqslant i \leqslant n} \max_{\substack{1 \leqslant s \leqslant d, \\ |M| = s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_{rq/(r-q)} \leqslant C_\beta \left(\frac{rq}{r-q} \right)^{1/\beta}.$$

Therefore taking $q = r - 1$, we get

$$\|\mathcal{E}_{2,\texttt{Norm}}\|_{r-1} \leqslant 3C_\beta + CC_\beta (r-1)(r(r-1))^{1/\beta} =: C_{\beta,r}.$$

Hence by Markov's inequality, we get with probability at least $1 - 1/t^{r-1}$, for any $1 \leqslant s \leqslant d$, for any model $M \subseteq \{1, 2, \ldots, d\}$ such that $|M| = s$,

$$\mathcal{E}_{M,2} \leqslant \frac{t C_{\beta,r} K_{n,r} \mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}}{n^{1-1/r}}. \tag{49}$$

Combining the bounds (43) and (49) yields: with probability at least $1 - 3e^{-t} - t^{-r+1}$, for any $1 \leqslant s \leqslant d$, for any model $M \subseteq \{1, 2, \ldots, d\}$ such that $|M| = s$,

$$\mathcal{E}_M \leqslant 7\sqrt{\frac{\mathfrak{V}_M (t + s \log(5e^2 d/s))}{n}} + \frac{C_\beta K_{n,r} \mathfrak{K}_\beta (\log(2n))^{1/\beta} (t + s \log(5e^2 d/s))^{\max\{1, 1/\beta\}}}{n^{1-1/r}} + \frac{t C_{\beta,r} K_{n,r} \mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}}{n^{1-1/r}}.$$

Combining all inequaliteies completes the proof of (35).