Probab. Theory Relat. Fields (2018) 172:983–1079
https://doi.org/10.1007/s00440-017-0824-7

Asymptotics for high dimensional regression *M*-estimates: fixed design results

Lihua Lei[1] · **Peter J. Bickel**[1] · **Noureddine El Karoui**[1,2]

Received: 26 January 2017 / Revised: 7 December 2017 / Published online: 9 February 2018

Abstract We investigate the asymptotic distributions of coordinates of regression M-estimates in the moderate p/n regime, where the number of covariates p grows proportionally with the sample size n. Under appropriate regularity conditions, we establish the coordinate-wise asymptotic normality of regression M-estimates assuming a fixed-design matrix. Our proof is based on the second-order Poincaré inequality and leave-one-out analysis. Some relevant examples are indicated to show that our regularity conditions are satisfied by a broad class of design matrices. We also show a counterexample, namely an ANOVA-type design, to emphasize that the technical assumptions are not just artifacts of the proof. Finally, numerical experiments confirm and complement our theoretical results.

Keywords M-estimation · Robust regression · High-dimensional statistics · Second order Poincaré inequality · Leave-one-out analysis

Mathematics Subject Classification Primary 62J99; Secondary 62E20

Peter J. Bickel and Lihua Lei gratefully acknowledge support from NSF DMS-1160319 and NSF DMS-1713083. Noureddine El Karoui gratefully acknowledges support from NSF grant DMS-1510172. He would also like to thank Criteo for providing a great research environment.

✉ Lihua Lei
lihua.lei@berkeley.edu

✉ Peter J. Bickel
bickelp@berkeley.edu

✉ Noureddine El Karoui
nkaroui@berkeley.edu; n.elkaroui@criteo.com

[1] Department of Statistics, University of California, Berkeley, Berkeley, CA, USA

[2] Criteo Research, Paris, France

1 Introduction

High-dimensional statistics has a long history [31,58,59] with considerable renewed interest over the last two decades. In many applications, the researcher collects data which can be represented as a matrix, called a design matrix and denoted by $X \in \mathbb{R}^{n\times p}$, as well as a response vector $y \in \mathbb{R}^n$ and aims to study the connection between X and y. The linear model is among the most popular models as a starting point of data analysis in various fields. A linear model assumes that

$$y = X\beta^* + \varepsilon, \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ is the coefficient vector which measures the marginal contribution of each predictor and ε is a random vector which captures the unobserved errors.

The aim of this article is to provide valid inferential results for features of β^*. For example, a researcher might be interested in testing whether a given predictor has a negligible effect on the response, or equivalently whether $\beta_j^* = 0$ for some j. Similarly, linear contrasts of β^* such as $\beta_1^* - \beta_2^*$ might be of interest in the case of the group comparison problem in which the first two predictors represent the same feature but are collected from two different groups.

An M-estimator, defined as

$$\hat{\beta}(\rho) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - x_i^T\beta\right) \tag{2}$$

where ρ denotes a loss function, is among the most popular estimators used in practice [31,47]. In particular, if $\rho(x) = \frac{1}{2}x^2$, $\hat{\beta}(\rho)$ is the famous Least Square Estimator (LSE). We intend to explore the distribution of $\hat{\beta}(\rho)$, based on which we can achieve the inferential goals mentioned above.

The most well-studied approach is the asymptotic analysis, which assumes that the scale of the problem grows to infinity and use the limiting result as an approximation. In regression problems, the scale parameter of a problem is the sample size n and the number of predictors p. The classical approach is to fix p and let n grow to infinity. It has been shown [30,31,47,61] that $\hat{\beta}(\rho)$ is consistent in terms of L_2 norm and asymptotically normal in this regime. The asymptotic variance can be then approximated by the bootstrap [8]. Later on, the studies are extended to the regime in which both n and p grow to infinity but p/n converges to 0 [39,42–45,62]. The consistency, in terms of the L_2 norm, the asymptotic normality and the validity of the bootstrap still hold in this regime. Based on these results, we can construct a 95% confidence interval for β_{0j} simply as $\hat{\beta}_j(\rho) \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j(\rho))}$ where $\widehat{\mathrm{Var}}(\hat{\beta}_j(\rho))$ is calculated by bootstrap. Similarly we can calculate p-values for the hypothesis testing procedure.

We ask whether the inferential results developed under the low-dimensional assumptions and the software built on top of them can be relied on for moderate and high-dimensional analysis? Concretely, if in a study $n = 50$ and $p = 40$, can the software built upon the assumption that $p/n \simeq 0$ be relied on when $p/n = .8$? Results in random matrix theory [40] already offer an answer in the negative side for many

PCA-related questions in multivariate statistics. The case of regression is more subtle: For instance for least-squares, standard degrees of freedom adjustments effectively take care of many dimensionality-related problems. But this nice property does not extend to more general regression M-estimates.

Once these questions are raised, it becomes very natural to analyze the behavior and performance of statistical methods in the regime where p/n is fixed. Indeed, it will help us to keep track of the inherent statistical difficulty of the problem when assessing the variability of our estimates. In other words, we assume in the current paper that $p/n \to \kappa > 0$ while let n grows to infinity. Due to identifiability issues, it is impossible to make inference on β^* if $p > n$ without further structural or distributional assumptions. We discuss this point in details in Sect. 2.3. Thus we consider the regime where $p/n \to \kappa \in (0, 1)$. We call it the moderate p/n regime. This regime is also the natural regime in random matrix theory [2,33,40,59]. It has been shown that the asymptotic results derived in this regime sometimes provide an extremely accurate approximation to finite sample distributions of estimators at least in certain cases [33] where n and p are both small.

1.1 Qualitatively different behavior of moderate p/n regime

First, $\hat{\beta}(\rho)$ is no longer consistent in terms of L_2 norm and the risk $\mathbb{E}\|\hat{\beta}(\rho) - \beta^*\|^2$ tends to a non-vanishing quantity determined by κ, the loss function ρ and the error distribution through a complicated system of non-linear equations [5,20–22]. This L_2-inconsistency prohibits the use of standard perturbation-analytic techniques to assess the behavior of the estimator. It also leads to qualitatively different behaviors for the residuals in moderate dimensions; in contrast to the low-dimensional case, they cannot be relied on to give accurate information about the distribution of the errors. However, this seemingly negative result does not exclude the possibility of inference since $\hat{\beta}(\rho)$ is still consistent in terms of $L_{2+\nu}$ norms for any $\nu > 0$ and in particular in L_∞ norm. Thus, we can at least hope to perform inference on each coordinate.

Second, classical optimality results do not hold in this regime. In the regime $p/n \to 0$, the maximum likelihood estimator is shown to be optimal [7,29,30]. In other words, if the error distribution is known then the M-estimator associated with the loss $\rho(\cdot) = -\log f_\varepsilon(\cdot)$ is asymptotically efficient, provided the design is of appropriate type, where $f_\varepsilon(\cdot)$ is the density of entries of ε. However, in the moderate p/n regime, it has been shown that the optimal loss is no longer the log-likehood but an other function with a complicated but explicit form [6], at least for certain designs. The suboptimality of maximum likelihood estimators suggests that classical techniques fail to provide valid intuition in the moderate p/n regime.

Third, the joint asymptotic normality of $\hat{\beta}(\rho)$, as a p-dimensional random vector, may be violated for a fixed design matrix X. This has been proved for least-squares by [31] in his pioneering work. For general M-estimators, this negative result is a simple consequence of the results of [22]: They exhibit an ANOVA design (see below) where even marginal fluctuations are not Gaussian. By contrast, for random design, they show that $\hat{\beta}(\rho)$ is jointly asymptotically normal when the design matrix is elliptical with general covariance by using the non-asymptotic stochastic representation for $\hat{\beta}(\rho)$ as

well as elementary properties of vectors uniformly distributed on the uniform sphere in $\mathbb{R}^p$; See section 2.2.3 of [22] or the supplementary material of [6] for details. This does not contradict [31]'s negative result in that it takes the randomness from both X and ε into account while [31]'s result only takes the randomness from ε into account. Later, [21] shows that each coordinate of $\hat{\beta}(\rho)$ is asymptotically normal for a broader class of random designs. This is also an elementary consequence of the analysis in [20]. However, to the best of our knowledge, beyond the ANOVA situation mentioned above, there are no distributional results for fixed design matrices. This is the topic of this article.

Last but not least, bootstrap inference fails in this moderate-dimensional regime. This has been shown by [9] for least-squares and residual bootstrap in their influential work. Recently, [24] studied the results to general M-estimators and showed that all commonly used bootstrapping schemes, including pairs-bootstrap, residual bootstrap and jackknife, fail to provide a consistent variance estimator and hence valid inferential statements. These latter results even apply to the marginal distributions of the coordinates of $\hat{\beta}(\rho)$. Moreover, there is no simple, design independent, modification to achieve consistency [24].

1.2 Our contributions

In summary, the behavior of the estimators we consider in this paper is completely different in the moderate p/n regime from its counterpart in the low-dimensional regime. As discussed in the next section, moving one step further in the moderate p/n regime is interesting from both the practical and theoretical perspectives. The main contribution of this article is to establish coordinate-wise asymptotic normality of $\hat{\beta}(\rho)$ for certain *fixed design matrices* X in this regime under technical assumptions. The following theorem informally states our main result.

Theorem 1.1 (Informal Version of Theorem 3.1 in Sect. 3) *Under appropriate conditions on the design matrix X, the distribution of ε and the loss function ρ, as $p/n \to \kappa \in (0, 1)$, while $n \to \infty$,*

$$\max_{1\le j\le p} d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j(\rho) - \mathbb{E}\hat{\beta}_j(\rho)}{\sqrt{\mathrm{Var}(\hat{\beta}_j(\rho))}}\right), N(0,1)\right) = o(1)$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total variation distance and $\mathscr{L}(\cdot)$ denotes the law.

It is worth mentioning that the above result can be extended to finite dimensional linear contrasts of $\hat{\beta}$. For instance, one might be interested in making inference on $\beta_1^* - \beta_2^*$ in the problems involving the group comparison. The above result can be extended to give the asymptotic normality of $\hat{\beta}_1 - \hat{\beta}_2$.

Besides the main result, we have several other contributions. First, we use a new approach to establish asymptotic normality. Our main technique is based on the second-order Poincaré inequality (SOPI), developed by [10] to derive, among many other results, the fluctuation behavior of linear spectral statistics of random matrices. In

contrast to classical approaches such as the Lindeberg–Feller central limit theorem, the second-order Poincaré inequality is capable of dealing with nonlinear and potentially implicit functions of independent random variables. Moreover, we use different expansions for $\hat{\beta}(\rho)$ and residuals based on double leave-one-out ideas introduced in [22], in contrast to the classical perturbation-analytic expansions. See aforementioned paper and follow-ups. An informal interpretation of the results of [10] is that if the Hessian of the nonlinear function of random variables under consideration is sufficiently small, this function acts almost linearly and hence a standard central limit theorem holds.

Second, to the best of our knowledge this is the first inferential result for fixed (non ANOVA-like) design in the moderate p/n regime. Fixed designs arise naturally from an experimental design or a conditional inference perspective. That is, inference is ideally carried out without assuming randomness in predictors; see Sect. 2.2 for more details. We clarify the regularity conditions for coordinate-wise asymptotic normality of $\hat{\beta}(\rho)$ explicitly, which are checkable for LSE and also checkable for general M-estimators if the error distribution is known. We also prove that these conditions are satisfied with by a broad class of designs.

The ANOVA-like design described in Sect. 3.3.4 exhibits a situation where the distribution of $\hat{\beta}_j(\rho)$ is not going to be asymptotically normal. As such the results of Theorem 3.1 below are somewhat surprising.

For complete inference, we need both the asymptotic normality and the asymptotic bias and variance. Under suitable symmetry conditions on the loss function and the error distribution, it can be shown that $\hat{\beta}(\rho)$ is unbiased (see Sect. 3.2.1 for details) and thus it is left to derive the asymptotic variance. As discussed at the end of Sect. 1.1, classical approaches, e.g. bootstrap, fail in this regime. For least-squares, classical results continue to hold and we discuss it in Sect. 5 for the sake of completeness. However, for M-estimators, there is no closed-form result. We briefly touch upon the variance estimation in Sect. 3.4.2. The derivation for general situations is beyond the scope of this paper and left to the future research.

1.3 Outline of paper

The rest of the paper is organized as follows: In Sect. 2, we clarify details which are mentioned in the current section. In Sect. 3, we state the main result (Theorem 3.1) formally and explain the technical assumptions. Then we show several examples of random designs which satisfy the assumptions with high probability. In Sect. 4, we introduce our main technical tool, second-order Poincaré inequality [10], and apply it on M-estimators as the first step to prove Theorem 3.1. Since the rest of the proof of Theorem 3.1 is complicated and lengthy, we illustrate the main ideas in "Appendix A". The rigorous proof is left to "Appendix B". In Sect. 5, we provide reminders about the theory of least-squares estimation for the sake of completeness, by taking advantage of its explicit form. In Sect. 6, we display the numerical results. The proof of other results are stated in "Appendix C" and more numerical experiments are presented in "Appendix D".

2 More details on background

2.1 Moderate p/n regime: a more informative type of asymptotics?

In Sect. 1, we mentioned that the ratio p/n measures the difficulty of statistical inference. The moderate p/n regime provides an approximation of finite sample properties with the difficulties fixed at the same level as the original problem. Intuitively, this regime should capture more variation in finite sample problems and provide a more accurate approximation. We will illustrate this via simulation.

Consider a study involving 50 participants and 40 variables; we can either use the asymptotics in which p is fixed to be 40, n grows to infinity or p/n is fixed to be 0.8, and n grows to infinity to perform approximate inference. Current software rely on low-dimensional asymptotics for inferential tasks, but there is no evidence that they yield more accurate inferential statements than the ones we would have obtained using moderate dimensional asymptotics. In fact, numerical evidence [6,23,33] show that the reverse is true.

We exhibit a further numerical simulation showing that. Consider a case that $n = 50$, ε has i.i.d. entries and X is one realization of a matrix generated with i.i.d. gaussian (mean 0, variance 1) entries. For $\kappa \in \{0.1, 0.2, \ldots, 0.9\}$ and different error distributions, we use the Kolmogorov–Smirnov (KS) statistics to quantify the distance between the finite sample distribution and two types of asymptotic approximation of the distribution of $\hat{\beta}_1(\rho)$.

Specifically, we use the Huber loss function $\rho_{\text{Huber},k}$ with default parameter $k = 1.345$ [32], i.e.

$$\rho_{\text{Huber},k}(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le k \\ k\left(|x| - \frac{1}{2}k\right) & |x| > k \end{cases}$$

Specifically, we generate three design matrices $X^{(0)}$, $X^{(1)}$ and $X^{(2)}$: $X^{(0)}$ for small sample case with a sample size $n = 50$ and a dimension $p = n\kappa$; $X^{(1)}$ for low-dimensional asymptotics (p fixed) with a sample size $n = 1000$ and a dimension $p = 50\kappa$; and $X^{(2)}$ for moderate-dimensional asymptotics (p/n fixed) with a sample size $n = 1000$ and a dimension $p = n\kappa$. Each of them is generated as one realization of an i.i.d. standard gaussian design and then treated as fixed across $K = 100$ repetitions. For each design matrix, vectors ε of appropriate length are generated with i.i.d. entries. The entry has either a standard normal distribution, or a t_3-distribution, or a standard Cauchy distribution, i.e. t_1. Then we use ε as the response, or equivalently assume $\beta^* = 0$, and obtain the M-estimators $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$. Repeating this procedure for $K = 100$ times results in K replications in three cases. Then we extract the first coordinate of each estimator, denoted by $\{\hat{\beta}^{(0)}_{k,1}\}_{k=1}^K$, $\{\hat{\beta}^{(1)}_{k,1}\}_{k=1}^K$, $\{\hat{\beta}^{(2)}_{k,1}\}_{k=1}^K$. Then the two-sample Kolmogorov–Smirnov statistics can be obtained by

$$\text{KS}_1 = \sqrt{\frac{n}{2}} \max_x \left|\hat{F}^{(0)}_n(x) - \hat{F}^{(1)}_n(x)\right|, \quad \text{KS}_2 = \sqrt{\frac{n}{2}} \max_x \left|\hat{F}^{(0)}_n(x) - \hat{F}^{(2)}_n(x)\right|,$$

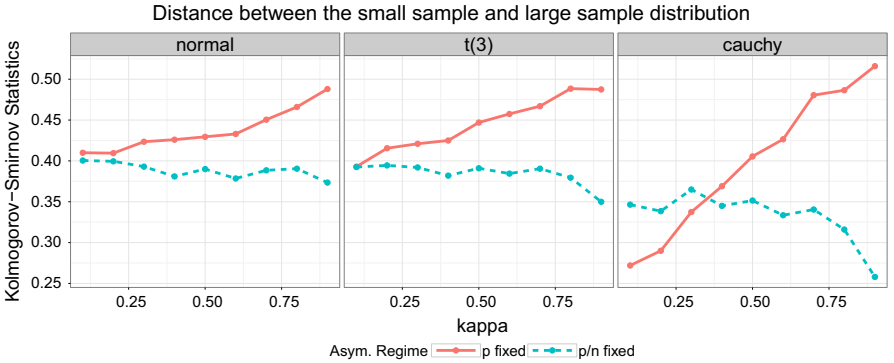

Fig. 1 Approximation accuracy of p-fixed asymptotics and p/n-fixed asymptotics: each column represents an error distribution; the x-axis represents the ratio κ of the dimension and the sample size and the y-axis represents the Kolmogorov–Smirnov statistic; the red solid line corresponds to p-fixed approximation and the blue dashed line corresponds to p/n-fixed approximation (color figure online)

where $\hat{F}_n^{(r)}$ is the empirical distribution of $\{\hat{\beta}_{k,1}^{(r)}\}_{k=1}^K$. We can then compare the accuracy of two asymptotic regimes by comparing KS_1 and KS_2. The smaller the value of KS_i, the better the approximation.

Figure 1 displays the results for these error distributions. We see that for gaussian errors and even t_3 errors, the p/n-fixed/moderate-dimensional approximation is uniformly more accurate than the widely used p-fixed/low-dimensional approximation. For Cauchy errors, the low-dimensional approximation performs better than the moderate-dimensional one when p/n is small but worsens when the ratio is large especially when p/n is close to 1. Moreover, when p/n grows, the two approximations have qualitatively different behaviors: the p-fixed approximation becomes less and less accurate while the p/n-fixed approximation does not suffer much deterioration when p/n grows. The qualitative and quantitative differences of these two approximations reveal the practical importance of exploring the p/n-fixed asymptotic regime. (See also [33]).

2.2 Random versus fixed design?

As discussed in Sect. 1.1, assuming a fixed design or a random design could lead to qualitatively different inferential results.

In the random design setting, X is considered as being generated from a super population. For example, the rows of X can be regarded as an i.i.d. sample from a distribution known, or partially known, to the researcher. In situations where one uses techniques such as cross-validation [54], pairs bootstrap in regression [17] or sample splitting [60], the researcher effectively assumes exchangeability of the data $(x_i^T, y_i)_{i=1}^n$. Naturally, this is only compatible with an assumption of random design. Given the extremely widespread use of these techniques in contemporary machine learning and statistics, one could argue that the random design setting is the one under which most of modern statistics is carried out, especially for prediction problems.

Furthermore, working under a random design assumption forces the researcher to take into account two sources of randomness as opposed to only one in the fixed design case. Hence working under a random design assumption should yield conservative confidence intervals for β_j^*.

In other words, in settings where the researcher collects data without control over the values of the predictors, the random design assumption is arguably the more natural one of the two.

However, it has now been understood for almost a decade that common random design assumptions in high-dimension (e.g. $x_i = \Sigma^{1/2} z_i$ where $z_{i,j}$'s are i.i.d with mean 0 and variance 1 and a few moments and Σ "well behaved") suffer from considerable geometric limitations, which have substantial impacts on the performance of the estimators considered in this paper [22]. As such, confidence statements derived from that kind of analysis can be relied on only after performing a few graphical tests on the data (see [19]). These geometric limitations are simple consequences of the concentration of measure phenomenon [36].

On the other hand, in the fixed design setting, X is considered a fixed matrix. In this case, the inference only takes the randomness of ε into consideration. This perspective is popular in several situations. The first one is the experimental design. The goal is to study the effect of a set of factors, which can be controlled by the experimenter, on the response. In contrast to the observational study, the experimenter can design the experimental condition ahead of time based on the inference target. For instance, a one-way ANOVA design encodes the covariates into binary variables (see Sect. 3.3.4 for details) and it is fixed prior to the experiment. Other examples include two-way ANOVA designs, factorial designs, Latin-square designs, etc. [52].

Another situation which is concerned with fixed design is the survey sampling where the inference is carried out conditioning on the data [13]. Generally, in order to avoid unrealistic assumptions, making inference conditioning on the design matrix X is necessary. Suppose the linear model (1) is true and identifiable (see Sect. 2.3 for details), then all information of β^* is contained in the conditional distribution $\mathscr{L}(y|X)$ and hence the information in the marginal distribution $\mathscr{L}(X)$ is redundant. The conditional inference framework is more robust to the data generating procedure due to the irrelevance of $\mathscr{L}(X)$.

Also, results based on fixed design assumptions may be preferable from a theoretical point of view in the sense that they could potentially be used to establish corresponding results for certain classes of random designs. Specifically, given a marginal distribution $\mathscr{L}(X)$, one only has to prove that $\mathscr{X}$ satisfies the assumptions for fixed design with high probability.

In conclusion, fixed and random design assumptions play complementary roles in moderate-dimensional settings. We focus on the least understood of the two, the fixed design case, in this paper.

2.3 Modeling and identification of parameters

The problem of identifiability is especially important in the fixed design case. Define $\beta^*(\rho)$ in the population as

$$\beta^*(\rho) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho\left(y_i - x_i^T \beta\right). \tag{3}$$

One may ask whether $\beta^*(\rho) = \beta^*$ regardless of ρ in the fixed design case. We provide an affirmative answer in the following proposition by assuming that ε_i has a symmetric distribution around 0 and ρ is even.

Proposition 2.1 *Suppose X has a full column rank and $\varepsilon_i \overset{d}{=} -\varepsilon_i$ for all i. Further assume ρ is an even convex function such that for any $i = 1, 2, \ldots$ and $\alpha \neq 0$,*

$$\frac{1}{2}\left(\mathbb{E}\rho(\varepsilon_i - \alpha) + \mathbb{E}\rho(\varepsilon_i + \alpha)\right) > \mathbb{E}\rho(\varepsilon_i). \tag{4}$$

Then $\beta^*(\rho) = \beta^*$ regardless of the choice of ρ.

The proof is left to "Appendix C". It is worth mentioning that Proposition 2.1 only requires the marginals of ε to be symmetric but does not impose any constraint on the dependence structure of ε. Further, if ρ is strongly convex, then for all $\alpha \neq 0$,

$$\frac{1}{2}\left(\rho(x - \alpha) + \rho(x + \alpha)\right) > \rho(x).$$

As a consequence, the condition (4) is satisfied provided that ε_i is non-zero with positive probability.

If ε is asymmetric, we may still be able to identify β^* if ε_i are i.i.d. random variables. In contrast to the last case, we should incorporate an intercept term as a shift towards the centroid of ρ. More precisely, we define $\alpha^*(\rho)$ and $\beta^*(\rho)$ as

$$(\alpha^*(\rho), \beta^*(\rho)) = \arg\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho\left(y_i - \alpha - x_i^T \beta\right).$$

Proposition 2.2 *Suppose $(\mathbf{1}, X)$ is of full column rank and ε_i are i.i.d. such that $\mathbb{E}\rho(\varepsilon_1 - \alpha)$ as a function of α has a unique minimizer $\alpha(\rho)$. Then $\beta^*(\rho)$ is uniquely defined with $\beta^*(\rho) = \beta^*$ and $\alpha^*(\rho) = \alpha(\rho)$.*

The proof is left to "Appendix C". For example, let $\rho(z) = |z|$. Then the minimizer of $\mathbb{E}\rho(\varepsilon_1 - a)$ is a median of ε_1, and is unique if ε_1 has a positive density. It is worth pointing out that incorporating an intercept term is essential for identifying β^*. For instance, in the least-square case, $\beta^*(\rho)$ no longer equals to β^* if $\mathbb{E}\varepsilon_i \neq 0$. Proposition 2.2 entails that the intercept term guarantees $\beta^*(\rho) = \beta^*$, although the intercept term itself depends on the choice of ρ unless more conditions are imposed.

If ε_i's are neither symmetric nor i.i.d., then β^* cannot be identified by the previous criteria because $\beta^*(\rho)$ depends on ρ. Nonetheless, from a modeling perspective, it is popular and reasonable to assume that ε_i's are symmetric or i.i.d. in many situations. Therefore, Propositions 2.1 and 2.2 justify the use of M-estimators in those cases and M-estimators derived from different loss functions can be compared because they are estimating the same parameter.

3 Main results

3.1 Notation and assumptions

Let $x_i^T \in \mathbb{R}^{1\times p}$ denote the i-th row of X and $X_j \in \mathbb{R}^{n\times 1}$ denote the j-th column of X. Throughout the paper we will denote by $X_{ij} \in \mathbb{R}$ the (i, j)-th entry of X, by $X_{[j]} \in \mathbb{R}^{n\times(p-1)}$ the design matrix X after removing the j-th column, and by $x_{i,[j]}^T \in \mathbb{R}^{1\times(p-1)}$ the vector x_i^T after removing j-th entry. The M-estimator $\hat{\beta}(\rho)$ associated with the loss function ρ is defined as

$$\hat{\beta}(\rho) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho\left(y_k - x_k^T\beta\right) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho\left(\varepsilon_k - x_k^T(\beta-\beta^*)\right) \quad (5)$$

We define $\psi = \rho'$ to be the first derivative of ρ. We will write $\hat{\beta}(\rho)$ simply $\hat{\beta}$ when no confusion can arise.

When the original design matrix X does not contain an intercept term, we can simply replace X by $(\mathbf{1}, X)$ and augment β into a $(p+1)$-dimensional vector $(\alpha, \beta^T)^T$. Although being a special case, we will discuss the question of intercept in Sect. 3.2.2 due to its important role in practice.

Equivariance and reduction to the null case

Notice that our target quantity $\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}$ is invariant to the choice of β^*, provided that β^* is identifiable as discussed in Sect. 2.3, we can assume $\beta^* = 0$ without loss of generality. In this case, we assume in particular that the design matrix X has full column rank. Then $y_k = \varepsilon_k$ and

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho\left(\varepsilon_k - x_k^T\beta\right).$$

Similarly we define the leave-j-th-predictor-out version as

$$\hat{\beta}_{[j]} = \underset{\beta\in\mathbb{R}^{p-1}}{\arg\min} \frac{1}{n}\sum_{k=1}^{n} \rho\left(\varepsilon_k - x_{k,[j]}^T\beta\right).$$

Based on these notations we define the full residuals R_k as

$$R_k = \varepsilon_k - x_k^T \hat{\beta}, \quad k = 1, 2, \ldots, n$$

and the leave-j-th-predictor-out residual as

$$r_{k,[j]} = \varepsilon_k - x_{k,[j]}^T \hat{\beta}_{[j]}, \quad k = 1, 2, \ldots, n, \quad j = 1, \ldots, p.$$

Three $n \times n$ diagonal matrices are defined as

$$D = \operatorname{diag}(\psi'(R_k))_{k=1}^n, \quad \tilde{D} = \operatorname{diag}(\psi''(R_k))_{k=1}^n, \quad D_{[j]} = \operatorname{diag}(\psi'(r_{k,[j]}))_{k=1}^n. \tag{6}$$

We say a random variable Z is σ^2-sub-gaussian if for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda Z} \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

In addition, we use $J_n \subset \{1, \ldots, p\}$ to represent the indices of parameters which are of interest. Intuitively, more entries in J_n would require more stringent conditions for the asymptotic normality.

Finally, we adopt Landau's notation ($O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, $o_p(\cdot)$). In addition, we say $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ and similarly, we say $a_n = \Omega_p(b_n)$ if $b_n = O_p(a_n)$. To simplify the logarithm factors, we use the symbol polyLog(n) to denote any factor that can be upper bounded by $(\log n)^\gamma$ for some $\gamma > 0$. Similarly, we use $\frac{1}{\text{polyLog(n)}}$ to denote any factor that can be lower bounded by $\frac{1}{(\log n)^{\gamma'}}$ for some $\gamma' > 0$.

3.2 Technical assumptions and main result

Before stating the assumptions, we need to define several quantities of interest. Let

$$\lambda_+ = \lambda_{\max}\left(\frac{X^T X}{n}\right), \quad \lambda_- = \lambda_{\min}\left(\frac{X^T X}{n}\right)$$

be the largest (resp. smallest) eigenvalue of the matrix $\frac{X^T X}{n}$. Let $e_i \in \mathbb{R}^n$ be the i-th canonical basis vector and

$$h_{j,0} \triangleq (\psi(r_{1,[j]}), \ldots, \psi(r_{n,[j]}))^T,$$
$$h_{j,1,i} \triangleq \left(I - D_{[j]}X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T\right) e_i.$$

Finally, let

$$\Delta_C = \max\left\{\max_{j\in J_n} \frac{\left|h_{j,0}^T X_j\right|}{\|h_{j,0}\|_2}, \max_{i\leq n, j\in J_n} \frac{\left|h_{j,1,i}^T X_j\right|}{\|h_{j,1,i}\|_2}\right\},$$
$$Q_j = \operatorname{Cov}(h_{j,0})$$

Based on the quantities defined above, we state our technical assumptions on the design matrix X followed by the main result. A detailed explanation of the assumptions follows.

A1 $\rho(0) = \psi(0) = 0$ and there exists positive numbers $K_0 = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$, $K_1, K_2 = O\,(\text{polyLog(n)})$, such that for any $x \in \mathbb{R}$,

$$K_0 \le \psi'(x) \le K_1, \quad \left|\frac{d}{dx}(\sqrt{\psi'}(x))\right| = \frac{|\psi''(x)|}{\sqrt{\psi'(x)}} \le K_2;$$

A2 $\varepsilon_i = u_i(W_i)$ where $(W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and u_i are smooth functions with $\|u_i'\|_\infty \le c_1$ and $\|u_i''\|_\infty \le c_2$ for some $c_1, c_2 = O(\text{polyLog(n)})$. Moreover, assume $\min_i \text{Var}(\varepsilon_i) = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$.

A3 $\lambda_+ = O(\text{polyLog(n)})$ and $\lambda_- = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$;

A4 $\min_{j\in J_n} \frac{X_j^T Q_j X_j}{\text{tr}(Q_j)} = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$;

A5 $\mathbb{E}\Delta_C^8 = O\,(\text{polyLog(n)})$.

Theorem 3.1 *Under assumptions A1–A5, as $p/n \to \kappa$ for some $\kappa \in (0, 1)$, while $n \to \infty$,*

$$\max_{j\in J_n} d_{\text{TV}}\left(\mathscr{L}\left(\frac{\hat\beta_j - \mathbb{E}\hat\beta_j}{\sqrt{\text{Var}(\hat\beta_j)}}\right), N(0, 1)\right) = o(1),$$

where $d_{\text{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$ is the total variation distance.

We provide several examples where our assumptions hold in Sect. 3.3. We also provide an example where the asymptotic normality does not hold in Sect. 3.3.4. This shows that our assumptions are not just artifacts of the proof technique we developed, but that there are (probably many) situations where asymptotic normality will not hold, even coordinate-wise.

3.2.1 Discussion of assumptions

Now we discuss assumptions **A**1–**A**5. Assumption **A**1 implies the boundedness of the first-order and the second-order derivatives of ψ. The upper bounds are satisfied by most loss functions including the L_2 loss, the smoothed L_1 loss, the smoothed Huber loss, etc. The non-zero lower bound K_0 implies the strong convexity of ρ and is required for technical reasons. It can be removed by considering first a ridge-penalized M-estimator and taking appropriate limits as in [20,21]. In addition, in this paper we consider the smooth loss functions and the results can be extended to non-smooth case via approximation.

For unregularized M-estimators, the strong convexity is also assumed by other works [15,20]. However, we believe that this assumption is unnecessary and can

be removed at least for well-behaved design matrices. In fact, we can extend our results to strictly convex loss functions, where ψ' is always positive by imposing slightly stronger assumptions on the designs. This includes the class of optimal loss functions in the moderate p/n regime derived in [6]. However, the proofs are very delicate and beyond the scope of this paper so we plan to leave it in our future works.

Assumption **A**2 was proposed in [10] when deriving the second-order Poincaré inequality discussed in Sect. 4.1. It means that the results apply to non-Gaussian distributions, such as the uniform distribution on [0, 1] by taking $u_i = \Phi$, the cumulative distribution function of standard normal distribution. Through the gaussian concentration [36], we see that **A**2 implies that ε_i are c_1^2-sub-gaussian. Thus **A**2 controls the tail behavior of ε_i. The bounds on the infinity norm of u_i' and u_i'' are required only for the direct application of Chatterjee's results. In fact, a look at his proof suggests that one can obtain a similar result to his Second-Order Poincaré inequality involving moment bounds on $u_i'(W_i)$ and $u_i''(W_i)$. This would be a way to weaken our assumptions to permit to have the heavy-tailed distributions expected in robustness studies. This requires substantial work and an extension of the main results of [10]. Because the technical part of the paper is already long, we leave this interesting statistical question to future works.

On the other hand, since we are considering strongly convex loss-functions, it is not completely unnatural to restrict our attention to light-tailed errors. Furthermore, efficiency—and not only robustness—questions are one of the main reasons to consider these estimators in the moderate-dimensional context. The potential gains in efficiency obtained by considering regression M-estimates [6] apply in the light-tailed context, which further justify our interest in this theoretical setup.

Assumption **A**3 is completely checkable since it only depends on X. It controls the singularity of the design matrix. Under **A**1 and **A**3, it can be shown that the objective function is strongly convex with curvature (the smallest eigenvalue of the Hessian matrix) lower bounded by $\Omega\left(\frac{1}{\text{polyLog(n)}}\right)$ everywhere.

Assumption **A**4 is controlling the left tail of quadratic forms. It is fundamentally connected to aspects of the concentration of measure phenomenon [36]. This condition is proposed and emphasized under the random design setting by [23]. Essentially, it means that for a matrix Q_j ,which does not depend on X_j, the quadratic form $X_j^T Q_j X_j$ should have the same order as $\text{tr}(Q_j)$.

Assumption **A**5 is proposed by [20] under the random design settings. It is motivated by leave-one-predictor-out analysis. Note that Δ_C is the maximum of linear contrasts of X_j, whose coefficients do not depend on X_j. It is easily checked for design matrix X which is a realization of a random matrix with i.i.d sub-gaussian entries for instance.

Remark 3.1 In certain applications, it is reasonable to make the following additional assumption:

A6 ρ is an even function and ε_i's have symmetric distributions.

Although assumption **A**6 is not necessary to Theorem 3.1, it can simplify the result. Under assumption **A**6, when X is full rank, we have, if $\overset{d}{=}$ denotes equality in distri-

bution,

$$\hat{\beta} - \beta^* = \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\varepsilon_i - x_i^T \eta\right) = \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho\left(-\varepsilon_i + x_i^T \eta\right)$$
$$\stackrel{d}{=} \arg\min_{\eta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\varepsilon_i + x_i^T \eta\right) = \beta^* - \hat{\beta}.$$

This implies that $\hat{\beta}$ is an unbiased estimator, provided it has a mean, which is the case here. Unbiasedness is useful in practice, since then Theorem 3.1 reads

$$\max_{j \in J_n} d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1)\ .$$

For inference, we only need to estimate the asymptotic variance.

3.2.2 An important remark concerning Theorem 3.1

When J_n is a subset of $\{1, \ldots, p\}$, the coefficients in J_n^c become nuisance parameters. Heuristically, in order for identifying $\beta_{J_n}^*$, one only needs the subspaces $\mathrm{span}(X_{J_n})$ and $\mathrm{span}(X_{J_n^c})$ to be distinguished and X_{J_n} has a full column rank. Here X_{J_n} denotes the sub-matrix of X with columns in J_n. Formally, let

$$\hat{\Sigma}_{J_n} = \frac{1}{n} X_{J_n}^T \left(I - X_{J_n^c}\left(X_{J_n^c}^T X_{J_n^c}\right)^{-} X_{J_n^c}^T\right) X_{J_n}$$

where A^- denotes the generalized inverse of A, and

$$\tilde{\lambda}_+ = \lambda_{\max}\left(\hat{\Sigma}_{J_n}\right), \quad \tilde{\lambda}_- = \lambda_{\min}\left(\hat{\Sigma}_{J_n}\right).$$

Then $\hat{\Sigma}_{J_n}$ characterizes the behavior of X_{J_n} after removing the effect of $X_{J_n^c}$. In particular, we can modify the assumption **A**3 by

A3* $\tilde{\lambda}_+ = O(\mathrm{polyLog(n)})$ and $\tilde{\lambda}_- = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$.

Then we are able to derive a stronger result in the case where $|J_n| < p$ than Theorem 3.1 as follows.

Corollary 3.1 *Under assumptions **A**1–2, **A**4–5 and **A**3*, as $p/n \to \kappa$ for some $\kappa \in (0, 1)$,*

$$\max_{j \in J_n} d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

It can be shown that $\tilde{\lambda}_+ \leq \lambda_+$ and $\tilde{\lambda}_- \geq \lambda_-$ and hence the assumption **A**3* is weaker than **A**3. It is worth pointing out that the assumption **A**3* even holds when $X^c_{J_n}$ does not have full column rank, in which case $\beta^*_{J_n}$ is still identifiable and $\hat{\beta}_{J_n}$ is still well-defined, although $\beta^*_{J_n^c}$ and $\hat{\beta}_{J_n^c}$ are not; see "Appendix C-2" for details.

3.3 Examples

Throughout this subsection (except Sect. 3.3.4), we consider the case where X is a realization of a random matrix, denoted by Z (to be distinguished from X). We will verify that the assumptions **A**3–**A**5 are satisfied with high probability under different regularity conditions on the distribution of Z. This is a standard way to justify the conditions for fixed design [42,43] in the literature on regression M-estimates.

3.3.1 Random design with independent entries

First we consider a random matrix Z with i.i.d. sub-gaussian entries.

Proposition 3.1 *Suppose Z has i.i.d. mean-zero σ^2-sub-gaussian entries with $\mathrm{Var}(Z_{ij}) = \tau^2 > 0$ for some $\sigma = O(\mathrm{polyLog(n)})$ and $\tau = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$, then, when X is a realization of Z, assumptions **A**3–**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

In practice, the assumption of identical distribution might be invalid. In fact the assumptions **A**4, **A**5 and the first part of **A**3 ($\lambda_+ = O(\mathrm{polyLog(n)})$) are still satisfied with high probability if we only assume the independence between entries and boundedness of certain moments. To control λ_-, we rely on [37] which assumes symmetry of each entry. We obtain the following result based on it.

Proposition 3.2 *Suppose Z has independent σ^2-sub-gaussian entries with*

$$Z_{ij} \overset{d}{=} -Z_{ij}, \quad \mathrm{Var}(Z_{ij}) > \tau^2$$

*for some $\sigma = O(\mathrm{polyLog(n)})$ and $\tau = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$, then, when X is a realization of Z, assumptions **A**3–**A**5 for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

Under the conditions of Proposition 3.2, we can add an intercept term into the design matrix. Adding an intercept allows us to remove the mean-zero assumption for Z_{ij}'s. In fact, suppose Z_{ij} is symmetric with respect to μ_j, which is potentially non-zero, for all i, then according to Sect. 3.2.2, we can replace Z_{ij} by $Z_{ij} - \mu_j$ and Proposition 3.3 can be then applied.

Proposition 3.3 *Suppose $Z = (\mathbf{1}, \tilde{Z})$ and $\tilde{Z} \in \mathbb{R}^{n\times(p-1)}$ has independent σ^2-sub-gaussian entries with*

$$\tilde{Z}_{ij} - \mu_j \overset{d}{=} \mu_j - \tilde{Z}_{ij}, \quad \mathrm{Var}(\tilde{Z}_{ij}) > \tau^2$$

for some $\sigma = O(\text{polyLog(n)})$, $\tau = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$ *and arbitrary* μ_j. *Then, when X is a realization of Z, assumptions* ***A3****, **A4** and **A5** for X are satisfied with high probability over Z for* $J_n = \{2, \ldots, p\}$.

3.3.2 Dependent gaussian design

To show that our assumptions handle a variety of situations, we now assume that the observations, namely the rows of Z, are i.i.d. random vectors with a covariance matrix Σ. In particular we show that the Gaussian design, i.e. $z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$, satisfies the assumptions with high probability.

Proposition 3.4 *Suppose* $z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$ *with* $\lambda_{\max}(\Sigma) = O(\text{polyLog(n)})$ *and* $\lambda_{\min}(\Sigma) = \Omega\left(\frac{1}{\text{polyLog(n)}}\right)$, *then, when X is a realization of Z, assumptions* **A3**–**A5** *for X are satisfied with high probability over Z for* $J_n = \{1, \ldots, p\}$.

This result extends to the matrix-normal design [41, Chapter 3], i.e. $(Z_{ij})_{i \le n, j \le p}$ is one realization of a np-dimensional random variable Z with multivariate gaussian distribution

$$\text{vec}(Z) \triangleq \left(z_1^T, z_2^T, \ldots, z_n^T\right) \sim N(0, \Lambda \otimes \Sigma),$$

and $\otimes$ is the Kronecker product. It turns out that assumptions **A**3–**A**5 are satisfied if both Λ and Σ are well-behaved.

Proposition 3.5 *Suppose Z is matrix-normal with* $\text{vec}(Z) \sim N(0, \Lambda \otimes \Sigma)$ *and*

$$\lambda_{\max}(\Lambda), \lambda_{\max}(\Sigma) = O(\text{polyLog(n)}), \quad \lambda_{\min}(\Lambda), \lambda_{\min}(\Sigma) = \Omega\left(\frac{1}{\text{polyLog(n)}}\right).$$

Then, when X is a realization of Z,assumptions ***A3–A5*** *for X are satisfied with high probability over Z for* $J_n = \{1, \ldots, p\}$.

In order to incorporate an intercept term, we need slightly more stringent condition on Λ. Instead of assumption **A**3, we prove that assumption **A**3*—see Sect. 3.2.2—holds with high probability.

Proposition 3.6 *Suppose Z contains an intercept term, i.e.* $Z = (\boldsymbol{1}, \tilde{Z})$ *and* $\tilde{Z}$ *satisfies the conditions of Proposition 3.5. Further assume that*

$$\frac{\max_i |(\Lambda^{-\frac{1}{2}}\boldsymbol{1})_i|}{\min_i |(\Lambda^{-\frac{1}{2}}\boldsymbol{1})_i|} = O(\text{polyLog(n)}). \tag{7}$$

Then, when X is a realization of Z, assumptions ***A3****, **A4** and **A5** for X are satisfied with high probability over Z for* $J_n = \{2, \ldots, p\}$.

When $\Lambda = I$, the condition (7) is satisfied. Another non-trivial example is the exchangeable case where Λ_{ij} are all equal for $i \neq j$. In this case, $\mathbf{1}$ is an eigenvector of Λ and hence it is also an eigenvector of $\Lambda^{-\frac{1}{2}}$. Thus $\Lambda^{-\frac{1}{2}}\mathbf{1}$ is a multiple of $\mathbf{1}$ and the condition (7) is satisfied.

3.3.3 Elliptical design

Furthermore, we can move from Gaussian-like structure to generalized elliptical models where $z_i = \zeta_i \Sigma^{1/2} \mathscr{Z}_i$ where $\{\zeta_i, \mathscr{Z}_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are independent random variables, $\mathscr{Z}_{ij}$ having for instance mean 0 and variance 1. The elliptical family is quite flexible in modeling data. It represents a type of data formed by a common driven factor and independent individual effects. It is widely used in multivariate statistics [1,55] and various fields, including finance [12] and biology [46]. In the context of high-dimensional statistics, this class of model was used to refute universality claims in random matrix theory [18]. In robust regression, [22] used elliptical models to show that the limit of $\|\hat{\beta}\|_2^2$ depends on the distribution of ζ_i and hence the geometry of the predictors. As such, studies limited to Gaussian-like design were shown to be of very limited statistical interest. See also the deep classical inadmissibility results [4,34]. However, as we will show in the next proposition, the common factors ζ_i do not distort the shape of the asymptotic distribution. A similar phenomenon happens in the random design case—see [6,23].

Proposition 3.7 *Suppose Z is generated from an elliptical model, i.e.*

$$Z_{ij} = \zeta_i \mathscr{Z}_{ij},$$

*where ζ_i are independent random variables taking values in $[a, b]$ for some $0 < a < b < \infty$ and $\mathscr{Z}_{ij}$ are independent random variables satisfying the conditions of Propositions 3.1 or 3.2. Further assume that $\{\zeta_i : i = 1, \ldots, n\}$ and $\{\mathscr{Z}_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are independent. Then, when X is a realization of Z, assumptions **A3–A5** for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

Thanks to the fact that ζ_i is bounded away from 0 and ∞, the proof of Proposition 3.7 is straightforward, as shown in "Appendix C". However, by a more refined argument and assuming identical distributions ζ_i, we can relax this condition.

Proposition 3.8 *Under the conditions of Proposition 3.7 (except the boundedness of ζ_i) and assume ζ_i are i.i.d. samples generated from some distribution F, independent of n, with*

$$P(\zeta_1 \geq t) \leq c_1 e^{-c_2 t^\alpha},$$

*for some fixed $c_1, c_2, \alpha > 0$ and $F^{-1}(q) > 0$ for any $q \in (0, 1)$ where F^{-1} is the quantile function of F and is continuous. Then, when X is a realization of Z, assumptions **A3–A5** for X are satisfied with high probability over Z for $J_n = \{1, \ldots, p\}$.*

3.3.4 A counterexample

Consider a one-way ANOVA situation. In other words, let the design matrix have exactly 1 non-zero entry per row, whose value is 1. Let $\{k_i\}_{i=1}^n$ be integers in $\{1, \ldots, p\}$. And let $X_{i,j} = 1(j = k_i)$. Furthermore, let us constrain $n_j = |\{i : k_i = j\}|$ to be such that $1 \le n_j \le 2\lfloor p/n \rfloor$. Taking for instance $k_i = (i \mod p)$ is an easy way to produce such a matrix. The associated statistical model is just $y_i = \varepsilon_i + \beta^*_{k_i}$.

It is easy to see that

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}} \sum_{i:k_i=j} \rho(y_i - \beta_j) = \arg\min_{\beta \in \mathbb{R}} \sum_{i:k_i=j} \rho(\varepsilon_i - (\beta_j - \beta^*_j)) \, .$$

This is of course a standard location problem. In the moderate-dimensional setting we consider, n_j remains finite as $n \to \infty$. So $\hat{\beta}_j$ is a non-linear function of finitely many random variables and will in general not be normally distributed.

For concreteness, one can take $\rho(x) = |x|$, in which case $\hat{\beta}_j$ is a median of $\{y_i\}_{\{i:k_i=j\}}$. The cdf of $\hat{\beta}_j$ is known exactly by elementary order statistics computations (see [14]) and is not that of a Gaussian random variable in general. In fact, the ANOVA design considered here violates the assumption **A**3 since $\lambda_- = \min_j n_j/n = O(1/n)$. Further, we can show that the assumption **A**5 is also violated, at least in the least-square case; see Sect. 5.1 for details.

3.4 Comments and discussions

3.4.1 Asymptotic normality in high dimensions

In the p-fixed regime, the asymptotic distribution is easily defined as the limit of $\mathscr{L}(\hat{\beta})$ in terms of weak topology [56]. However, in regimes where the dimension p grows, the notion of asymptotic distribution is more delicate. a conceptual question arises from the fact that the dimension of the estimator $\hat{\beta}$ changes with n and thus there is no well-defined distribution which can serve as the limit of $\mathscr{L}(\hat{\beta})$, where $\mathscr{L}(\cdot)$ denotes the law. One remedy is proposed by [38]. Under this framework, a triangular array $\{W_{n,j}, j = 1, 2, \ldots, p_n\}$, with $\mathbb{E}W_{n,j} = 0$, $\mathbb{E}W^2_{n,j} = 1$, is called jointly asymptotically normal if for any deterministic sequence $a_n \in \mathbb{R}^{p_n}$ with $\|a_n\|_2 = 1$,

$$\mathscr{L}\left(\sum_{j=1}^{p_n} a_{n,j} W_{n,j}\right) \to N(0, 1).$$

When the zero mean and unit variance are not satisfied, it is easy to modify the definition by normalizing random variables.

Definition 3.1 (*joint asymptotic normality*)
$\{W_n : W_n \in \mathbb{R}^{p_n}\}$ is jointly asymptotically normal if and only if for any sequence $\{a_n : a_n \in \mathbb{R}^{p_n}\}$,

$$\mathscr{L}\left(\frac{a_n^T(W_n - \mathbb{E}W_n)}{\sqrt{a_n^T \operatorname{Cov}(W_n) a_n}}\right) \to N(0,1).$$

The above definition of asymptotic normality is strong and appealing but was shown not to hold for least-squares in the moderate p/n regime [31]. In fact, [31] shows that $\hat{\beta}^{LS}$ is jointly asymtotically normal only if

$$\max_i (X(X^TX)^{-1}X^T)_{i,i} \to 0.$$

When $p/n \to \kappa \in (0,1)$, provided X is full rank,

$$\max_i (X(X^TX)^{-1}X^T)_{i,i} \geq \frac{1}{n}\operatorname{tr}(X(X^TX)^{-1}X^T) = \frac{p}{n} \to \kappa > 0.$$

In other words, in moderate p/n regime, the asymptotic normality cannot hold for all linear contrasts, even in the case of least-squares.

In applications, however, it is usually not necessary to consider all linear contrasts but instead a small subset of them, e.g. all coordinates or low dimensional linear contrasts such as $\beta_1^* - \beta_2^*$. We can naturally modify Definition 3.1 and adapt to our needs by imposing constraints on a_n. A popular concept, which we use in Sect. 1 informally, is called coordinate-wise asymptotic normality and defined by restricting a_n to be the canonical basis vectors, which have only one non-zero element. An equivalent definition is stated as follows.

Definition 3.2 (*coordinate-wise asymptotic normal*) $\{W_n : W_n \in \mathbb{R}^{p_n}\}$ is coordinate-wise asymptotically normal if and only if for any sequence $\{j_n : j_n \in \{1, \ldots, p_n\}\}$,

$$\mathscr{L}\left(\frac{W_{n,j_n} - \mathbb{E}W_{n,j_n}}{\sqrt{\operatorname{Var}(W_{n,j_n})}}\right) \to N(0,1).$$

A more convenient way to define the coordinate-wise asymptotic normality is to introduce a metric $d(\cdot,\cdot)$, e.g. Kolmogorov distance and total variation distance, which induces the weak convergence topology. Then W_n is coordinate-wise asymptotically normal if and only if

$$\max_j d\left(\mathscr{L}\left(\frac{W_{n,j} - \mathbb{E}W_{n,j}}{\sqrt{\operatorname{Var}(W_{n,j})}}\right), N(0,1)\right) = o(1).$$

3.4.2 Variance and bias estimation

To complete the inference, we need to compute the bias and variance. As discussed in Remark 3.1, the M-estimator is unbiased if the loss function and the error distribution

are symmetric. For the variance, it is easy to get a conservative estimate via resampling methods such as Jackknife as a consequence of Efron–Stein's inequality; see [20,24] for details. Moreover, by the variance decomposition formula,

$$\mathrm{Var}(\hat{\beta}_j) = \mathbb{E}\left[\mathrm{Var}(\hat{\beta}_j|X)\right] + \mathrm{Var}\left[\mathbb{E}(\hat{\beta}_j|X)\right] \geq \mathbb{E}\left[\mathrm{Var}(\hat{\beta}_j|X)\right],$$

the unconditional variance, when X is a random design matrix, is a conservative estimate. The unconditional variance can be calculated by solving a non-linear system; see [15,20].

However, estimating the exact variance is known to be hard. [24] show that the existing resampling schemes, including jacknife, pairs-bootstrap, residual bootstrap, etc., are either too conservative or too anti-conservative when p/n is large. The challenge, as mentioned in [20,24], is due to the fact that the residuals $\{R_i\}$ do not mimic the behavior of $\{\varepsilon_i\}$ and that the resampling methods effectively modifies the geometry of the dataset from the point of view of the statistics of interest. We believe that variance estimation in moderate p/n regime should rely on different methodologies from the ones used in low-dimensional estimation.

4 Proof sketch

Since the proof of Theorem 3.1 is somewhat technical, we illustrate the main idea in this section.

First notice that the M-estimator $\hat{\beta}$ is an implicit function of independent random variables $\varepsilon_1, \ldots, \varepsilon_n$, which is determined by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \psi(\varepsilon_i - x_i\hat{\beta}) = 0. \tag{8}$$

The Hessian matrix of the loss function in (5) is $\frac{1}{n}X^T DX \succeq D_0\lambda_- I_p$ under the notation introduced in Sect. 3.1. The assumption **A**3 then implies that the loss function is strongly convex, in which case $\hat{\beta}$ is unique. Then $\hat{\beta}$ can be seen as a non-linear function of ε_i's. A powerful central limit theorem for this type of statistics is the second-order Poincaré inequality (SOPI), developed in [10] and used there to re-prove central limit theorems for linear spectral statistics of large random matrices. We recall one of the main results for the convenience of the reader.

Proposition 4.1 (SOPI; [10]) *Let* $\mathscr{W} = (\mathscr{W}_1, \ldots, \mathscr{W}_n) = (u_1(W_1), \ldots, u_n(W_n))$ *where* $W_i \overset{i.i.d.}{\sim} N(0,1)$ *and* $\|u_i'\|_\infty \leq c_1$, $\|u_i''\|_\infty \leq c_2$. *Take any* $g \in C^2(\mathbb{R}^n)$ *and let* $\nabla_i g$, ∇g *and* $\nabla^2 g$ *denote the* i*-th partial derivative, gradient and Hessian of* g. *Let*

$$\kappa_0 = \left(\mathbb{E}\sum_{i=1}^{n}\left|\nabla_i g(\mathscr{W})\right|^4\right)^{\frac{1}{2}}, \quad \kappa_1 = (\mathbb{E}\|\nabla g(\mathscr{W})\|_2^4)^{\frac{1}{4}}, \quad \kappa_2 = (\mathbb{E}\|\nabla^2 g(\mathscr{W})\|_{\mathrm{op}}^4)^{\frac{1}{4}},$$

and $U = g(\mathscr{W})$. If U has finite fourth moment, then

$$d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{U-\mathbb{E}U}{\sqrt{\mathrm{Var}(U)}}\right), N(0,1)\right) \le \frac{2\sqrt{5}(c_1c_2\kappa_0 + c_1^3\kappa_1\kappa_2)}{\mathrm{Var}(U)}.$$

From (8), it is not hard to compute the gradient and Hessian of $\hat{\beta}_j$ with respect to ε. Recalling the definitions in Eq. (6) on p. 9, we have

Lemma 4.1 *Suppose $\psi \in C^2(\mathbb{R}^n)$, then*

$$\frac{\partial \hat{\beta}_j}{\partial \varepsilon^T} = e_j^T (X^T DX)^{-1} X^T D \tag{9}$$

$$\frac{\partial \hat{\beta}_j}{\partial \varepsilon \partial \varepsilon^T} = G^T \,\mathrm{diag}\left(e_j^T (X^T DX)^{-1} X^T \tilde{D}\right) G \tag{10}$$

where e_j is the j-th cononical basis vectors in $\mathbb{R}^p$ and

$$G = I - X(X^T DX)^{-1} X^T D.$$

Recalling the definitions of K_i's in Assumption **A**1 on p. 10, we can bound κ_0, κ_1 and κ_2 as follows.

Lemma 4.2 *Let $\kappa_{0j}, \kappa_{1j}, \kappa_{2j}$ defined as in Proposition 4.1 by setting $\mathscr{W} = \varepsilon$ and $g(\mathscr{W}) = \hat{\beta}_j$. Let*

$$M_j = \mathbb{E}\left\| e_j^T (X^T DX)^{-1} X^T D^{\frac{1}{2}} \right\|_\infty, \tag{11}$$

then

$$\kappa_{0j}^2 \le \frac{K_1^2}{(nK_0\lambda_-)^{\frac{3}{2}}} \cdot M_j, \quad \kappa_{1j}^4 \le \frac{K_1^2}{(nK_0\lambda_-)^2}, \quad \kappa_{2j}^4 \le \frac{K_2^4}{(nK_0\lambda_-)^{\frac{3}{2}}} \cdot \left(\frac{K_1}{K_0}\right)^4 \cdot M_j.$$

As a consequence of the second-order Poincaré inequality, we can bound the total variation distance between $\hat{\beta}_j$ and a normal distribution by M_j and $\mathrm{Var}(\hat{\beta}_j)$. More precisely, we prove the following Lemma.

Lemma 4.3 *Under Assumptions **A**1–**A**3,*

$$\max_j d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = O_p\left(\frac{\max_j (nM_j^2)^{\frac{1}{8}}}{n \cdot \min_j \mathrm{Var}(\hat{\beta}_j)} \cdot \mathrm{polyLog(n)}\right).$$

Lemma 4.3 is the key to prove Theorem 3.1. To obtain the coordinate-wise asymptotic normality, it is left to establish an upper bound for M_j and a lower bound for $\mathrm{Var}(\hat{\beta}_j)$. In fact, we can prove that

Lemma 4.4 *Under assumptions A1–A5,*

$$\max_j M_j = O\left(\frac{\text{polyLog(n)}}{n}\right), \quad \min_j \text{Var}(\hat{\beta}_j) = \Omega\left(\frac{1}{n \cdot \text{polyLog(n)}}\right).$$

Then Lemmas 4.3 and 4.4 together imply that

$$\max_j d_{\text{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = O\left(\frac{\text{polyLog(n)}}{n^{\frac{1}{8}}}\right) = o(1).$$

"Appendix A", provides a roadmap of the proof of Lemma 4.4 under a special case where the design matrix X is one realization of a random matrix with i.i.d. sub-gaussian entries. It also serves as an outline of the rigorous proof in "Appendix B".

4.1 Comment on the second-order Poincaré inequality

Notice that when g is a linear function such that $g(z) = \sum_{i=1}^n a_i z_i$, then the Berry–Esseen inequality [25] implies that

$$d_K\left(\mathscr{L}\left(\frac{W - \mathbb{E}W}{\sqrt{\text{Var}(W)}}\right), N(0,1)\right) \preceq \frac{\sum_{i=1}^n |a_i|^3}{\left(\sum_{i=1}^n a_i^2\right)^{\frac{3}{2}}},$$

where

$$d_K(F, G) = \sup_x |F(x) - G(x)|.$$

On the other hand, the second-order Poincaré inequality implies that

$$\begin{aligned} d_K\left(\mathscr{L}\left(\frac{W - \mathbb{E}W}{\sqrt{\text{Var}(W)}}\right), N(0,1)\right) &\leq d_{\text{TV}}\left(\mathscr{L}\left(\frac{W - \mathbb{E}W}{\sqrt{\text{Var}(W)}}\right), N(0,1)\right) \\ &\preceq \frac{\left(\sum_{i=1}^n a_i^4\right)^{\frac{1}{2}}}{\sum_{i=1}^n a_i^2}. \end{aligned}$$

This is slightly worse than the Berry–Esseen bound and requires stronger conditions on the distributions of variates but provides bounds for TV metric instead of Kolmogorov metric. This comparison shows that second-order Poincaré inequality can be regarded as a generalization of the Berry–Esseen bound for non-linear transformations of independent random variables.

5 Least-squares estimator

The Least-Squares Estimator is a special case of an M-estimator with $\rho(x) = \frac{1}{2}x^2$. Because the estimator can then be written explicitly, the analysis of its properties

is extremely simple and it has been understood for several decades (see arguments in e.g. [31, Lemma 2.1] and [32, Proposition 2.2]). In this case, the hat matrix $H = X(X^TX)^{-1}X^T$ captures all the problems associated with dimensionality in the problem. In particular, proving the asymptotic normality simply requires an application of the Lindeberg–Feller theorem.

It is however somewhat helpful to compare the conditions required for asymptotic normality in this simple case and the ones we required in the more general setup of Theorem 3.1. We do so briefly in this section.

5.1 Coordinate-wise asymptotic normality of LSE

Under the linear model (1), when X is full rank,

$$\hat{\beta}^{LS} = \beta^* + (X^TX)^{-1}X^T\varepsilon,$$

thus each coordinate of $\hat{\beta}^{LS}$ is a linear contrast of ε with zero mean. Instead of assumption **A**2, which requires ε_i to be sub-gaussian, we only need to assume $\max_i \mathbb{E}|\varepsilon_i|^3 < \infty$, under which the Berry–Essen bound for non-i.i.d. data [25] implies that

$$d_K\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) \preceq \frac{\|e_j(X^TX)^{-1}X^T\|_3^3}{\|e_j^T(X^TX)^{-1}X^T\|_2^3} \leq \frac{\|e_j(X^TX)^{-1}X^T\|_\infty}{\|e_j(X^TX)^{-1}X^T\|_2}.$$

This motivates us to define a matrix specific quantity $S_j(X)$ such that

$$S_j(X) = \frac{\left\|e_j^T(X^TX)^{-1}X^T\right\|_\infty}{\left\|e_j^T(X^TX)^{-1}X^T\right\|_2} \tag{12}$$

then the Berry–Esseen bound implies that $\max_{j\in J_n} S_j(X)$ determines the coordinate-wise asymptotic normality of $\hat{\beta}^{LS}$.

Theorem 5.1 *If* $\max_i \mathbb{E}|\varepsilon_i|^3 < \infty$, *then*

$$\max_{j\in J_n} d_K\left(\frac{\hat{\beta}_{LS,j} - \beta_{0,j}}{\sqrt{\mathrm{Var}(\hat{\beta}_{LS,j})}}, N(0,1)\right) \leq A \cdot \max_i \frac{\mathbb{E}|\varepsilon_i|^3}{\left(\mathbb{E}\varepsilon_i^2\right)^{\frac{3}{2}}} \cdot \max_{j\in J_n} S_j(X),$$

where A is an absolute constant and $d_K(\cdot,\cdot)$ is the Kolmogorov distance, defined as

$$d_K(F,G) = \sup_x |F(x) - G(x)|.$$

It turns out that $\max_{j\in J_n} S_j(X)$ plays in the least-squares setting the role of Δ_C in assumption **A**5. Since it has been known that a condition like $S_j(X) \to 0$ is necessary

for asymptotic normality of least-square estimators [31, Proposition 2.2], this shows in particular that our Assumption **A**5, or a variant, is also needed in the general case. See "Appendix C-4.1" for details.

5.2 Discussion

Naturally, checking the conditions for asymptotic normality is much easier in the least-squares case than in the general case under consideration in this paper. In particular:

1. Asymptotic normality conditions can be checked for a broader class of random design matrices. See "Appendix C-4.2" for details.
2. For orthogonal design matrices, i.e $X^T X = c\mathrm{Id}$ for some $c > 0$, $S_j(X) = \frac{\|X_j\|_\infty}{\|X_j\|_2}$. Hence, the condition $S_j(X) = o(1)$ is true if and only if no entry dominates the j-th row of X.
3. The ANOVA-type counterexample we gave in Sect. 3.3.4 still provides a counter-example. The reason now is different: namely the sum of finitely many independent random variables is evidently in general non-Gaussian. In fact, in this case, $S_j(X) = \frac{1}{\sqrt{n_j}}$ is bounded away from 0.

Inferential questions are also extremely simple in this context and essentially again dimension-independent for the reasons highlighted above. Theorem 5.1 naturally reads,

$$\frac{\hat{\beta}_j - \beta_j^*}{\sigma\sqrt{e_j^T (X^T X)^{-1} e_j}} \xrightarrow{d} N(0, 1). \tag{13}$$

Estimating σ is still simple under minimal conditions provided $n - p \to \infty$: see [9, Theorem 1.3] or standard computations concerning the normalized residual sum-of-squares (using variance computations for the latter may require up to 4 moments for ε_i's). Then we can replace σ in (13) by $\hat{\sigma}$ with

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{k=1}^{n} R_k^2$$

where $R_k = y_k - x_k^T \hat{\beta}$ and construct confidence intervals for β_j^* based on $\hat{\sigma}$. If $n - p$ does not tend to ∞, the normalized residual sum of squares is evidently not consistent even in the case of Gaussian errors, so this requirement may not be dispensed of.

6 Numerical results

As seen in the previous sections and related papers, there are five important factors that affect the distribution of $\hat{\beta}$: the design matrix X, the error distribution $\mathscr{L}(\varepsilon)$, the sample size n, the ratio κ, and the loss function ρ. The aim of this section is to assess the quality of the agreement between the asymptotic theoretical results of

Theorem 3.1 and the empirical, finite-dimensional properties of $\hat{\beta}(\rho)$. We also perform a few simulations where some of the assumptions of Theorem 3.1 are violated to get an intuitive sense of whether those assumptions appear necessary or whether they are simply technical artifacts associated with the method of proof we developed. As such, the numerical experiments we report on in this section can be seen as a complement to Theorem 3.1 rather than only a simple check of its practical relevance.

The design matrices we consider are one realization of random design matrices of the following three types:

(i.i.d. design) $X_{ij} \overset{i.i.d.}{\sim} F$;
(elliptical design) $X_{ij} = \zeta_i \tilde{X}_{ij}$, where $\tilde{X}_{ij} \overset{i.i.d.}{\sim} N(0, 1)$ and $\zeta_i \overset{i.i.d.}{\sim} F$. In addition, $\{\zeta_i\}$ is independent of $\{\tilde{X}_{ij}\}$;
(partial Hadamard design) a matrix formed by a random set of p columns of a $n \times n$ Hadamard matrix, i.e. a $n \times n$ matrix whose columns are orthogonal with entries restricted to ± 1.

Here we consider two candidates for F in i.i.d. design and elliptical design: standard normal distribution $N(0, 1)$ and t-distribution with two degrees of freedom (denoted t_2). For the error distribution, we assume that ε has i.i.d. entries with one of the above two distributions, namely $N(0, 1)$ and t_2. The t-distribution violates our assumption **A**2.

To evaluate the finite sample performance, we consider the sample sizes $n \in \{100, 200, 400, 800\}$ and $\kappa \in \{0.5, 0.8\}$. In this section we will consider a Huber loss with $k = 1.345$ [32], i.e.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le k \\ kx - \frac{k^2}{2} & |x| > k \end{cases}$$

$k = 1.345$ is the default in `R` and yields 95% relative efficiency for Gaussian errors in low-dimensional problems. We also carried out the numerical work for L_1-regression, i.e. $\rho(x) = |x|$. See "Appendix D" for details.

6.1 Asymptotic normality of a single coordinate

First we simulate the finite sample distribution of $\hat{\beta}_1$, the first coordinate of $\hat{\beta}$. For each combination of sample size n (100, 200, 400 and 800), type of design (i.i.d, elliptical and Hadamard), entry distribution F (normal and t_2) and error distribution $\mathscr{L}(\varepsilon)$ (normal and t_2), we run 50 simulations with each consisting of the following steps:

(Step 1) Generate one design matrix X;
(Step 2) Generate 300 error vectors ε;
(Step 3) Regress each $Y = \varepsilon$ on the design matrix X and end up with 300 random samples of $\hat{\beta}_1$, denoted by $\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(300)}$;
(Step 4) Estimate the standard deviation of $\hat{\beta}_1$ by the sample standard error $\widehat{\mathrm{sd}}$;

(Step 5) Construct a confidence interval $\mathscr{I}^{(k)} = \left[\hat{\beta}_1^{(k)} - 1.96 \cdot \widehat{\mathrm{sd}}, \hat{\beta}_1^{(k)} + 1.96 \cdot \widehat{\mathrm{sd}}\right]$ for each $k = 1, \ldots, 300$;

(Step 6) Calculate the empirical 95% coverage by the proportion of confidence intervals which cover the true $\beta_1 = 0$.

Finally, we display the boxplots of the empirical 95% coverages of $\hat{\beta}_1$ for each case in Fig. 2. It is worth mentioning that our theories cover two cases: (1) i.i.d design with normal entries and normal errors (orange bars in the first row and the first column), see Proposition 3.1; (2) elliptical design with normal factors ζ_i and normal errors (orange bars in the second row and the first column), see Proposition 3.7.

We first discuss the case $\kappa = 0.5$. In this case, there are only two samples per parameter. Nonetheless, we observe that the coverage is quite close to 0.95, even with a sample size as small as 100, in both cases that are covered by our theories. For other cases, it is interesting to see that the coverage is valid and most stable in the partial hadamard design case and is not sensitive to the distribution of multiplicative factor in elliptical design case even when the error has a t_2 distribution. For i.i.d. designs, the coverage is still valid and stable when the entry is normal. By contrast, when the entry has a t_2 distribution, the coverage has a large variation in small samples. The average coverage is still close to 0.95 in the i.i.d. normal design case but is slightly lower than 0.95 in the i.i.d. t_2 design case. In summary, the finite sample distribution of $\hat{\beta}_1$ is more sensitive to the entry distribution than the error distribution. This indicates that

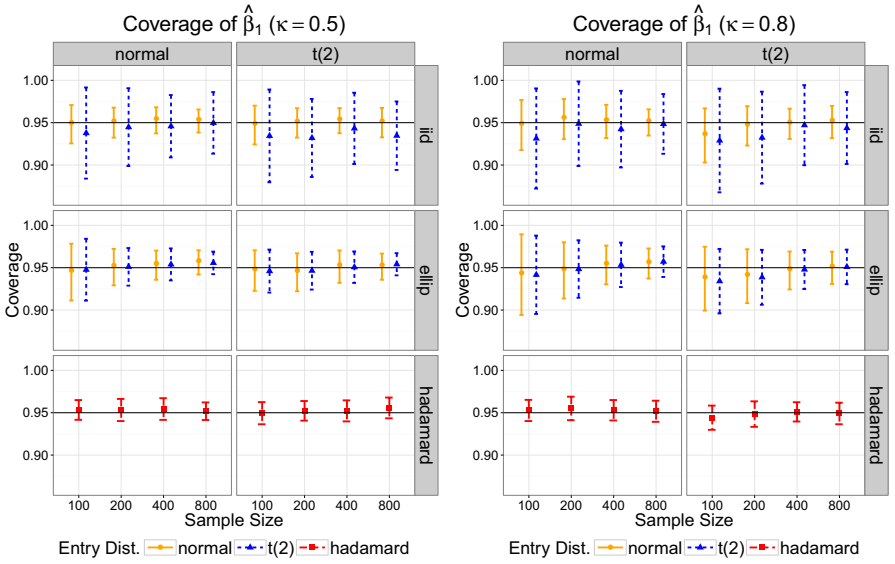

Fig. 2 Empirical 95% coverage of $\hat{\beta}_1$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\mathrm{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \mathrm{Normal}$; the blue dotted bar corresponds to the case $F = t_2$; the red dashed bar represents the Hadamard design (color figure online)

the assumptions on the design matrix are not just artifacts of the proof but are quite essential.

The same conclusion can be drawn from the case where $\kappa = 0.8$ except that the variation becomes larger in most cases when the sample size is small. However, it is worth pointing out that even in this case where there is 1.25 samples per parameter, the sample distribution of $\hat{\beta}_1$ is well approximated by a normal distribution with a moderate sample size ($n \geq 400$). This is in contrast to the classical rule of thumb which suggests that 5–10 samples are needed per parameter.

6.2 Asymptotic normality for multiple marginals

Since our theory holds for general J_n, it is worth checking the approximation for multiple coordinates in finite samples. For illustration, we consider 10 coordinates, namely $\hat{\beta}_1 \sim \hat{\beta}_{10}$, simultaneously and calculate the minimum empirical 95% coverage. To avoid the finite sample dependence between coordinates involved in the simulation, we estimate the empirical coverage independently for each coordinate. Specifically, we run 50 simulations with each consisting of the following steps:

(Step 1) Generate one design matrix X;
(Step 2) Generate 3000 error vectors ε;
(Step 3) Regress each $Y = \varepsilon$ on the design matrix X and end up with 300 random samples of $\hat{\beta}_j$ for each $j = 1, \ldots, 10$ by using the $(300(j-1)+1)$-th to $300j$-th response vector Y;
(Step 4) Estimate the standard deviation of $\hat{\beta}_j$ by the sample standard error $\widehat{\mathrm{sd}}_j$ for $j = 1, \ldots, 10$;
(Step 5) Construct a confidence interval $\mathscr{I}_j^{(k)} = \left[\hat{\beta}_j^{(k)} - 1.96 \cdot \widehat{\mathrm{sd}}_j, \hat{\beta}_j^{(k)} + 1.96 \cdot \widehat{\mathrm{sd}}_j\right]$ for each $j = 1, \ldots, 10$ and $k = 1, \ldots, 300$;
(Step 6) Calculate the empirical 95% coverage by the proportion of confidence intervals which cover the true $\beta_j = 0$, denoted by C_j, for each $j = 1, \ldots, 10$,
(Step 7) Report the minimum coverage $\min_{1 \leq j \leq 10} C_j$.

If the assumptions **A**1–**A**5 are satisfied, $\min_{1 \leq j \leq 10} C_j$ should also be close to 0.95 as a result of Theorem 3.1. Thus, $\min_{1 \leq j \leq 10} C_j$ is a measure for the approximation accuracy for multiple marginals. Figure 3 displays the boxplots of this quantity under the same scenarios as the last subsection. In two cases that our theories cover, the minimum coverage is increasingly closer to the true level 0.95. Similar to the last subsection, the approximation is accurate in the partial hadamard design case and is insensitive to the distribution of multiplicative factors in the elliptical design case. However, the approximation is very inaccurate in the i.i.d. t_2 design case. Again, this shows the evidence that our technical assumptions are not artifacts of the proof.

On the other hand, the Fig. 3 suggests using a conservative variance estimator, e.g. the Jackknife estimator, or corrections on the confidence level in order to make simultaneous inference on multiple coordinates. Here we investigate the validity of Bonferroni correction by modifying the step 5 and step 6. The confidence interval after Bonferroni correction is obtained by

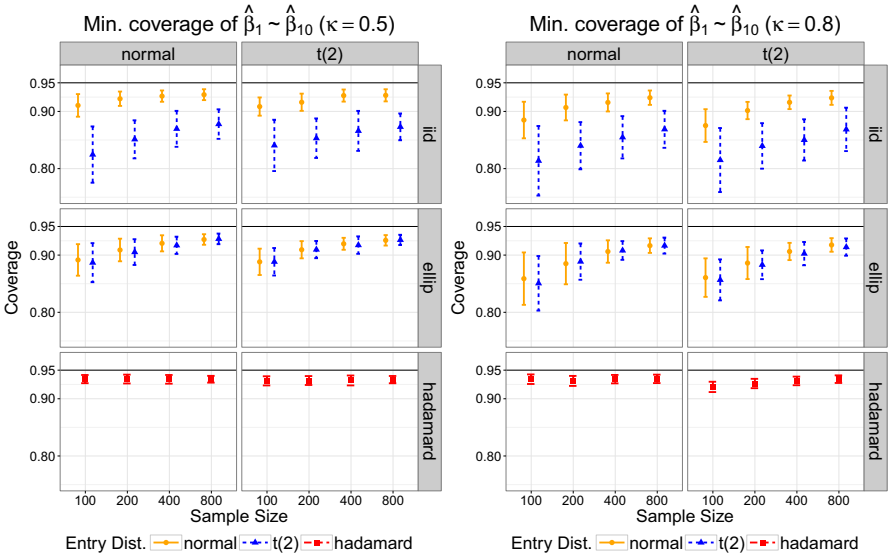

Fig. 3 Mininum empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\text{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the minimum empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design (color figure online)

$$\mathscr{I}_j^{(k)} = \left[\hat{\beta}_j^{(k)} - z_{1-\alpha/20} \cdot \widehat{\text{sd}}_j, \hat{\beta}_j^{(k)} + z_{1-\alpha/20} \cdot \widehat{\text{sd}}_j\right] \tag{14}$$

where $\alpha = 0.05$ and z_γ is the γ-th quantile of a standard normal distribution. The proportion of k such that $0 \in \mathscr{I}_j^{(k)}$ for all $j \le 10$ should be at least 0.95 if the marginals are all close to a normal distribution. We modify the confidence intervals in step 5 by (14) and calculate the proportion of k such that $0 \in \mathscr{I}_j^{(k)}$ for all j in step 6. Figure 4 displays the boxplots of this coverage. It is clear that the Bonferroni correction gives the valid coverage except when $n = 100, \kappa = 0.8$ and the error has a t_2 distribution.

7 Conclusion

We have proved coordinate-wise asymptotic normality for regression M-estimates in the moderate-dimensional asymptotic regime $p/n \to \kappa \in (0, 1)$, for fixed design matrices under appropriate technical assumptions. Our design assumptions are satisfied with high probability for a broad class of random designs. The main novel ingredient of the proof is the use of the second-order Poincaré inequality. Numerical experiments confirm and complement our theoretical results.

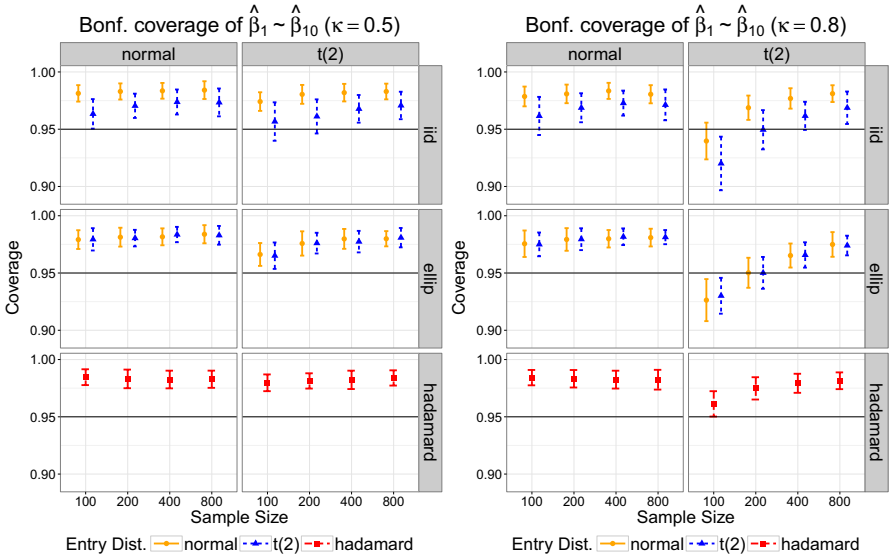

Fig. 4 Empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ after Bonferroni correction with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using $\text{Huber}_{1.345}$ loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical uniform 95% coverage after Bonferroni correction. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design (color figure online)

Appendix

A Proof sketch of Lemma 4.4

In this Appendix, we provide a roadmap for proving Lemma 4.4 by considering a special case where X is one realization of a random matrix Z with i.i.d. mean-zero σ^2-sub-gaussian entries. Random matrix theory [3,26,53] implies that $\lambda_+ = (1 + \sqrt{\kappa})^2 + o_p(1) = O_p(1)$ and $\lambda_- = (1 - \sqrt{\kappa})^2 + o_p(1) = \Omega_p(1)$. Thus, the assumption **A**3 is satisfied with high probability. Thus, the Lemma 4.3 in p. 17 holds with high probability. It remains to prove the following lemma to obtain Theorem 3.1.

Lemma A-1 *Let Z be a random matrix with i.i.d. mean-zero σ^2-sub-gaussian entries and X be one realization of Z. Then under assumptions **A**1 and **A**2,*

$$\max_{1 \le j \le p} M_j = O_p\left(\frac{\text{polyLog(n)}}{n}\right), \quad \min_{1 \le j \le p} \text{Var}(\hat{\beta}_j) = \Omega_p\left(\frac{1}{n \cdot \text{polyLog(n)}}\right),$$

where M_j is defined in (11) in p. 17 and the randomness in $o_p(\cdot)$ and $O_p(\cdot)$ comes from Z.

Note that we prove in Proposition 3.1 that assumptions **A**4 and **A**5 are satisfied with high probability in this case. However, we will not use them directly but prove Lemma

A-1 from the scratch instead, in order to clarify why assump3tions in forms of **A**4 and **A**5 are needed in the proof.

A-1 Upper bound of M_j

First by Proposition E.3,

$$\lambda_+ = O_p(1), \quad \lambda_- = \Omega_p(1).$$

In the rest of the proof, the symbol $\mathbb{E}$ and Var denotes the expectation and the variance conditional on Z. Let $\tilde{Z} = D^{\frac{1}{2}}Z$, then $M_j = \mathbb{E}\|e_j^T(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\|_\infty$. Let $\tilde{H}_j = I - \tilde{Z}_{[j]}(\tilde{Z}_{[j]}^T\tilde{Z}_{[j]})^{-1}\tilde{Z}_{[j]}^T$, then by block matrix inversion formula (see Proposition E.1), which we state as Proposition E.1 in "Appendix E".

$$\begin{aligned}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T &= \begin{pmatrix} \tilde{Z}_1^T\tilde{Z}_1 & \tilde{Z}_1^T\tilde{Z}_{[1]} \\ \tilde{Z}_{[1]}^T\tilde{Z}_1 & \tilde{Z}_{[1]}^T\tilde{Z}_{[1]} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_{[1]} \end{pmatrix} \\ &= \frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} \begin{pmatrix} 1 & -\tilde{Z}_1^T\tilde{Z}_{[1]}(\tilde{Z}_{[1]}^T\tilde{Z}_{[1]})^{-1} \\ * & * \end{pmatrix} \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_{[1]} \end{pmatrix} \\ &= \frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} \begin{pmatrix} \tilde{Z}_1^T(I-\tilde{H}_1) \\ * \end{pmatrix}.\end{aligned}$$

This implies that

$$M_1 = \mathbb{E}\frac{\left\|\tilde{Z}_1^T(I-\tilde{H}_1)\right\|_\infty}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1}. \tag{A-1}$$

Since $Z^TDZ/n \succeq K_0\lambda_- I$, we have

$$\begin{aligned}\frac{1}{\tilde{Z}_1^T(I-\tilde{H}_1)\tilde{Z}_1} &= e_1^T(\tilde{Z}^T\tilde{Z})^{-1}e_1 = e_1^T(Z^TDZ)^{-1}e_1 \\ &= \frac{1}{n}e_1^T\left(\frac{Z^TDZ}{n}\right)^{-1}e_1 \le \frac{1}{nK_0\lambda_-}\end{aligned}$$

and we obtain a bound for M_1 as

$$M_1 \le \frac{\mathbb{E}\left\|\tilde{Z}_1^T(I-\tilde{H}_1)\right\|_\infty}{nK_0\lambda_-} = \frac{\mathbb{E}\left\|Z_1^TD^{\frac{1}{2}}(I-\tilde{H}_1)\right\|_\infty}{nK_0\lambda_-}.$$

Similarly,

$$M_j \le \frac{\mathbb{E}\left\|Z_j^TD^{\frac{1}{2}}(I-\tilde{H}_j)\right\|_\infty}{nK_0\lambda_-}$$

$$= \frac{\mathbb{E}\left\| Z_j^T D^{\frac{1}{2}} \left(I - D^{\frac{1}{2}} Z_{[j]}^T \left(Z_{[j]}^T D Z_{[j]} \right)^{-1} Z_{[j]} D^{\frac{1}{2}} \right) \right\|_\infty}{n K_0 \lambda_-}. \tag{A-2}$$

The vector in the numerator is a linear contrast of Z_j and Z_j has mean-zero i.i.d. sub-gaussian entries. For any fixed matrix $A \in \mathbb{R}^{n\times n}$, denote A_k by its k-th column, then $A_k^T Z_j$ is $\sigma^2 \|A_k\|_2^2$-sub-gaussian (see Section 5.2.3 of [57] for a detailed discussion) and hence by definition of sub-Gaussianity,

$$P\left(\left| A_k^T Z_j \right| \geq \sigma \|A_k\|_2 t \right) \leq 2e^{-\frac{t^2}{2}}.$$

Therefore, by a simple union bound, we conclude that

$$P(\|A^T Z_j\|_\infty \geq \sigma \max_k \|A_k\|_2 t) \leq 2ne^{-\frac{t^2}{2}}.$$

Let $t = 2\sqrt{\log n}$,

$$P(\|A^T Z_j\|_\infty \geq 2\sigma \max_k \|A_k\|_2 \sqrt{\log n}) \leq \frac{2}{n} = o(1).$$

This entails that

$$\|A^T Z_j\|_\infty = O_p\left(\max_k \|A_k\|_2 \cdot \text{polyLog(n)} \right) = O_p\left(\|A\|_{\text{op}} \cdot \text{polyLog(n)} \right). \tag{A-3}$$

with high probability. In M_j, the coefficient matrix $(I - H_j) D^{\frac{1}{2}}$ depends on Z_j through D and hence we cannot use (A-3) directly. However, the dependence can be removed by replacing D by $D_{[j]}$ since $r_{i,[j]}$ does not depend on Z_j.

Since Z has i.i.d. sub-gaussian entries, no column is highly influential. In other words, the estimator will not change drastically after removing j-th column. This would suggest $R_i \approx r_{i,[j]}$. It is proved by [20] that

$$\sup_{i,j} |R_i - r_{i,[j]}| = O_p\left(\frac{\text{polyLog(n)}}{\sqrt{n}} \right).$$

It can be rigorously proved that

$$\left| \|Z_j^T D(I - \tilde{H}_j)\|_\infty - \|Z_j^T D_{[j]}(I - H_j)\|_\infty \right| = O_p\left(\frac{\text{polyLog(n)}}{n} \right),$$

where $H_j = I - D_{[j]}^{\frac{1}{2}} Z_{[j]} (Z_{[j]}^T D_{[j]} Z_{[j]})^{-1} Z_{[j]}^T D_{[j]}^{\frac{1}{2}}$; see "Appendix A-1" for details. Since $D_{[j]}(I - H_j)$ is independent of Z_j and

$$\|D_{[j]}(I - H_j)\|_{\text{op}} \leq \|D_{[j]}\|_{\text{op}} \leq K_1 = O\left(\text{polyLog(n)}\right),$$

it follows from (A-2) and (A-3) that

$$\left\| Z_j^T D_{[j]}(I - H_j) \right\|_\infty = O_p\left(\frac{\text{polyLog(n)}}{n}\right).$$

In summary,

$$M_j = O_p\left(\frac{\text{polyLog(n)}}{n}\right). \tag{A-4}$$

A-2 Lower bound of $\text{Var}(\hat{\beta}_j)$

A-2.1 Approximating $\text{Var}(\hat{\beta}_j)$ *by* $\text{Var}(b_j)$

It is shown by [20][1] that

$$\hat{\beta}_j \approx b_j \triangleq \frac{1}{\sqrt{n}} \frac{N_j}{\xi_j} \tag{A-5}$$

where

$$N_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_{ij} \psi(r_{i,[j]}),$$
$$\xi_j = \frac{1}{n} Z_j^T \left(D_{[j]} - D_{[j]} Z_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} Z_{[j]}^T D_{[j]} \right) Z_j.$$

It has been shown by [20] that

$$\max_j |\hat{\beta}_j - b_j| = O_p\left(\frac{\text{polyLog(n)}}{n}\right).$$

Thus, $\text{Var}(\hat{\beta}_j) \approx \text{Var}(b_j)$ and a more refined calculation in "Appendix A-2.1" shows that

$$|\text{Var}(\hat{\beta}_j) - \text{Var}(b_j)| = O_p\left(\frac{\text{polyLog(n)}}{n^{\frac{3}{2}}}\right).$$

It is left to show that

$$\text{Var}(b_j) = \Omega_p\left(\frac{1}{n \cdot \text{polyLog(n)}}\right). \tag{A-6}$$

[1] [20] considers a ridge regularized M estimator, which is different from our setting. However, this argument still holds in our case and proved in "Appendix B".

A-2.2 Bounding $\mathrm{Var}(b_j)$ *via* $\mathrm{Var}(N_j)$

By definition of b_j,

$$\mathrm{Var}(b_j) = \Omega_p\left(\frac{\mathrm{polyLog(n)}}{n}\right) \Longleftrightarrow \mathrm{Var}\left(\frac{N_j}{\xi_j}\right) = \Omega_p\left(\mathrm{polyLog(n)}\right).$$

As will be shown in "Appendix B-6.4",

$$\mathrm{Var}(\xi_j) = O_p\left(\frac{\mathrm{polyLog(n)}}{n}\right).$$

As a result, $\xi_j \approx \mathbb{E}\xi_j$ and

$$\mathrm{Var}\left(\frac{N_j}{\xi_j}\right) \approx \mathrm{Var}\left(\frac{N_j}{\mathbb{E}\xi_j}\right) = \frac{\mathrm{Var}(N_j)}{(\mathbb{E}\xi_j)^2}.$$

As in the previous paper [20], we rewrite ξ_j as

$$\xi_j = \frac{1}{n} Z_j^T D_{[j]}^{\frac{1}{2}} \left(I - D_{[j]}^{\frac{1}{2}} Z_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} Z_{[j]}^T D_{[j]}^{\frac{1}{2}}\right) D_{[j]}^{\frac{1}{2}} Z_j.$$

The middle matrix is idempotent and hence positive semi-definite. Thus,

$$\xi_j \leq \frac{1}{n} Z_j^T D_{[j]} Z_j \leq K_1 \lambda_+ = O_p\left(\mathrm{polyLog(n)}\right).$$

Then we obtain that

$$\frac{\mathrm{Var}(N_j)}{(\mathbb{E}\xi_j)^2} = \Omega_p\left(\frac{\mathrm{Var}(N_j)}{\mathrm{polyLog(n)}}\right),$$

and it is left to show that

$$\mathrm{Var}(N_j) = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right). \tag{A-7}$$

A-2.3 Bounding $\mathrm{Var}(N_j)$ *via* $\mathrm{tr}(Q_j)$

Recall the definition of N_j (A-5), and that of Q_j (see Sect. 3.1 in p. 8), we have

$$\mathrm{Var}(N_j) = \frac{1}{n} Z_j^T Q_j Z_j$$

Notice that Z_j is independent of $r_{i,[j]}$ and hence the conditional distribution of Z_j given Q_j remains the same as the marginal distribution of Z_j. Since Z_j has i.i.d. sub-gaussian entries, the Hanson-Wright inequality ([27,51]; see Proposition E.2), shown

in Proposition E.2, implies that any quadratic form of Z_j, denoted by $Z_j^T Q_j Z_j$ is concentrated on its mean, i.e.

$$Z_j^T Q_j Z_j \approx \mathbb{E}_{Z_j,\varepsilon} Z_j^T Q_j Z_j = \left(\mathbb{E} Z_{1j}^2\right) \cdot \mathrm{tr}(Q_j).$$

As a consequence, it is left to show that

$$\mathrm{tr}(Q_j) = \Omega_p\left(\frac{n}{\mathrm{polyLog(n)}}\right). \tag{A-8}$$

A-2.4 Lower bound of $\mathrm{tr}(Q_j)$

By definition of Q_j,

$$\mathrm{tr}(Q_j) = \sum_{i=1}^{n} \mathrm{Var}(\psi(r_{i,[j]})).$$

To lower bounded the variance of $\psi(r_{i,[j]})$, recall that for any random variable W,

$$\mathrm{Var}(W) = \frac{1}{2}\mathbb{E}(W - W')^2. \tag{A-9}$$

where W' is an independent copy of W. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function such that $|g'(x)| \geq c$ for all x, then (A-9) implies that

$$\mathrm{Var}(g(W)) = \frac{1}{2}\mathbb{E}(g(W) - g(W'))^2 \geq \frac{c^2}{2}\mathbb{E}(W - W')^2 = c^2\,\mathrm{Var}(W). \tag{A-10}$$

In other words, (A-10) entails that $\mathrm{Var}(W)$ is a lower bound for $\mathrm{Var}(g(W))$ provided that the derivative of g is bounded away from 0. As an application, we see that

$$\mathrm{Var}(\psi(r_{i,[j]})) \geq K_0^2\,\mathrm{Var}(r_{i,[j]})$$

and hence

$$\mathrm{tr}(Q_j) \geq K_0^2 \sum_{i=1}^{n} \mathrm{Var}(r_{i,[j]}).$$

By the variance decomposition formula,

$$\mathrm{Var}(r_{i,[j]}) = \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\varepsilon_{(i)}\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(r_{i,[j]}\middle|\varepsilon_{(i)}\right)\right) \geq \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\varepsilon_{[i]}\right)\right),$$

where $\varepsilon_{(i)}$ includes all but i-th entry of ε. Given $\varepsilon_{(i)}$, $r_{i,[j]}$ is a function of ε_i. Using (A-10), we have

$$\mathrm{Var}(r_{i,[j]}|\varepsilon_{(i)}) \geq \inf_{\varepsilon_i} \left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2 \cdot \mathrm{Var}(\varepsilon_i|\varepsilon_{(i)}) \geq \inf_{\varepsilon_i} \left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2 \cdot \mathrm{Var}(\varepsilon_i).$$

This implies that

$$\mathrm{Var}(r_{i,[j]}) \geq \mathbb{E}\left(\mathrm{Var}\left(r_{i,[j]}\middle|\varepsilon_{[i]}\right)\right) \geq \mathbb{E}\inf_{\varepsilon} \left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2 \cdot \min_i \mathrm{Var}(\varepsilon_i).$$

Summing $\mathrm{Var}(r_{i,[j]})$ over $i = 1, \ldots, n$, we obtain that

$$\mathrm{tr}(Q_j) = \sum_{i=1}^{n} \mathrm{Var}(r_{i,[j]}) \geq \mathbb{E}\left(\sum_i \inf_{\varepsilon} \left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2\right) \cdot \min_i \mathrm{Var}(\varepsilon_i).$$

It will be shown in "Appendix B-6.3" that under assumptions **A**1–**A**3,

$$\mathbb{E}\sum_i \inf_{\varepsilon} \left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2 = \Omega_p\left(\frac{n}{\mathrm{polyLog(n)}}\right). \tag{A-11}$$

This proves (A-8) and as a result,

$$\min_j \mathrm{Var}(\hat{\beta}_j) = \Omega_p\left(\frac{1}{n \cdot \mathrm{polyLog(n)}}\right).$$

B Proof of Theorem 3.1

B-1 Notation

To be self-contained, we summarize our notations in this subsection. The model we considered here is

$$y = X\beta^* + \varepsilon$$

where $X \in \mathbb{R}^{n\times p}$ be the design matrix and ε is a random vector with independent entries. Notice that the target quantity $\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}$ is shift invariant, we can assume $\beta^* = 0$ without loss of generality provided that X has full column rank; see Sect. 3.1 for details.

Let $x_i^T \in \mathbb{R}^{1\times p}$ denote the i-th row of X and $X_j \in \mathbb{R}^{n\times 1}$ denote the j-th column of X. Throughout the paper we will denote by $X_{ij} \in \mathbb{R}$ the (i, j)-th entry of X, by $X_{(i)} \in \mathbb{R}^{(n-1)\times p}$ the design matrix X after removing the i-th row, by $X_{[j]} \in \mathbb{R}^{n\times(p-1)}$ the design matrix X after removing the j-th column, by $X_{(i),[j]} \in \mathbb{R}^{(n-1)\times(p-1)}$ the design matrix after removing both i-th row and j-th column, and by $x_{i,[j]} \in \mathbb{R}^{1\times(p-1)}$

the vector x_i after removing j-th entry. The M-estimator $\hat{\beta}$ associated with the loss function ρ is defined as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{k=1}^{n} \rho\left(\varepsilon_k - x_k^T \beta\right). \tag{B-12}$$

Similarly we define the leave-j-th-predictor-out version as

$$\hat{\beta}_{[j]} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{k=1}^{n} \rho\left(\varepsilon_k - x_{k,[j]}^T \beta\right). \tag{B-13}$$

Based on these notation we define the full residual R_k as

$$R_k = \varepsilon_k - x_k^T \hat{\beta}, \quad k = 1, 2, \ldots, n \tag{B-14}$$

the leave-j-th-predictor-out residual as

$$r_{k,[j]} = \varepsilon_k - x_{k,[j]}^T \hat{\beta}_{[j]}, \quad k = 1, 2, \ldots, n, \ \ j \in J_n. \tag{B-15}$$

Four diagonal matrices are defined as

$$D = \operatorname{diag}(\psi'(R_k)), \quad \tilde{D} = \operatorname{diag}(\psi''(R_k)), \tag{B-16}$$

$$D_{[j]} = \operatorname{diag}(\psi'(r_{k,[j]})), \quad \tilde{D}_{[j]} = \operatorname{diag}(\psi''(r_{k,[j]})). \tag{B-17}$$

Further we define G and $G_{[j]}$ as

$$G = I - X(X^T D X)^{-1} X^T D, \quad G_{[j]} = I - X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}. \tag{B-18}$$

Let J_n denote the indices of coefficients of interest. We say $a \in]a_1, a_2[$ if and only if $a \in [\min\{a_1, a_2\}, \max\{a_1, a_2\}]$. Regarding the technical assumptions, we need the following quantities

$$\lambda_+ = \lambda_{\max}\left(\frac{X^T X}{n}\right), \quad \lambda_- = \lambda_{\min}\left(\frac{X^T X}{n}\right) \tag{B-19}$$

be the largest (resp. smallest) eigenvalue of the matrix $\frac{X^T X}{n}$. Let $e_i \in \mathbb{R}^n$ be the i-th canonical basis vector and

$$h_{j,0} = (\psi(r_{1,[j]}), \ldots, \psi(r_{n,[j]}))^T, \quad h_{j,1,i} = G_{[j]}^T e_i. \tag{B-20}$$

Finally, let

$$\Delta_C = \max\left\{ \max_{j \in J_n} \frac{\left|h_{j,0}^T X_j\right|}{\|h_{j,0}\|}, \max_{i \le n, j \in J_n} \frac{\left|h_{j,1,i}^T X_j\right|}{\|h_{j,1,i}\|} \right\}, \tag{B-21}$$

$$Q_j = \mathrm{Cov}(h_{j,0}). \tag{B-22}$$

We adopt Landau's notation ($O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, $o_p(\cdot)$). In addition, we say $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ and similarly, we say $a_n = \Omega_p(b_n)$ if $b_n = O_p(a_n)$. To simplify the logarithm factors, we use the symbol polyLog(n) to denote any factor that can be upper bounded by $(\log n)^\gamma$ for some $\gamma > 0$. Similarly, we use $\frac{1}{\mathrm{polyLog(n)}}$ to denote any factor that can be lower bounded by $\frac{1}{(\log n)^{\gamma'}}$ for some $\gamma' > 0$.

Finally we restate all the technical assumptions:

A1 $\rho(0) = \psi(0) = 0$ and there exists $K_0 = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$, $K_1, K_2 = O\,(\mathrm{polyLog(n)})$, such that for any $x \in \mathbb{R}$,

$$K_0 \le \psi'(x) \le K_1, \quad \left|\frac{d}{dx}(\sqrt{\psi'}(x))\right| = \frac{|\psi''(x)|}{\sqrt{\psi'(x)}} \le K_2;$$

A2 $\varepsilon_i = u_i(W_i)$ where $(W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and u_i are smooth functions with $\|u_i'\|_\infty \le c_1$ and $\|u_i''\|_\infty \le c_2$ for some $c_1, c_2 = O(\mathrm{polyLog(n)})$. Moreover, assume $\min_i \mathrm{Var}(\varepsilon_i) = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$.

A3 $\lambda_+ = O(\mathrm{polyLog(n)})$ and $\lambda_- = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$;

A4 $\min_{j\in J_n} \frac{X_j^T Q_j X_j}{\mathrm{tr}(Q_j)} = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$;

A5 $\mathbb{E}\Delta_C^8 = O\,(\mathrm{polyLog(n)})$.

B-2 Deterministic approximation results

In "Appendix A", we use several approximations under random designs, e.g. $R_i \approx r_{i,[j]}$. To prove them, we follow the strategy of [20] which establishes the deterministic results and then apply the concentration inequalities to obtain high probability bounds. Note that $\hat\beta$ is the solution of

$$0 = f(\beta) \triangleq \frac{1}{n}\sum_{i=1}^n x_i \psi\left(\varepsilon_i - x_i^T\beta\right),$$

we need the following key lemma to bound $\|\beta_1 - \beta_2\|_2$ by $\|f(\beta_1) - f(\beta_2)\|_2$, which can be calculated explicily.

Lemma B-1 [20, Proposition 2.1] *For any β_1 and β_2,*

$$\|\beta_1 - \beta_2\|_2 \le \frac{1}{K_0\lambda_-}\|f(\beta_1) - f(\beta_2)\|_2.$$

Proof By the mean value theorem, there exists $\nu_i \in]\varepsilon_i - x_i^T\beta_1, \varepsilon_i - x_i^T\beta_2[$ such that

$$\psi\left(\varepsilon_i - x_i^T\beta_1\right) - \psi\left(\varepsilon_i - x_i^T\beta_2\right) = \psi'(\nu_i)\cdot x_i^T(\beta_2 - \beta_1).$$

Then

$$\begin{aligned}\|f(\beta_1) - f(\beta_2)\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^{n} \psi'(v_i) x_i x_i^T \, (\beta_1 - \beta_2) \right\|_2 \\ &\geq \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \psi'(v_i) x_i x_i^T \right) \cdot \|\beta_1 - \beta_2\|_2 \\ &\geq K_0 \lambda_- \, \|\beta_1 - \beta_2\|_2 \, .\end{aligned}$$

□

Based on Lemma B-1, we can derive the deterministic results informally stated in "Appendix A". Such results are shown by [20] for ridge-penalized M-estimates and here we derive a refined version for unpenalized M-estimates. Throughout this subsection, we only assume assumption **A**1. This implies the following lemma,

Lemma B-2 *Under assumption A1, for any x and y,*

$$\begin{aligned}&|\psi(x)| \leq K_1|x|, \quad |\sqrt{\psi'}(x) - \sqrt{\psi'}(y)| \leq K_2|x-y|, \\ &|\psi'(x) - \psi'(y)| \leq 2\sqrt{K_1}K_2|x-y| \triangleq K_3|x-y|.\end{aligned}$$

To state the result, we define the following quantities.

$$T = \frac{1}{\sqrt{n}} \max\left\{ \max_i \|x_i\|_2, \max_{j \in J_n} \|X_j\|_2 \right\}, \quad \mathscr{E} = \frac{1}{n}\sum_{i=1}^{n} \rho(\varepsilon_i), \tag{B-23}$$

$$U = \left\| \frac{1}{n}\sum_{i=1}^{n} x_i(\psi(\varepsilon_i) - \mathbb{E}\psi(\varepsilon_i)) \right\|_2, \quad U_0 = \left\| \frac{1}{n}\sum_{i=1}^{n} x_i \mathbb{E}\psi(\varepsilon_i) \right\|_2. \tag{B-24}$$

The following proposition summarizes all deterministic results which we need in the proof.

Proposition B.1 *Under Assumption A*1,

(i) *The norm of M estimator is bounded by*

$$\|\hat{\beta}\|_2 \leq \frac{1}{K_0\lambda_-}(U + U_0);$$

(ii) *Define* b_j *as*

$$b_j = \frac{1}{\sqrt{n}} \frac{N_j}{\xi_j}$$

where

$$N_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij} \psi(r_{i,[j]}),$$

$$\xi_j = \frac{1}{n} X_j^T \left(D_{[j]} - D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]} \right) X_j,$$

Then

$$\max_{j \in J_n} |b_j| \le \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2K_1}}{K_0 \lambda_-} \cdot \Delta_C \cdot \sqrt{\mathscr{E}},$$

(iii) *The difference between* $\hat{\beta}_j$ *and* b_j *is bounded by*

$$\max_{j \in J_n} |\hat{\beta}_j - b_j| \le \frac{1}{n} \cdot \frac{2K_1^2 K_3 \lambda_+ T}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E}.$$

(iv) *The difference between the full and the leave-one-predictor-out residual is bounded by*

$$\max_{j \in J_n} \max_i |R_i - r_{i,[j]}| \le \frac{1}{\sqrt{n}} \left(\frac{2K_1^2 K_3 \lambda_+ T^2}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E} + \frac{\sqrt{2} K_1}{K_0^{\frac{3}{2}} \lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathscr{E}} \right).$$

Proof (i) By Lemma B-1,

$$\|\hat{\beta}\|_2 \le \frac{1}{K_0 \lambda_-} \|f(\hat{\beta}) - f(0)\|_2 = \frac{\|f(0)\|_2}{K_0 \lambda_-},$$

since $\hat{\beta}$ is a zero of $f(\beta)$. By definition,

$$f(0) = \frac{1}{n} \sum_{i=1}^n x_i \psi(\varepsilon_i) = \frac{1}{n} \sum_{i=1}^n x_i (\psi(\varepsilon_i) - \mathbb{E}\psi(\varepsilon_i)) + \frac{1}{n} \sum_{i=1}^n x_i \mathbb{E}\psi(\varepsilon_i).$$

This implies that

$$\|f(0)\|_2 \le U + U_0.$$

(ii) First we prove that

$$\xi_j \ge K_0 \lambda_-. \tag{B-25}$$

Since all diagonal entries of $D_{[j]}$ is lower bounded by K_0, we conclude that

$$\lambda_{\min} \left(\frac{X^T D_{[j]} X}{n} \right) \ge K_0 \lambda_-.$$

Note that ξ_j is the Schur's complement ([28], chapter 0.8) of $\frac{X^T D_{[j]} X}{n}$, we have

$$\xi_j^{-1} = e_j^T \left(\frac{X^T D_{[j]} X}{n} \right)^{-1} e_j \leq \frac{1}{K_0 \lambda_-},$$

which implies (B-25). As for N_j, we have

$$N_j = \frac{X_j^T h_{j,0}}{\sqrt{n}} = \frac{\|h_{j,0}\|_2}{\sqrt{n}} \cdot \frac{X_j^T h_{j,0}}{\|h_{j,0}\|_2}. \tag{B-26}$$

The the second term is bounded by Δ_C by definition, see (B-21). For the first term, the assumption **A**1 that $\psi'(x) \leq K_1$ implies that

$$\rho(x) = \rho(x) - \rho(0) = \int_0^x \psi(y) dy \geq \int_0^x \frac{\psi'(y)}{K_1} \cdot \psi(y) dy = \frac{1}{2K_1} \psi^2(x).$$

Here we use the fact that $\text{sign}(\psi(y)) = \text{sign}(y)$. Recall the definition of $h_{j,0}$, we obtain that

$$\frac{\|h_{j,0}\|_2}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^n \psi(r_{i,[j]})^2}{n}} \leq \sqrt{2K_1} \cdot \sqrt{\frac{\sum_{i=1}^n \rho(r_{i,[j]})}{n}}.$$

Since $\hat{\beta}_{[j]}$ is the minimizer of the loss function $\sum_{i=1}^n \rho(\varepsilon_i - x_{i,[j]}^T \beta_{[j]})$, it holds that

$$\frac{1}{n} \sum_{i=1}^n \rho(r_{i,[j]}) \leq \frac{1}{n} \sum_{i=1}^n \rho(\varepsilon_i) = \mathscr{E}.$$

Putting together the pieces, we conclude that

$$|N_j| \leq \sqrt{2K_1} \cdot \Delta_C \sqrt{\mathscr{E}}. \tag{B-27}$$

By definition of b_j,

$$|b_j| \leq \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2K_1}}{K_0 \lambda_-} \Delta_C \sqrt{\mathscr{E}}.$$

(iii) The proof of this result is almost the same as [20]. We state it here for the sake of completeness. Let $\tilde{\mathbf{b}}_\mathbf{j} \in \mathbb{R}^p$ with

$$(\tilde{\mathbf{b}}_\mathbf{j})_j = b_j, \quad (\tilde{\mathbf{b}}_\mathbf{j})_{[j]} = \hat{\beta}_{[j]} - b_j \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]} X_j \tag{B-28}$$

where the subscript j denotes the j-th entry and the subscript $[j]$ denotes the sub-vector formed by all but j-th entry. Furthermore, define γ_j with

$$(\gamma_j)_j = -1, \quad (\gamma_j)_{[j]} = \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]} X_j. \tag{B-29}$$

Then we can rewrite $\tilde{\mathbf{b}}_\mathbf{j}$ as

$$(\tilde{\mathbf{b}}_\mathbf{j})_j = -b_j(\gamma_j)_j, \quad (\tilde{\mathbf{b}}_\mathbf{j})_{[j]} = \hat{\beta}_{[j]} - b_j(\gamma_j)_{[j]}.$$

By definition of $\hat{\beta}_{[j]}$, we have $[f(\hat{\beta}_{[j]})]_{[j]} = 0$ and hence

$$\begin{aligned}[f(\tilde{\mathbf{b}}_\mathbf{j})]_{[j]} &= [f(\tilde{\mathbf{b}}_\mathbf{j})]_{[j]} - [f(\hat{\beta}_{[j]})]_{[j]} \\ &= \frac{1}{n}\sum_{i=1}^n x_{i,[j]}\left[\psi(\varepsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}) - \psi(\varepsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]})\right].\end{aligned} \tag{B-30}$$

By mean value theorem, there exists $v_{i,j} \in]\varepsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}, \varepsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]}[$ such that

$$\begin{aligned}&\psi\left(\varepsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}\right) - \psi\left(\varepsilon_i - x_{i,[j]}^T\hat{\beta}_{[j]}\right) = \psi'(v_{i,j})\left(x_{i,[j]}^T\hat{\beta}_{[j]} - x_i^T\tilde{\mathbf{b}}_\mathbf{j}\right) \\ &= \psi'(v_{i,j})\left(x_{i,[j]}^T\hat{\beta}_{[j]} - x_{i,[j]}^T(\tilde{\mathbf{b}}_\mathbf{j})_{[j]} - X_{ij}b_j\right) \\ &= \psi'(v_{i,j})\cdot b_j\cdot\left[x_{i,[j]}^T\left(X_{[j]}^T D_{[j]}X_{[j]}\right)^{-1}X_{[j]}^T D_{[j]}X_j - X_{ij}\right]\end{aligned}$$

Let

$$d_{i,j} = \psi'(v_{i,j}) - \psi'(r_{i,[j]}) \tag{B-31}$$

and plug the above result into (B-30), we obtain that

$$\begin{aligned}\left[f(\tilde{\mathbf{b}}_\mathbf{j})\right]_{[j]} &= \frac{1}{n}\sum_{i=1}^n x_{i,[j]}\cdot\left(\psi'(r_{i,[j]}) + d_{i,j}\right)\cdot b_j\cdot\left[x_{i,[j]}^T\left(X_{[j]}^T D_{[j]}X_{[j]}\right)^{-1}X_{[j]}^T D_{[j]}X_j - X_{ij}\right] \\ &= b_j\cdot\frac{1}{n}\sum_{i=1}^n\psi'(r_{i,[j]})x_{i,[j]}\left[x_{i,[j]}^T\left(X_{[j]}^T D_{[j]}X_{[j]}\right)^{-1}X_{[j]}^T D_{[j]}X_j - X_{ij}\right] \\ &\quad + b_j\cdot\frac{1}{n}\sum_{i=1}^n d_{i,j}x_{i,[j]}\left(x_{i,[j]}^T\left(X_{[j]}^T D_{[j]}X_{[j]}\right)^{-1}X_{[j]}^T D_{[j]}X_j - X_{ij}\right) \\ &= b_j\cdot\frac{1}{n}\left[X_{[j]}^T D_{[j]}X_{[j]}\left(X_{[j]}^T D_{[j]}X_{[j]}\right)^{-1}X_{[j]}^T D_{[j]}X_j - X_{[j]}^T D_{[j]}X_j\right] \\ &\quad + b_j\cdot\frac{1}{n}\sum_{i=1}^n d_{i,j}x_{i,[j]}\cdot x_i^T\gamma_j \\ &= b_j\cdot\frac{1}{n}\left(\sum_{i=1}^n d_{i,j}x_{i,[j]}x_i^T\right)\gamma_j.\end{aligned}$$

Now we calculate $[f(\tilde{\mathbf{b}}_\mathbf{j})]_j$, the j-th entry of $f(\tilde{\mathbf{b}}_\mathbf{j})$. Note that

$$\begin{aligned}
\left[f(\tilde{\mathbf{b}}_\mathbf{j})\right]_j &= \frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi\left(\varepsilon_i - x_i^T\tilde{\mathbf{b}}_\mathbf{j}\right) = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}) \\
&\quad + b_j \cdot \frac{1}{n}\sum_{i=1}^{n} X_{ij}(\psi'(r_{i,[j]}) + d_{i,j}) \\
&\quad \cdot \left[x_{i,[j]}^T (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi(r_{i,[j]}) + b_j \\
&\quad \cdot \frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[j]}) X_{ij}\left[x_{i,[j]}^T (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right] \\
&\quad + b_j \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T\right)\gamma_j = \frac{1}{\sqrt{n}} N_j + b_j \\
&\quad \cdot \left(\frac{1}{n} X_j^T D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - \frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[j]}) X_{ij}^2\right) \\
&\quad + b_j \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T\right)\gamma_j = \frac{1}{\sqrt{n}} N_j - b_j \cdot \xi_j \\
&\quad + b_j \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T\right)\gamma_j = b_j \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_{i,j} X_{ij} x_i^T\right)\gamma_j
\end{aligned}$$

where the second last line uses the definition of b_j. Putting the results together, we obtain that

$$f(\tilde{\mathbf{b}}_\mathbf{j}) = b_j \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_{i,j} x_i x_i^T\right) \cdot \gamma_j.$$

This entails that

$$\|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2 \le |b_j| \cdot \max_i |d_{i,j}| \cdot \lambda_+ \cdot \|\gamma_j\|_2. \tag{B-32}$$

Now we derive a bound for $\max_i |d_{i,j}|$, where $d_{i,j}$ is defined in (B-36). By Lemma B-2,

$$|d_{i,j}| = |\psi'(v_{i,j}) - \psi'(r_{i,[j]})| \le K_3 \left|v_{i,j} - r_{i,[j]}\right| = K_3 |x_{i,[j]}^T \hat{\beta}_{[j]} - x_i^T \tilde{\mathbf{b}}_\mathbf{j}\Big|.$$

By definition of $\tilde{\mathbf{b}}_\mathbf{j}$ and $h_{j,1,i}$,

$$|x_{i,[j]}^T \hat{\beta}_{[j]} - x_i^T \tilde{\mathbf{b}}_\mathbf{j}| = |b_j| \cdot \left|x_{i,[j]}^T (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]} X_j - X_{ij}\right|$$

$$= |b_j| \cdot \left| e_i^T (I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}) X_j \right|$$
$$= |b_j| \cdot \left| h_{j,1,i}^T X_j \right| \le |b_j| \cdot \Delta_C \left\| h_{j,1,i} \right\|_2, \tag{B-33}$$

where the last inequality is derived by definition of Δ_C, see (B-21). Since $h_{j,1,i}$ is the i-th column of matrix $I - D_{[j]}X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T$, its L_2 norm is upper bounded by the operator norm of this matrix. Notice that

$$I - D_{[j]}X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T$$
$$= D_{[j]}^{\frac{1}{2}} \left(I - D_{[j]}^{\frac{1}{2}} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} \right) D_{[j]}^{-\frac{1}{2}}.$$

The middle matrix in RHS of the displayed atom is an orthogonal projection matrix and hence

$$\left\| I - D_{[j]}X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T \right\|_{\text{op}} \le \left\| D_{[j]}^{\frac{1}{2}} \right\|_{\text{op}} \cdot \left\| D_{[j]}^{-\frac{1}{2}} \right\|_{\text{op}} \le \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}. \tag{B-34}$$

Therefore,

$$\max_{i,j} \|h_{j,1,i}\|_2 \le \max_{j \in J_n} \left\| I - D_{[j]}X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T \right\|_{\text{op}} \le \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}, \tag{B-35}$$

and thus

$$\max_i |d_{i,j}| \le K_3 \sqrt{\frac{K_1}{K_0}} \cdot |b_j| \cdot \Delta_C. \tag{B-36}$$

As for γ_j, we have

$$K_0 \lambda_- \|\gamma_j\|_2^2 \le \gamma_j^T \left(\frac{X^T D_{[j]} X}{n} \right) \gamma_j$$
$$= (\gamma_j)_j^2 \cdot \frac{X_j^T D_j X_j}{n} + (\gamma_j)_{[j]}^T \left(\frac{X_{[j]}^T D_{[j]} X_{[j]}}{n} \right) (\gamma_j)_{[j]}$$
$$+ 2\gamma_j \frac{X_j^T D_{[j]} X_{[j]}}{n} (\gamma_j)_{[j]}$$

Recall the definition of γ_j in (B-37), we have

$$(\gamma_j)_{[j]}^T \left(\frac{X_{[j]}^T D_{[j]} X_{[j]}}{n} \right) (\gamma_j)_{[j]} = \frac{1}{n} X_j^T D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]} X_j$$

and

$$\gamma_j \frac{X_j^T D_{[j]} X_{[j]}}{n}(\gamma_j)_{[j]} = -\frac{1}{n} X_j^T D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]} X_j.$$

As a result,

$$\begin{aligned} K_0\lambda_- \|\gamma_j\|_2^2 &\le \frac{1}{n} X_j^T D_{[j]}^{\frac{1}{2}} \left(I - D_{[j]}^{\frac{1}{2}} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}\right) D_{[j]}^{\frac{1}{2}} X_j \\ &\le \frac{\left\| D_{[j]}^{\frac{1}{2}} X_j \right\|_2^2}{n} \cdot \left\| I - D_{[j]}^{\frac{1}{2}} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} \right\|_{op} \\ &\le \frac{\left\| D_{[j]}^{\frac{1}{2}} X_j \right\|_2^2}{n} \le \frac{K_1 \|X_j\|_2^2}{n} \le T^2 K_1, \end{aligned}$$

where T is defined in (B-23). Therefore we have

$$\|\gamma_j\|_2 \le \sqrt{\frac{K_1}{K_0\lambda_-}}\, T. \tag{B-37}$$

Putting (B-32), (B-36), (B-37) and part (ii) together, we obtain that

$$\begin{aligned} \|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2 &\le \lambda_+ \cdot |b_j| \cdot K_3 \sqrt{\frac{K_1}{K_0}} \Delta_C |b_j| \cdot \sqrt{\frac{K_1}{K_0\lambda_-}}\, T \\ &\le \lambda_+ \cdot \frac{1}{n} \frac{2K_1}{(K_0\lambda_-)^2} \Delta_C^2 \mathscr{E} \cdot K_3 \sqrt{\frac{K_1}{K_0}} \Delta_C \cdot \sqrt{\frac{K_1}{K_0\lambda_-}}\, T \\ &= \frac{1}{n} \cdot \frac{2K_1^2 K_3 \lambda_+ T}{K_0^3 \lambda_-^{\frac{5}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E}. \end{aligned}$$

By Lemma B-1,

$$\|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 \le \frac{\|f(\hat{\beta}) - f(\tilde{\mathbf{b}}_\mathbf{j})\|_2}{K_0\lambda_-} = \frac{\|f(\tilde{\mathbf{b}}_\mathbf{j})\|_2}{K_0\lambda_-} \le \frac{1}{n} \cdot \frac{2K_1^2 K_3 \lambda_+ T}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E}.$$

Since $\hat{\beta}_j - b_j$ is the j-th entry of $\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}$, we have

$$|\hat{\beta}_j - b_j| \le \|\hat{\beta} - \tilde{\mathbf{b}}_\mathbf{j}\|_2 \le \frac{1}{n} \cdot \frac{2K_1^2 K_3 \lambda_+ T}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E}.$$

(iv) Similar to part (iii), this result has been shown by [20]. Here we state a refined version for the sake of completeness. Let $\tilde{\mathbf{b}}_{\mathbf{j}}$ be defined as in (B-28), then

$$\begin{aligned} |R_i - r_{i,[j]}| &= \left| x_i^T \hat{\beta} - x_{i,[j]}^T \hat{\beta}_{[j]} \right| = \left| x_i^T (\hat{\beta} - \tilde{\mathbf{b}}_{\mathbf{j}}) + x_i^T \tilde{\mathbf{b}}_{\mathbf{j}} - x_{i,[j]}^T \hat{\beta}_{[j]} \right| \\ &\leq \|x_i\|_2 \cdot \|\hat{\beta} - \tilde{\mathbf{b}}_{\mathbf{j}}\|_2 + \left| x_i^T \tilde{\mathbf{b}}_{\mathbf{j}} - x_{i,[j]}^T \hat{\beta}_{[j]} \right|. \end{aligned}$$

Note that $\|x_i\|_2 \leq \sqrt{n}T$, by part (iii), we have

$$\|x_i\|_2 \cdot \|\hat{\beta} - \tilde{\mathbf{b}}_{\mathbf{j}}\|_2 \leq \frac{1}{\sqrt{n}} \frac{2K_1^2 K_3 \lambda_+ T^2}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E}. \tag{B-38}$$

On the other hand, similar to (B-36), by (B-33),

$$\left| x_i^T \tilde{\mathbf{b}}_{\mathbf{j}} - x_{i,[j]}^T \hat{\beta}_{[j]} \right| \leq \sqrt{\frac{K_1}{K_0}} \cdot |b_j| \cdot \Delta_C \leq \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}} \lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathscr{E}}. \tag{B-39}$$

Therefore,

$$|R_i - r_{i,[j]}| \leq \frac{1}{\sqrt{n}} \left(\frac{2K_1^2 K_3 \lambda_+ T^2}{K_0^4 \lambda_-^{\frac{7}{2}}} \cdot \Delta_C^3 \cdot \mathscr{E} + \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}} \lambda_-} \cdot \Delta_C^2 \cdot \sqrt{\mathscr{E}} \right).$$

□

B-3 Summary of approximation results

Under our technical assumptions, we can derive the rate for approximations via Proposition B.1. This justifies all approximations in "Appendix A".

Theorem B.1 *Under the assumptions **A**1–**A**5,*

(i)

$$T \leq \lambda_+ = O\left(\text{polyLog(n)}\right);$$

(ii)

$$\max_{j \in J_n} |\hat{\beta}_j| \leq \|\hat{\beta}\|_2 = O_{L^4}\left(\text{polyLog(n)}\right);$$

(iii)

$$\max_{j \in J_n} |b_j| = O_{L^2}\left(\frac{\text{polyLog(n)}}{\sqrt{n}} \right);$$

(iv)

$$\max_{j\in J_n} |\hat{\beta}_j - b_j| = O_{L^2}\left(\frac{\text{polyLog(n)}}{n}\right);$$

(v)

$$\max_{j\in J_n}\max_i |R_i - r_{i,[j]}| = O_{L^2}\left(\frac{\text{polyLog(n)}}{\sqrt{n}}\right).$$

Proof (i) Notice that $X_j = Xe_j$, where e_j is the j-th canonical basis vector in $\mathbb{R}^p$, we have

$$\frac{\|X_j\|^2}{n} = e_j^T \frac{X^T X}{n} e_j \le \lambda_+.$$

Similarly, consider the X^T instead of X, we conclude that

$$\frac{\|x_i\|^2}{n} \le \lambda_{\max}\left(\frac{XX^T}{n}\right) = \lambda_+.$$

Recall the definition of T in (B-23), we conclude that

$$T \le \sqrt{\lambda_+} = O\left(\text{polyLog(n)}\right).$$

(ii) Since $\varepsilon_i = u_i(W_i)$ with $\|u_i'\|_\infty \le c_1$, the gaussian concentration property ([36], chapter 1.3) implies that ε_i is c_1^2-sub-gaussian and hence $\mathbb{E}|\varepsilon_i|^k = O(c_1^k)$ for any finite $k > 0$. By Lemma B-2, $|\psi(\varepsilon_i)| \le K_1|\varepsilon_i|$ and hence for any finite k,

$$\mathbb{E}|\psi(\varepsilon_i)|^k \le K_1^k \mathbb{E}|\varepsilon_i|^k = O\left(c_1^k\right).$$

By part (i) of Proposition B.1, using the convexity of x^4 and hence $\left(\frac{a+b}{2}\right)^4 \le \frac{a^4+b^4}{2}$,

$$\mathbb{E}\|\hat{\beta}\|_2^4 \le \frac{1}{(K_0\lambda_-)^4}\mathbb{E}(U+U_0)^4 \le \frac{8}{(K_0\lambda_-)^4}\left(\mathbb{E}U^4 + U_0^4\right).$$

Recall (B-24) that $U = \left\|\frac{1}{n}\sum_{i=1}^n x_i(\psi(\varepsilon_i) - \mathbb{E}\psi(\varepsilon_i))\right\|_2$,

$$U^4 = (U^2)^2 = \frac{1}{n^4}\left(\sum_{i,i'=1}^n x_i^T x_{i'}(\psi(\varepsilon_i) - \mathbb{E}\psi(\varepsilon_i))(\psi(\varepsilon_{i'}) - \mathbb{E}\psi(\varepsilon_{i'}))\right)^2$$
$$= \frac{1}{n^4}\left(\sum_{i=1}^n \|x_i\|_2^2(\psi(\varepsilon_i) - \mathbb{E}\psi(\varepsilon_i))^2\right.$$

$$
\begin{aligned}
&+\sum_{i\neq i'}|x_i^T x_{i'}|(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))(\psi(\varepsilon_{i'})-\mathbb{E}\psi(\varepsilon_{i'}))\Bigg)^2\\
&=\frac{1}{n^4}\Bigg\{\sum_{i=1}^n\|x_i\|_2^4(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^4\\
&+\sum_{i\neq i'}(2|x_i^T x_{i'}|^2+\|x_i\|_2^2\|x_{i'}\|_2^2)(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^2(\psi(\varepsilon_{i'})-\mathbb{E}\psi(\varepsilon_{i'}))^2\\
&+\sum_{\text{others}}|x_i^T x_{i'}|\cdot|x_k^T x_{k'}|\cdot(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))(\psi(\varepsilon_{i'})\\
&-\mathbb{E}\psi(\varepsilon_{i'}))(\psi(\varepsilon_k)-\mathbb{E}\psi(\varepsilon_k))(\psi(\varepsilon_{k'})-\mathbb{E}\psi(\varepsilon_{k'}))\Bigg\}
\end{aligned}
$$

Since $\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i)$ has a zero mean, we have

$$
\begin{aligned}
&\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))(\psi(\varepsilon_{i'})-\mathbb{E}\psi(\varepsilon_{i'}))(\psi(\varepsilon_k)\\
&-\mathbb{E}\psi(\varepsilon_k))(\psi(\varepsilon_{k'})-\mathbb{E}\psi(\varepsilon_{k'}))=0
\end{aligned}
$$

for any $(i,i')\neq(k,k')$ or (k',k) and $i\neq i'$. As a consequence,

$$
\begin{aligned}
\mathbb{E}U^4&=\frac{1}{n^4}\Bigg(\sum_{i=1}^n\|x_i\|_2^4\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^4\\
&+\sum_{i\neq i'}(2|x_i^T x_{i'}|_2^2+\|x_i\|_2^2\|x_{i'}\|_2^2)\mathbb{E}(\psi(\varepsilon_i)\\
&-\mathbb{E}\psi(\varepsilon_i))^2\mathbb{E}(\psi(\varepsilon_{i'})-\mathbb{E}\psi(\varepsilon_{i'}))^2\Bigg)\\
&\leq\frac{1}{n^4}\Bigg(\sum_{i=1}^n\|x_i\|_2^4\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^4\\
&+3\sum_{i\neq i'}\|x_i\|_2^2\|x_{i'}\|_2^2\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^2\mathbb{E}(\psi(\varepsilon_{i'})-\mathbb{E}\psi(\varepsilon_{i'}))^2\Bigg).
\end{aligned}
$$

For any i, using the convexity of x^4, hence $(\frac{a+b}{2})^4\leq\frac{a^4+b^4}{2}$, we have

$$
\begin{aligned}
\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^4&\leq 8\mathbb{E}\left(\psi(\varepsilon_i)^4+(\mathbb{E}\psi(\varepsilon_i))^4\right)\leq 16\mathbb{E}\psi(\varepsilon_i)^4\\
&\leq 16\max_i\mathbb{E}\psi(\varepsilon_i)^4.
\end{aligned}
$$

By Cauchy–Schwartz inequality,

$$
\mathbb{E}(\psi(\varepsilon_i)-\mathbb{E}\psi(\varepsilon_i))^2\leq\mathbb{E}\psi(\varepsilon_i)^2\leq\sqrt{\mathbb{E}\psi(\varepsilon_i)^4}\leq\sqrt{\max_i\mathbb{E}\psi(\varepsilon_i)^4}.
$$

Recall (B-23) that $\|x_i\|_2^2 \le nT^2$ and thus,

$$\begin{aligned}\mathbb{E}U^4 &\le \frac{1}{n^4}\left(16n\cdot n^2T^4 + 3n^2\cdot n^2T^4\right)\cdot \max_i \mathbb{E}\psi(\varepsilon_i)^4 \\ &\le \frac{1}{n^4}\cdot(16n^3+3n^4)T^4 \max_i \mathbb{E}\psi(\varepsilon_i)^4 = O\left(\text{polyLog(n)}\right).\end{aligned}$$

On the other hand, let $\mu^T = (\mathbb{E}\psi(\varepsilon_1),\ldots,\mathbb{E}\psi(\varepsilon_n))$, then $\|\mu\|_2^2 = O(n\cdot \text{polyLog(n)})$ and hence by definition of U_0 in (B-24),

$$U_0 = \frac{\|\mu^T X\|_2}{n} = \frac{1}{n}\sqrt{\mu^T XX^T\mu} \le \sqrt{\frac{\|\mu\|_2^2}{n}\cdot\lambda_+} = O\left(\text{polyLog(n)}\right).$$

In summary,

$$\mathbb{E}\|\hat\beta\|_2^4 = O\left(\text{polyLog(n)}\right).$$

(iii) By mean-value theorem, there exists $a_x \in (0,x)$ such that

$$\rho(x) = \rho(0) + x\psi(0) + \frac{x^2}{2}\psi'(a_x).$$

By assumption **A**1 and Lemma B-2, we have

$$\rho(x) = \frac{x^2}{2}\psi'(a_x) \le \frac{x^2}{2}\|\psi'\|_\infty \le \frac{K_3x^2}{2},$$

where K_3 is defined in Lemma B-2. As a result,

$$\mathbb{E}\rho(\varepsilon_i)^8 \le \left(\frac{K_3}{2}\right)^8 \mathbb{E}\varepsilon_i^{16} = O\left(c_1^{16}\right).$$

Recall the definition of $\mathscr{E}$ in (B-23) and the convexity of x^8, we have

$$\mathbb{E}\mathscr{E}^8 \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\rho(\varepsilon_i)^8 = O(c_1^{16}) = O\left(\text{polyLog(n)}\right). \tag{B-40}$$

Under assumption **A**5, by Cauchy–Schwartz inequality,

$$\mathbb{E}(\Delta_C\sqrt{\mathscr{E}})^2 = \mathbb{E}\Delta_C^2\mathscr{E} \le \sqrt{\mathbb{E}\Delta_C^4}\cdot\sqrt{\mathbb{E}\mathscr{E}^2} = O\left(\text{polyLog(n)}\right).$$

Under assumptions **A**1 and **A**3,

$$\frac{\sqrt{2K_1}}{K_0\lambda_-} = O\left(\text{polyLog(n)}\right).$$

Putting all the pieces together, we obtain that

$$\max_{j\in J_n}|b_j| = O_{L^2}\left(\frac{\text{polyLog(n)}}{\sqrt{n}}\right).$$

(iv) Similarly, by Holder's inequality,

$$\mathbb{E}\left(\Delta_C^3\mathscr{E}\right)^2 = \mathbb{E}\Delta_C^6\mathscr{E}^2 \le \left(\mathbb{E}\Delta_C^8\right)^{\frac{3}{4}}\cdot\left(\mathbb{E}\mathscr{E}^8\right)^{\frac{1}{4}} = O\left(\text{polyLog(n)}\right),$$

and under assumptions **A**1 and **A**3,

$$\frac{2K_1^2K_3\lambda_+T}{K_0^4\lambda_-^{\frac{7}{2}}} = O\left(\text{polyLog(n)}\right).$$

Therefore,

$$\max_{j\in J_n}|\hat{\beta}_j - b_j| = O_{L^2}\left(\frac{\text{polyLog(n)}}{n}\right).$$

(v) It follows from the previous part that

$$\mathbb{E}(\Delta_C^2\cdot\sqrt{\mathscr{E}})^2 = O\left(\text{polyLog(n)}\right).$$

Under assumptions **A**1 and **A**3, the multiplicative factors are also O (polyLog(n)), i.e.

$$\frac{2K_1^2K_3\lambda_+T^2}{K_0^4\lambda_-^{\frac{7}{2}}} = O\left(\text{polyLog(n)}\right), \quad \frac{\sqrt{2}K_1}{K_0^{\frac{3}{2}}\lambda_-} = O\left(\text{polyLog(n)}\right).$$

Therefore,

$$\max_{j\in J_n}\max_i|R_i - r_{i,[j]}| = O_{L^2}\left(\frac{\text{polyLog(n)}}{\sqrt{n}}\right).$$

□

B-4 Controlling gradient and Hessian

Proof (Proof of Lemma 4.1) Recall that $\hat{\beta}$ is the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^n x_i\psi\left(\varepsilon_i - x_i^T\hat{\beta}\right) = 0. \tag{B-41}$$

Taking derivative of (B-41), we have

$$X^T D\left(I - X\frac{\partial\hat{\beta}}{\partial\varepsilon^T}\right) = 0 \Longrightarrow \frac{\partial\hat{\beta}}{\partial\varepsilon^T} = (X^T DX)^{-1}X^T D.$$

This establishes (9). To establishes (10), note that (9) can be rewritten as

$$(X^T DX)\frac{\partial\hat{\beta}}{\partial\varepsilon^T} = X^T D. \tag{B-42}$$

Fix $k \in \{1, \ldots, n\}$. Note that

$$\frac{\partial R_i}{\partial\varepsilon_k} = \frac{\partial\varepsilon_i}{\partial\varepsilon_k} - x_i^T\frac{\partial\hat{\beta}}{\partial\varepsilon_k} = I(i=k) - x_i^T(X^T DX)^{-1}X^T D.$$

Recall that $G = I - X(X^T DX)^{-1}X^T D$, we have

$$\frac{\partial R_i}{\partial\varepsilon_k} = e_i^T G e_k, \tag{B-43}$$

where e_i is the i-th canonical basis of $\mathbb{R}^n$. As a result,

$$\frac{\partial D}{\partial\varepsilon_k} = \tilde{D}\,\mathrm{diag}(Ge_k). \tag{B-44}$$

Taking derivative of (B-42), we have

$$\begin{aligned}
&X^T\frac{\partial D}{\partial\varepsilon_k}X\frac{\partial\hat{\beta}}{\partial\varepsilon^T} + (X^T DX)\frac{\partial\hat{\beta}}{\partial\varepsilon_k\partial\varepsilon^T} = X^T\frac{\partial D}{\partial\varepsilon_k}\\
\Longrightarrow\ &\frac{\partial\hat{\beta}}{\partial\varepsilon_k\partial\varepsilon^T} = (X^T DX)^{-1}X^T\frac{\partial D}{\partial\varepsilon_k}\left(I - X(X^T DX)^{-1}X^T D\right)\\
\Longrightarrow\ &\frac{\partial\hat{\beta}}{\partial\varepsilon_k\partial\varepsilon^T} = (X^T DX)^{-1}X^T\tilde{D}\,\mathrm{diag}(Ge_k)G,
\end{aligned}$$

where $G = I - X(X^T DX)^{-1}X^T D$ is defined in (B-18) in p. 31. Then for each $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, n\}$,

$$\frac{\partial\hat{\beta}_j}{\partial\varepsilon_k\partial\varepsilon^T} = e_j^T(X^T DX)^{-1}X^T\tilde{D}\,\mathrm{diag}(Ge_k)G = e_k^T G^T\,\mathrm{diag}\left(e_j^T(X^T DX)^{-1}X^T\tilde{D}\right)G$$

where we use the fact that $a^T\,\mathrm{diag}(b) = b^T\,\mathrm{diag}(a)$ for any vectors a, b. This implies that

$$\frac{\partial\hat{\beta}_j}{\partial\varepsilon\partial\varepsilon^T} = G^T\,\mathrm{diag}\left(e_j^T(X^T DX)^{-1}X^T\tilde{D}\right)G$$

□

Proof (Proof of Lemma 4.2) Throughout the proof we are using the simple fact that $\|a\|_\infty \le \|a\|_2$. Based on it, we found that

$$\begin{aligned}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty &\le \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2 \\ &= \sqrt{e_j^T(X^TDX)^{-1}X^TDX(X^TDX)^{-1}e_j} \\ &= \sqrt{e_j^T(X^TDX)^{-1}e_j} \le \frac{1}{(nK_0\lambda_-)^{\frac{1}{2}}}. \end{aligned} \tag{B-45}$$

Thus for any $m > 1$, recall that $M_j = \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty$,

$$\begin{aligned}&\mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^m \\ &\le \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty \cdot \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^{m-1} \\ &\le \frac{M_j}{(nK_0\lambda_-)^{\frac{m-1}{2}}}. \end{aligned} \tag{B-46}$$

We should emphasize that we cannot use the naive bound that

$$\begin{aligned}&\mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^m \le \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^m \le \frac{1}{(nK_0\lambda_-)^{\frac{m}{2}}}, \\ &\Longrightarrow \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_{L^m}\left(\frac{\text{polyLog(n)}}{\sqrt{n}}\right) \end{aligned} \tag{B-47}$$

since it fails to guarantee the convergence of TV distance. We will address this issue after deriving Lemma 4.3.

By contrast, as proved below,

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_p(M_j) = O_p\left(\frac{\text{polyLog(n)}}{n}\right) \ll \frac{1}{\sqrt{nK_0\lambda_-}}. \tag{B-48}$$

Thus (B-46) produces a slightly tighter bound

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty = O_{L^m}\left(\frac{\text{polyLog(n)}}{n^{\frac{m+1}{2m}}}\right).$$

It turns out that the above bound suffices to prove the convergence. Although (B-48) implies the possibility to sharpen the bound from $n^{-\frac{m+1}{2m}}$ to n^{-1} using refined analysis, we do not explore this to avoid extra conditions and notation.

- **Bound for κ_{0j}**

First we derive a bound for κ_{0j}. By definition,

$$\kappa_{0j}^2 = \mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_4^4 \le \mathbb{E}\left(\left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_\infty^2 \cdot \left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_2^2\right).$$

By Lemma 4.1 and (B-46) with $m = 2$,

$$\mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_\infty^2 \le \mathbb{E}\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^2 \cdot K_1 = \frac{K_1M_j}{(nK_0\lambda_-)^{\frac{1}{2}}}.$$

On the other hand, it follows from (B-45) that

$$\left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_2^2 = \left\|e_j^T(X^TDX)^{-1}X^TD\right\|_2^2 \le K_1 \cdot \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_2^2 \le \frac{K_1}{nK_0\lambda_-}. \tag{B-49}$$

Putting the above two bounds together we have

$$\kappa_{0j}^2 \le \frac{K_1^2}{(nK_0\lambda_-)^{\frac{3}{2}}} \cdot M_j. \tag{B-50}$$

- **Bound for** κ_{1j}
  As a by-product of (B-49), we obtain that

$$\kappa_{1j}^4 = \mathbb{E}\left\|\frac{\partial\hat{\beta}_j}{\partial\varepsilon^T}\right\|_2^4 \le \frac{K_1^2}{(nK_0\lambda_-)^2}. \tag{B-51}$$

- **Bound for** κ_{2j}
  Finally, we derive a bound for κ_{2j}. By Lemma 4.1, κ_{2j} involves the operator norm of a symmetric matrix with form G^TMG where M is a diagonal matrix. Then by the triangle inequality,

$$\left\|G^TMG\right\|_{op} \le \|M\|_{\text{op}} \cdot \left\|G^TG\right\|_{op} = \|M\|_{\text{op}} \cdot \|G\|_{op}^2.$$

Note that

$$D^{\frac{1}{2}}GD^{-\frac{1}{2}} = I - D^{\frac{1}{2}}X(X^TDX)^{-1}X^TD^{\frac{1}{2}}$$

is a projection matrix, which is idempotent. This implies that

$$\left\|D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right\|_{op} = \lambda_{\max}\left(D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right) \le 1.$$

Write G as $D^{-\frac{1}{2}}(D^{\frac{1}{2}}GD^{-\frac{1}{2}})D^{\frac{1}{2}}$, then we have

$$\|G\|_{op} \le \left\|D^{-\frac{1}{2}}\right\|_{op} \cdot \left\|D^{\frac{1}{2}}GD^{-\frac{1}{2}}\right\|_{op} \cdot \left\|D^{\frac{1}{2}}\right\|_{op} \le \sqrt{\frac{K_1}{K_0}}.$$

Returning to κ_{2j}, we obtain that

$$\begin{aligned}
\kappa_{2j}^4 &= \mathbb{E}\left\|G^T \operatorname{diag}(e_j^T(X^TDX)^{-1}X^T\tilde{D})G\right\|_{op}^4 \\
&\le \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^T\tilde{D}\right\|_\infty^4 \cdot \|G\|_{op}^8\right) \\
&\le \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^T\tilde{D}\right\|_\infty^4\right)\left(\frac{K_1}{K_0}\right)^4 \\
&= \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}D^{-\frac{1}{2}}\tilde{D}\right\|_\infty^4\right)\cdot\left(\frac{K_1}{K_0}\right)^4
\end{aligned}$$

Assumption **A**1 implies that

$$\forall i,\ \frac{|\psi''(R_i)|}{\sqrt{\psi'(R_i)}} \le K_2 \ \&\ \text{hence}\|D^{-\frac{1}{2}}\tilde{D}\|_{\text{op}} \le K_2.$$

Therefore,

$$\left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}D^{-\frac{1}{2}}\tilde{D}\right\|_\infty^4 \le K_2^4 \cdot \left\|e_j^T(X^TDX)^{-1}X^TD^{\frac{1}{2}}\right\|_\infty^4.$$

By (B-46) with $m = 4$,

$$\kappa_{2j}^4 \le \frac{K_2^4}{(n\lambda_-)^{\frac{3}{2}}} \cdot \left(\frac{K_1}{K_0}\right)^4 \cdot M_j. \tag{B-52}$$

□

Proof (Proof of Lemma 4.3) By Theorem B.1, for any j,

$$\mathbb{E}\hat{\beta}_j^4 \le \mathbb{E}\|\hat{\beta}\|_2^4 < \infty.$$

Then using the second-order Poincaré inequality (Proposition 4.1),

$$\max_{j\in J_n} d_{TV}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\operatorname{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = O\left(\frac{c_1c_2\kappa_{0j} + c_1^3\kappa_{1j}\kappa_{2j}}{\operatorname{Var}(\hat{\beta}_j)}\right)$$

$$= O\left(\frac{\frac{M_j^{\frac{1}{2}}}{n^{\frac{3}{4}}} + \frac{M_j^{\frac{1}{4}}}{n^{\frac{7}{8}}}}{\mathrm{Var}(\hat{\beta}_j)} \cdot \mathrm{polyLog(n)}\right) = O\left(\frac{(nM_j^2)^{\frac{1}{4}} + (nM_j^2)^{\frac{1}{8}}}{n\,\mathrm{Var}(\hat{\beta}_j)} \cdot \mathrm{polyLog(n)}\right).$$

It follows from (B-45) that $nM_j^2 = O(\mathrm{polyLog(n)})$ and the above bound can be simplified as

$$\max_{j\in J_n} d_{TV}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = O\left(\frac{(nM_j^2)^{\frac{1}{8}}}{n\,\mathrm{Var}(\hat{\beta}_j)} \cdot \mathrm{polyLog(n)}\right).$$

□

Remark B.1 If we use the naive bound (B-47), by repeating the above derivation, we obtain a worse bound for $\kappa_{0,j} = O(\frac{\mathrm{polyLog(n)}}{n})$ and $\kappa_2 = O(\frac{\mathrm{polyLog(n)}}{\sqrt{n}})$, in which case,

$$\max_{j\in J_n} d_{TV}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = O\left(\frac{\mathrm{polyLog(n)}}{n\,\mathrm{Var}(\hat{\beta}_j)}\right).$$

However, we can only prove that $\mathrm{Var}(\hat{\beta}_j) = \Omega(\frac{1}{n})$. Without the numerator $(nM_j^2)^{\frac{1}{8}}$, which will be shown to be $O(n^{-\frac{1}{8}}\mathrm{polyLog(n)})$ in the next subsection, the convergence cannot be proved.

B-5 Upper bound of M_j

As mentioned in "Appendix A", we should approximate D by $D_{[j]}$ to remove the functional dependence on X_j. To achieve this, we introduce two terms, $M_j^{(1)}$ and $M_j^{(2)}$, defined as

$$\begin{aligned} M_j^{(1)} &= \mathbb{E}\left(\left\|e_j^T(X^TDX)^{-1}X^TD_{[j]}^{\frac{1}{2}}\right\|_\infty\right), \quad M_j^{(2)} \\ &= \mathbb{E}\left(\left\|e_j^T(X^TD_{[j]}X)^{-1}X^TD_{[j]}^{\frac{1}{2}}\right\|_\infty\right). \end{aligned}$$

We will first prove that both $|M_j - M_j^{(1)}|$ and $|M_j^{(1)} - M_j^{(2)}|$ are negligible and then derive an upper bound for $M_j^{(2)}$.

B-5.1 Controlling $|M_j - M_j^{(1)}|$

By Lemma B-2,

$$\left\| D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right\|_\infty \le K_2 \max_i |R_i - r_{i,[j]}| \triangleq K_2 \mathscr{R}_j,$$

and by Theorem B.1,

$$\sqrt{\mathbb{E}\mathscr{R}_j^2} = O\left(\frac{\text{polyLog(n)}}{\sqrt{n}}\right).$$

Then we can bound $|M_j - M_j^{(1)}|$ via the fact that $\|a\|_\infty \le \|a\|_2$ and algebra as follows.

$$\begin{aligned}
|M_j - M_j^{(1)}| &\le \mathbb{E}\left(\left\| e_j^T (X^T D X)^{-1} X^T \left(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right) \right\|_\infty \right) \\
&\le \mathbb{E}\left(\left\| e_j^T (X^T D X)^{-1} X^T \left(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right) \right\|_2 \right) \\
&\le \sqrt{\mathbb{E}\left(\left\| e_j^T (X^T D X)^{-1} X^T \left(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right) \right\|_2^2 \right)} \\
&= \sqrt{\mathbb{E}\left(e_j^T (X^T D X)^{-1} X^T \left(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right)^2 X (X^T D X)^{-1} e_j \right)}.
\end{aligned}$$

By Lemma B-2,

$$\left| \sqrt{\psi'(R_i)} - \sqrt{\psi'(r_{i,[j]})} \right| \le K_2 |R_i - r_{i,[j]}| \le K_2 \mathscr{R}_j,$$

thus

$$\left(D^{\frac{1}{2}} - D_{[j]}^{\frac{1}{2}} \right)^2 \preceq K_2^2 \mathscr{R}_j^2 I \preceq \frac{K_2^2}{K_0} \mathscr{R}_j^2 D.$$

This entails that

$$\begin{aligned}
\left| M_j - M_j^{(1)} \right| &\le K_2 K_0^{-\frac{1}{2}} \sqrt{\mathbb{E}\left(\mathscr{R}_j^2 \cdot e_j^T (X^T D X)^{-1} X^T D X (X^T D X)^{-1} e_j \right)} \\
&= K_2 K_0^{-\frac{1}{2}} \sqrt{\mathbb{E}\left(\mathscr{R}_j^2 \cdot e_j^T (X^T D X)^{-1} e_j \right)} \\
&\le \frac{K_2}{\sqrt{n} K_0 \sqrt{\lambda_-}} \sqrt{\mathbb{E}\left(\mathscr{R}_j^2 \right)} = O\left(\frac{\text{polyLog(n)}}{n} \right).
\end{aligned}$$

B-5.2 Bound of $|M_j^{(1)} - M_j^{(2)}|$

First we prove a useful lemma.

Lemma B-3 *For any symmetric matrix* N *with* $\|N\|_{\text{op}} < 1$,

$$(I - (I+N)^{-1})^2 \preceq \frac{N^2}{(1-\|N\|_{\text{op}})^2}.$$

Proof First, notice that

$$I - (I+N)^{-1} = (I+N-I)(I+N)^{-1} = N(I+N)^{-1},$$

and therefore

$$(I - (I+N)^{-1})^2 = N(I+N)^{-2}N.$$

Since $\|N\|_{\text{op}} < 1$, $I+N$ is positive semi-definite and

$$(I+N)^{-2} \preceq \frac{1}{(1-\|N\|_{\text{op}})^2} I.$$

Therefore,

$$N(I+N)^{-2}N \preceq \frac{N^2}{(1-\|N\|_{\text{op}})^2}.$$

□

We now back to bounding $|M_j^{(1)} - M_j^{(2)}|$. Let $A_j = X^T D_{[j]} X$, $B_j = X^T(D - D_{[j]})X$. By Lemma B-2,

$$\|D - D_{[j]}\|_\infty \le K_3 \max_i |R_i - r_{i,[j]}| = K_3 \mathscr{R}_j$$

and hence

$$\|B_j\|_{\text{op}} \le K_3 \mathscr{R}_j \cdot n\lambda_+ I \triangleq n\eta_j.$$

where $\eta_j = K_3 \lambda_+ \cdot \mathscr{R}_j$. Then by Theorem B.1.(v),

$$\mathbb{E}(\eta_j^2) = O\left(\frac{\text{polyLog(n)}}{n}\right).$$

Using the fact that $\|a\|_\infty \le \|a\|_2$, we obtain that

$$\left|M_j^{(1)} - M_j^{(2)}\right| \le \mathbb{E}\left(\left\| e_j^T A_j^{-1} X^T D_{[j]}^{\frac{1}{2}} - e_j^T (A_j + B_j)^{-1} X^T D_{[j]}^{\frac{1}{2}} \right\|_\infty\right)$$

$$\leq \sqrt{\mathbb{E}\left(\|e_j^T A_j^{-1} X^T D_{[j]}^{\frac{1}{2}} - e_j^T (A_j + B_j)^{-1} X^T D_{[j]}^{\frac{1}{2}}\|_2^2\right)}$$
$$= \sqrt{\mathbb{E}\left[e_j^T (A_j^{-1} - (A_j + B_j)^{-1}) X^T D_{[j]} X (A_j^{-1} - (A_j + B_j)^{-1}) e_j\right]}$$
$$= \sqrt{\mathbb{E}\left[e_j^T (A_j^{-1} - (A_j + B_j)^{-1}) A_j (A_j^{-1} - (A_j + B_j)^{-1}) e_j\right]}$$

The inner matrix can be rewritten as

$$\begin{aligned}
&\left(A_j^{-1} - (A_j + B_j)^{-1}\right) A_j \left(A_j^{-1} - (A_j + B_j)^{-1}\right) \\
&= A_j^{-\frac{1}{2}} \left(I - \left(I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}}\right)^{-1}\right) A_j^{-\frac{1}{2}} A_j A_j^{-\frac{1}{2}} (I \\
&-(I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}})^{-1}\Big) A_j^{-\frac{1}{2}} \\
&= A_j^{-\frac{1}{2}} \left(I - \left(I + A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}}\right)^{-1}\right)^2 A_j^{-\frac{1}{2}}. \qquad \text{(B-53)}
\end{aligned}$$

Let $N_j = A_j^{-\frac{1}{2}} B_j A_j^{-\frac{1}{2}}$, then

$$\begin{aligned}
\|N_j\|_{\text{op}} &\leq \left\|A_j^{-\frac{1}{2}}\right\|_{\text{op}} \cdot \|B_j\|_{\text{op}} \cdot \left\|A_j^{-\frac{1}{2}}\right\|_{\text{op}} \leq (nK_0\lambda_-)^{-\frac{1}{2}} \cdot n\eta_j \cdot (nK_0\lambda_-)^{-\frac{1}{2}} \\
&= \frac{\eta_j}{K_0\lambda_-}.
\end{aligned}$$

On the event $\{\eta_j \leq \frac{1}{2} K_0\lambda_-\}$, $\|N_j\|_{\text{op}} \leq \frac{1}{2}$. By Lemma B-3,

$$(I - (I + N_j)^{-1})^2 \preceq 4N_j^2.$$

This together with (B-53) entails that

$$\begin{aligned}
&e_j^T \left(A_j^{-1} - (A_j + B_j)^{-1}\right) A_j \left(A_j^{-1} - (A_j + B_j)^{-1}\right) e_j \\
&= e_j^T A_j^{-\frac{1}{2}} (I - (I + N_j)^{-1})^2 A_j^{-\frac{1}{2}} e_j \\
&\leq 4 e_j^T A_j^{-\frac{1}{2}} N_j^2 A_j^{-\frac{1}{2}} e_j = e_j^T A_j^{-1} B_j A_j^{-1} B_j A_j^{-1} e_j \leq \left\|A_j^{-1} B_j A_j^{-1} B_j A_j^{-1}\right\|_{\text{op}}.
\end{aligned}$$

Since $A_j \succeq nK_0\lambda_- I$, and $\|B_j\|_{\text{op}} \leq n\eta_j$, we have

$$\left\|A_j^{-1} B_j A_j^{-1} B_j A_j^{-1}\right\|_{\text{op}} \leq \left\|A_j^{-1}\right\|_{\text{op}}^3 \cdot \|B_j\|_{\text{op}}^2 \leq \frac{1}{n} \cdot \frac{1}{(K_0\lambda_-)^3} \cdot \eta_j^2.$$

Thus,

$$\begin{aligned}&\mathbb{E}\left[e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\cdot I\left(\eta_j\le\frac{K_0\lambda_-}{2}\right)\right]\\&\quad\le\mathbb{E}\left[e_j^TA_j^{-1}B_jA_j^{-1}B_jA_j^{-1}e_j\right]\le\frac{1}{n}\cdot\frac{1}{(K_0\lambda_-)^3}\cdot\mathbb{E}\eta_j^2=O\left(\frac{\text{polyLog(n)}}{n^2}\right).\end{aligned}$$

On the event $\{\eta_j > \frac{1}{2}K_0\lambda_-\}$, since $nK_0\lambda_- I \preceq A_j \preceq nK_1\lambda_+ I$ and $A_j + B_j \succeq nK_0\lambda_- I$,

$$\begin{aligned}&\left|e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\right|\\&\le nK_1\lambda_+\cdot\left|e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)^2e_j\right|\\&\le nK_1\lambda_+\cdot\left(2\left|e_j^TA_j^{-2}e_j\right|+2\left|e_j^T(A_j+B_j)^{-2}e_j\right|\right)\\&\le\frac{4nK_1\lambda_+}{(nK_0\lambda_-)^2}=\frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}.\end{aligned}$$

This together with Markov inequality implies htat

$$\begin{aligned}&\mathbb{E}\left[e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\cdot I\left(\eta_j>\frac{K_0\lambda_-}{2}\right)\right]\\&\le\frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}\cdot P\left(\eta_j>\frac{K_0\lambda_-}{2}\right)\\&\le\frac{1}{n}\cdot\frac{4K_1\lambda_+}{(K_0\lambda_-)^2}\cdot\frac{4}{(K_0\lambda_-)^2}\cdot\mathbb{E}\eta_j^2\\&=O\left(\frac{\text{polyLog(n)}}{n^2}\right).\end{aligned}$$

Putting pieces together, we conclude that

$$\begin{aligned}|M_j^{(1)}-M_j^{(2)}|&\le\sqrt{\mathbb{E}\left[e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\right]}\\&\le\sqrt{\mathbb{E}\left[e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\cdot I\left(\eta_j>\frac{K_0\lambda_-}{2}\right)\right]}\\&+\sqrt{\mathbb{E}\left[e_j^T\left(A_j^{-1}-(A_j+B_j)^{-1}\right)A_j\left(A_j^{-1}-(A_j+B_j)^{-1}\right)e_j\cdot I\left(\eta_j\le\frac{K_0\lambda_-}{2}\right)\right]}\\&=O\left(\frac{\text{polyLog(n)}}{n}\right).\end{aligned}$$

B-5.3 Bound of $M_j^{(2)}$

Similar to (A-1), by block matrix inversion formula (See Proposition E.1),

$$e_j^T (X^T D_{[j]} X)^{-1} X^T D_{[j]}^{\frac{1}{2}} = \frac{X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j)}{X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j) D_{[j]}^{\frac{1}{2}} X_j},$$

where $H_j = D_{[j]}^{\frac{1}{2}} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}$. Recall that $\xi_j \geq K_0 \lambda_-$ by (B-25), so we have

$$X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j) D_{[j]}^{\frac{1}{2}} X_j = n\xi_j \geq n\lambda_-.$$

As for the numerator, recalling the definition of $h_{j,1,i}$, we obtain that

$$\begin{aligned} \|X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j)\|_\infty &= \left\| \frac{1}{n} X_j^T (I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}) \cdot D_{[j]}^{\frac{1}{2}} \right\|_\infty \\ &\leq \sqrt{K_1} \cdot \left\| \frac{1}{n} X_j^T (I - D_{[j]} X_{[j]} (X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}) \right\|_\infty \\ &= \sqrt{K_1} \max_i \left| h_{j,1,i}^T X_j \right| \leq \sqrt{K_1} \Delta_C \max_i \|h_{j,1,i}\|_2. \end{aligned}$$

As proved in (B-35),

$$\max_i \|h_{j,1,i}\|_2 \leq \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}.$$

This entails that

$$\left\| X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j) \right\|_\infty \leq \frac{K_1}{\sqrt{K_0}} \cdot \Delta_C = O_{L^1}(\text{polyLog(n)}).$$

Putting the pieces together we conclude that

$$M_j^{(2)} \leq \frac{\mathbb{E} \left\| X_j^T D_{[j]}^{\frac{1}{2}} (I - H_j) \right\|_\infty}{n\lambda_-} = O\left(\frac{\text{polyLog(n)}}{n} \right).$$

B-5.4 Summary

Based on results from Sections B.5.1–B.5.3, we have

$$M_j = O\left(\frac{\text{polyLog(n)}}{n} \right).$$

Note that the bounds we obtained do not depend on j, so we conclude that

$$\max_{j\in J_n} M_j = O\left(\frac{\text{polyLog(n)}}{n}\right).$$

B-6 Lower Bound of $\text{Var}(\hat{\boldsymbol{\beta}}_j)$

B-6.1 Approximating $\text{Var}(\hat{\beta}_j)$ *by* $\text{Var}(b_j)$

By Theorem B.1,

$$\max_j \mathbb{E}(\hat{\beta}_j - b_j)^2 = O\left(\frac{\text{polyLog(n)}}{n^2}\right), \quad \max_j \mathbb{E}b_j^2 = O\left(\frac{\text{polyLog(n)}}{n}\right).$$

Using the fact that

$$\hat{\beta}_j^2 - b_j^2 = (\hat{\beta}_j - b_j + b_j)^2 - b_j^2 = (\hat{\beta}_j - b_j)^2 + 2(\hat{\beta}_j - b_j)b_j,$$

we can bound the difference between $\mathbb{E}\hat{\beta}_j^2$ and $\mathbb{E}b_j^2$ by

$$\begin{aligned}\left|\mathbb{E}\hat{\beta}_j^2 - \mathbb{E}b_j^2\right| &= \mathbb{E}(\hat{\beta}_j - b_j)^2 + 2|\mathbb{E}(\hat{\beta}_j - b_j)b_j| \le \mathbb{E}(\hat{\beta}_j - b_j)^2 \\ &\quad + 2\sqrt{\mathbb{E}(\hat{\beta}_j - b_j)^2}\sqrt{\mathbb{E}b_j^2} = O\left(\frac{\text{polyLog(n)}}{n^{\frac{3}{2}}}\right).\end{aligned}$$

Similarly, since $|a^2 - b^2| = |a - b| \cdot |a + b| \le |a - b|(|a - b| + 2|b|)$,

$$|(\mathbb{E}\hat{\beta}_j)^2 - (\mathbb{E}b_j)^2| \le \mathbb{E}|\hat{\beta}_j - b_j| \cdot \left(\mathbb{E}|\hat{\beta}_j - b_j| + 2\mathbb{E}|b_j|\right) = O\left(\frac{\text{polyLog(n)}}{n^{\frac{3}{2}}}\right).$$

Putting the above two results together, we conclude that

$$\left|\text{Var}(\hat{\beta}_j) - \text{Var}(b_j)\right| = O\left(\frac{\text{polyLog(n)}}{n^{\frac{3}{2}}}\right). \tag{B-54}$$

Then it is left to show that

$$\text{Var}(b_j) = \Omega\left(\frac{1}{n \cdot \text{polyLog(n)}}\right).$$

B-6.2 Controlling $\text{Var}(b_j)$ *by* $\text{Var}(N_j)$

Recall that

$$b_j = \frac{1}{\sqrt{n}}\frac{N_j}{\xi_j}$$

where

$$N_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij} \psi(r_{i,[j]}), \quad \xi_j$$
$$= \frac{1}{n} X_j^T \left(D_{[j]} - D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]} \right) X_j.$$

Then

$$n \operatorname{Var}(b_j) = \mathbb{E} \left(\frac{N_j}{\xi_j} - \mathbb{E} \frac{N_j}{\xi_j} \right)^2 = \mathbb{E} \left(\frac{N_j - \mathbb{E} N_j}{\xi_j} + \frac{\mathbb{E} N_j}{\xi_j} - \mathbb{E} \frac{N_j}{\xi_j} \right)^2.$$

Using the fact that $(a+b)^2 - (\frac{1}{2}a^2 - b^2) = \frac{1}{2}(a+2b)^2 \geq 0$, we have

$$n \operatorname{Var}(b_j) \geq \frac{1}{2} \mathbb{E} \left(\frac{N_j - \mathbb{E} N_j}{\xi_j} \right)^2 - \mathbb{E} \left(\frac{\mathbb{E} N_j}{\xi_j} - \mathbb{E} \frac{N_j}{\xi_j} \right)^2 \triangleq \frac{1}{2} I_1 - I_2. \tag{B-55}$$

B-6.3 Controlling I_1

The Assumption **A**4 implies that

$$\operatorname{Var}(N_j) = \frac{1}{n} X_j^T Q_j X_j = \Omega \left(\frac{\operatorname{tr}(\operatorname{Cov}(h_{j,0}))}{n \cdot \operatorname{polyLog(n)}} \right).$$

It is left to show that $\operatorname{tr}(\operatorname{Cov}(h_{j,0}))/n = \Omega \left(\frac{1}{\operatorname{polyLog(n)}} \right)$. Since this result will also be used later in "Appendix C", we state it in the following the lemma.

Lemma B-4 *Under assumptions **A**1 - **A**3,*

$$\frac{\operatorname{tr}(\operatorname{Cov}(\psi(h_{j,0})))}{n} \geq \frac{K_0^4}{K_1^2} \cdot \left(\frac{n-p+1}{n} \right)^2 \cdot \min_i \operatorname{Var}(\varepsilon_i) = \Omega \left(\frac{1}{\operatorname{polyLog(n)}} \right).$$

Proof The (A-10) implies that

$$\operatorname{Var}(\psi(r_{i,[j]})) \geq K_0^2 \operatorname{Var}(r_{i,[j]}). \tag{B-56}$$

Note that $r_{i,[j]}$ is a function of ε, we can apply (A-10) again to obtain a lower bound for $\operatorname{Var}(r_{i,[j]})$. In fact, by variance decomposition formula, using the independence of ε_i's,

$$\operatorname{Var}(r_{i,[j]}) = \mathbb{E} \left(\operatorname{Var} \left(r_{i,[j]} \middle| \varepsilon_{(i)} \right) \right) + \operatorname{Var} \left(\mathbb{E} \left(r_{i,[j]} \middle| \varepsilon_{(i)} \right) \right) \geq \mathbb{E} \left(\operatorname{Var} \left(r_{i,[j]} \middle| \varepsilon_{(i)} \right) \right),$$

where $\varepsilon_{(i)}$ includes all but the i-th entry of ε. Apply A-10 again,

$$\operatorname{Var} \left(r_{i,[j]} \middle| \varepsilon_{(i)} \right) \geq \inf_{\varepsilon_i} \left| \frac{\partial r_{i,[j]}}{\partial \varepsilon_i} \right|^2 \cdot \operatorname{Var}(\varepsilon_i),$$

and hence

$$\mathrm{Var}(r_{i,[j]}) \geq \mathbb{E}\,\mathrm{Var}\left(r_{i,[j]}\middle|\varepsilon_{(i)}\right) \geq \mathbb{E}\inf_{\varepsilon}\left|\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}\right|^2 \cdot \mathrm{Var}(\varepsilon_i). \tag{B-57}$$

Now we compute $\frac{\partial r_{i,[j]}}{\partial \varepsilon_i}$. Similar to (B-43) in p. 40, we have

$$\frac{\partial r_{k,[j]}}{\partial \varepsilon_i} = e_i^T G_{[j]} e_k, \tag{B-58}$$

where $G_{[j]}$ is defined in (B-18) in p. 31. When $k = i$,

$$\frac{\partial r_{i,[j]}}{\partial \varepsilon_i} = e_i^T G_{[j]} e_i = e_i^T D_{[j]}^{-\frac{1}{2}} D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} D_{[j]}^{\frac{1}{2}} e_i = e_i^T D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} e_i. \tag{B-59}$$

By definition of $G_{[j]}$,

$$D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} = I - D_{[j]}^{\frac{1}{2}} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}}.$$

Let $\tilde{X}_{[j]} = D_{[j]}^{\frac{1}{2}} X_{[j]}$ and $H_j = \tilde{X}_{[j]}(\tilde{X}_{[j]}^T \tilde{X}_{[j]})^{-1} \tilde{X}_{[j]}^T$. Denote by $\tilde{X}_{(i),[j]}$ the matrix $\tilde{X}_{[j]}$ after removing i-th row, then by block matrix inversion formula (See Proposition E.1),

$$\begin{aligned}
e_i^T H_j e_i &= \tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]} + \tilde{x}_{i,[j]} \tilde{x}_{i,[j]}^T\right)^{-1} \tilde{x}_{i,[j]} \\
&= \tilde{x}_{i,[j]}^T \left(\left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \right. \\
&\quad \left. - \frac{\left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \tilde{x}_{i,[j]} \tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1}}{1 + \tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \tilde{x}_{i,[j]}} \right) \tilde{x}_{i,[j]} \\
&= \frac{\tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \tilde{x}_{i,[j]}}{1 + \tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \tilde{x}_{i,[j]}}.
\end{aligned}$$

This implies that

$$\begin{aligned}
e_i^T D_{[j]}^{\frac{1}{2}} G_{[j]} D_{[j]}^{-\frac{1}{2}} e_i = e_i^T (I - H_j) e_i &= \frac{1}{1 + \tilde{x}_{i,[j]}^T \left(\tilde{X}_{(i),[j]}^T \tilde{X}_{(i),[j]}\right)^{-1} \tilde{x}_{i,[j]}} \\
&= \frac{1}{1 + e_i^T D_{[j]}^{\frac{1}{2}} X_{[j]} \left(X_{(i),[j]}^T D_{(i),[j]} X_{(i),[j]}\right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} e_i}
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{1+K_0^{-1}e_i^T D_{[j]}^{\frac{1}{2}} X_{[j]}\left(X_{(i),[j]}^T X_{(i),[j]}\right)^{-1} X_{[j]}^T D_{[j]}^{\frac{1}{2}} e_i} \\
&= \frac{1}{1+K_0^{-1}(D_{[j]})_{i,i}\cdot e_i^T X_{[j]}\left(X_{(i),[j]}^T X_{(i),[j]}\right)^{-1} X_{[j]}^T e_i} \\
&\geq \frac{1}{1+K_0^{-1}K_1 e_i^T X_{[j]}\left(X_{(i),[j]}^T X_{(i),[j]}\right)^{-1} X_{[j]}^T e_i} \\
&\geq \frac{K_0}{K_1}\cdot\frac{1}{1+e_i^T X_{[j]}\left(X_{(i),[j]}^T X_{(i),[j]}\right)^{-1} X_{[j]}^T e_i}. \qquad \text{(B-60)}
\end{aligned}$$

Apply the above argument that replaces H_j by $X_{[j]}(X_{[j]}^T X_{[j]})^{-1}X_{[j]}^T$, we have

$$\frac{1}{1+e_i^T X_{[j]}^T\left(X_{(i),[j]}^T X_{(i),[j]}\right)^{-1} X_{[j]} e_i} = e_i^T\left(I - X_{[j]}\left(X_{[j]}^T X_{[j]}\right)^{-1} X_{[j]}^T\right) e_i.$$

Thus, by (B-56) and (B-57),

$$\mathrm{Var}(\psi(r_{i,[j]})) \geq \frac{K_0^4}{K_1^2}\cdot\left[e_i^T\left(I - X_{[j]}\left(X_{[j]}^T X_{[j]}\right)^{-1} X_{[j]}^T\right) e_i\right]^2.$$

Summing i over $1,\ldots,n$, we obtain that

$$\begin{aligned}
\frac{\mathrm{tr}(\mathrm{Cov}(h_{j,0}))}{n} &\geq \frac{K_0^4}{K_1^2}\cdot\frac{1}{n}\sum_{i=1}^n\left[e_i^T\left(I - X_{[j]}\left(X_{[j]}^T X_{[j]}\right)^{-1} X_{[j]}^T\right) e_i\right]^2\cdot\min_i \mathrm{Var}(\varepsilon_i) \\
&\geq \frac{K_0^4}{K_1^2}\cdot\left(\frac{1}{n}\,\mathrm{tr}\left(I - X_{[j]}\left(X_{[j]}^T X_{[j]}\right)^{-1} X_{[j]}^T\right)\right)^2\cdot\min_i \mathrm{Var}(\varepsilon_i) \\
&= \frac{K_0^4}{K_1^2}\cdot\left(\frac{n-p+1}{n}\right)^2\cdot\min_i \mathrm{Var}(\varepsilon_i)
\end{aligned}$$

Since $\min_i \mathrm{Var}(\varepsilon_i) = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right)$ by assumption **A**2, we conclude that

$$\frac{\mathrm{tr}(\mathrm{Cov}(h_{j,0}))}{n} = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

□

In summary,

$$\mathrm{Var}(N_j) = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

Recall that

$$\begin{aligned}\xi_j &= \frac{1}{n} X_j^T \left(D_{[j]} - D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]} \right)^{-1} X_{[j]}^T D_{[j]} \right) X_j \\ &\leq \frac{1}{n} X_j^T D_{[j]} X_j \leq K_1 T^2,\end{aligned}$$

we conclude that

$$I_1 \geq \frac{\mathrm{Var}(N_j)}{(K_1 T^2)^2} = \Omega\left(\frac{1}{\mathrm{polyLog(n)}} \right). \tag{B-61}$$

B-6.4 Controlling I_2

By definition,

$$\begin{aligned}I_2 &= \mathbb{E}\left(\mathbb{E}N_j \left(\frac{1}{\xi_j} - \mathbb{E}\frac{1}{\xi_j} \right) + \mathbb{E}N_j \mathbb{E}\frac{1}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j} \right)^2 \\ &= \mathrm{Var}\left(\frac{\mathbb{E}N_j}{\xi_j} \right) + \left(\mathbb{E}N_j \mathbb{E}\frac{1}{\xi_j} - \mathbb{E}\frac{N_j}{\xi_j} \right)^2 \\ &= (\mathbb{E}N_j)^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j} \right) + \mathrm{Cov}\left(N_j, \frac{1}{\xi_j} \right)^2 \\ &\leq (\mathbb{E}N_j)^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j} \right) + \mathrm{Var}(N_j)\, \mathrm{Var}\left(\frac{1}{\xi_j} \right) \\ &= \mathbb{E}N_j^2 \cdot \mathrm{Var}\left(\frac{1}{\xi_j} \right).\end{aligned} \tag{B-62}$$

By (B-27) in the proof of Theorem B.1,

$$\mathbb{E}N_j^2 \leq 2K_1 \mathbb{E}\left(\mathscr{E} \cdot \Delta_C^2 \right) \leq 2K_1 \sqrt{\mathbb{E}\mathscr{E}^2 \cdot \mathbb{E}\Delta_C^4} = O\left(\mathrm{polyLog(n)} \right),$$

where the last equality uses the fact that $\mathscr{E} = O_{L^2}(\mathrm{polyLog(n)})$ as proved in (B-40). On the other hand, let $\tilde{\xi}_j$ be an independent copy of ξ_j, then

$$\mathrm{Var}\left(\frac{1}{\xi_j} \right) = \frac{1}{2}\mathbb{E}\left(\frac{1}{\xi_j} - \frac{1}{\tilde{\xi}_j} \right)^2 = \frac{1}{2}\mathbb{E}\frac{(\xi_j - \tilde{\xi}_j)^2}{\xi_j^2 \tilde{\xi}_j^2}.$$

Since $\xi_j \geq K_0 \lambda_-$ as shown in (B-25), we have

$$\mathrm{Var}\left(\frac{1}{\xi_j} \right) \leq \frac{1}{2(K_0\lambda_-)^4}\mathbb{E}(\xi_j - \tilde{\xi}_j)^2 = \frac{1}{(K_0\lambda_-)^4} \cdot \mathrm{Var}(\xi_j). \tag{B-63}$$

To bound $\mathrm{Var}(\xi_j)$, we propose to using the standard Poincaré inequality [11], which is stated as follows.

Proposition B.2 *Let $W = (W_1, \ldots, W_n) \sim N(0, I_{n\times n})$ and f be a twice differentiable function, then*

$$\mathrm{Var}(f(W)) \le \mathbb{E}\left\| \frac{\partial f(W)}{\partial W} \right\|_2^2.$$

In our case, $\varepsilon_i = u_i(W_i)$, and hence for any twice differentiable function g,

$$\mathrm{Var}(g(\varepsilon)) \le \mathbb{E}\left\| \frac{\partial g(\varepsilon)}{\partial W} \right\|_2^2 = \mathbb{E}\left\| \frac{\partial g(\varepsilon)}{\partial \varepsilon} \cdot \frac{\partial \varepsilon}{\partial W^T} \right\|_2^2 \le \max_i \left\|u_i'\right\|_\infty^2 \cdot \mathbb{E}\left\| \frac{\partial g(\varepsilon)}{\partial \varepsilon} \right\|_2^2.$$

Applying it to ξ_j, we have

$$\mathrm{Var}(\xi_j) \le c_1^2 \cdot \mathbb{E}\left\| \frac{\partial \xi_j}{\partial \varepsilon} \right\|_2^2. \tag{B-64}$$

For given $k \in \{1, \ldots, n\}$, using the chain rule and the fact that $dB^{-1} = -B^{-1}dB\,B^{-1}$ for any square matrix B, we obtain that

$$\begin{aligned}
&\frac{\partial}{\partial \varepsilon_k}\left(D_{[j]} - D_{[j]}X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}\right)\\
&= \frac{\partial D_{[j]}}{\partial \varepsilon_k} - \frac{\partial D_{[j]}}{\partial \varepsilon_k} X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}\\
&- D_{[j]}X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k}\\
&+ D_{[j]}X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} X_{[j]}\left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T D_{[j]}\\
&= G_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} G_{[j]}
\end{aligned}$$

where $G_{[j]} = I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}$ as defined in last subsection. This implies that

$$\frac{\partial \xi_j}{\partial \varepsilon_k} = \frac{1}{n} X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} G_{[j]} X_j.$$

Then (B-64) entails that

$$\mathrm{Var}(\xi_j) \le \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\left(X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} G_{[j]} X_j\right)^2 \tag{B-65}$$

First we compute $\frac{\partial D_{[j]}}{\partial \varepsilon_k}$. Similar to (B-44) in p. 40 and recalling the definition of $D_{[j]}$ in (B-17) and that of $G_{[j]}$ in (B-18) in p. 31, we have

$$\frac{\partial D_{[j]}}{\partial \varepsilon_k} = \tilde{D}_{[j]} \operatorname{diag}(G_{[j]} e_k) \operatorname{diag}(\tilde{D}_{[j]} G_{[j]} e_k),$$

Let $\mathscr{X}_j = G_{[j]} X_j$ and $\tilde{\mathscr{X}_j} = \mathscr{X}_j \circ \mathscr{X}_j$ where $\circ$ denotes Hadamard product. Then

$$\begin{aligned} X_j^T G_{[j]}^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} G_{[j]} X_j = \mathscr{X}_j^T \frac{\partial D_{[j]}}{\partial \varepsilon_k} \mathscr{X}_j &= \mathscr{X}_j^T \operatorname{diag}(\tilde{D}_{[j]} G_{[j]} e_k) \mathscr{X}_j \\ &= \tilde{\mathscr{X}_j}^T \tilde{D}_{[j]} G_{[j]} e_k. \end{aligned}$$

Here we use the fact that for any vectors $x, a \in \mathbb{R}^n$,

$$x^T \operatorname{diag}(a) x = \sum_{i=1}^n a_i x_i^2 = (x \circ x)^T a.$$

This together with (B-65) imply that

$$\begin{aligned} \operatorname{Var}(\xi_j) &\le \frac{1}{n^2} \sum_{k=1}^n \mathbb{E} \left(\tilde{\mathscr{X}_j}^T \tilde{D}_{[j]} G_{[j]} e_k \right)^2 = \frac{1}{n^2} \mathbb{E} \left\| \tilde{\mathscr{X}_j}^T \tilde{D}_{[j]} G_{[j]} \right\|_2^2 \\ &= \frac{1}{n^2} \mathbb{E} \left(\tilde{\mathscr{X}_j}^T \tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]} \tilde{\mathscr{X}_j} \right) \end{aligned}$$

Note that $G_{[j]} G_{[j]}^T \preceq \|G_{[j]}\|_{\text{op}}^2 I$, and $\tilde{D}_{[j]} \preceq K_3 I$ by Lemma B-2 in p. 32. Therefore we obtain that

$$\begin{aligned} \operatorname{Var}(\xi_j) &\le \frac{1}{n^2} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \tilde{\mathscr{X}_j}^T \tilde{D}_{[j]}^2 \tilde{\mathscr{X}_j} \right) \le \frac{K_3^2}{n^2} \cdot \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \| \tilde{\mathscr{X}_j} \|_2^2 \right) \\ &= \frac{K_3^2}{n^2} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \| \mathscr{X}_j \|_4^4 \right) \le \frac{K_3^2}{n} \mathbb{E} \left(\left\| G_{[j]} \right\|_{op}^2 \cdot \| \mathscr{X}_j \|_\infty^4 \right) \end{aligned}$$

As shown in (B-34),

$$\|G_{[j]}\|_{\text{op}} \le \left(\frac{K_1}{K_0} \right)^{\frac{1}{2}}.$$

On the other hand, notice that the i-th row of $G_{[j]}$ is $h_{j,1,i}$ (see (B-20) for definition), by definition of Δ_C we have

$$\| \mathscr{X}_j \|_\infty = \| G_{[j]} X_j \|_\infty = \max_i \left| h_{j,1,i}^T X_j \right| \le \Delta_C \cdot \max \| h_{j,1,i} \|_2.$$

By (B-35) and assumption **A5**,

$$\|\mathscr{X}_j\|_\infty \leq \Delta_C \cdot \left(\frac{K_1}{K_0}\right)^{\frac{1}{2}} = O_{L^4}\left(\text{polyLog(n)}\right).$$

This entails that

$$\text{Var}(\xi_j) = O\left(\frac{\text{polyLog(n)}}{n}\right).$$

Combining with (B-62) and (B-63), we obtain that

$$I_2 = O\left(\frac{\text{polyLog(n)}}{n}\right).$$

B-6.5 Summary

Putting (B-55), (B-61) and (B-62) together, we conclude that

$$\begin{aligned} n\,\text{Var}(b_j) &= \Omega\left(\frac{1}{\text{polyLog(n)}}\right) - O\left(\frac{1}{n\cdot\text{polyLog(n)}}\right) \\ &= \Omega\left(\frac{1}{\text{polyLog(n)}}\right) \Longrightarrow \text{Var}(b_j) = \Omega\left(\frac{1}{n\cdot\text{polyLog(n)}}\right). \end{aligned}$$

Combining with (B-54),

$$\text{Var}(\hat{\beta}_j) = \Omega\left(\frac{1}{n\cdot\text{polyLog(n)}}\right).$$

C Proof of other results

C-1 Proofs of propositions in Section 2.3

Proof (Proof of Proposition 2.1) Let $H_i(\alpha) = \mathbb{E}\rho(\varepsilon_i - \alpha)$. First we prove that the conditions imply that 0 is the unique minimizer of $H_i(\alpha)$ for all i. In fact, since $\varepsilon_i \overset{d}{=} -\varepsilon_i$,

$$H_i(\alpha) = \mathbb{E}\rho(\varepsilon_i - \alpha) = \frac{1}{2}\left(\mathbb{E}\rho(\varepsilon_i - \alpha) + \rho(-\varepsilon_i - \alpha)\right).$$

Using the fact that ρ is even, we have

$$H_i(\alpha) = \mathbb{E}\rho(\varepsilon_i - \alpha) = \frac{1}{2}\left(\mathbb{E}\rho(\varepsilon_i - \alpha) + \rho(\varepsilon_i + \alpha)\right).$$

By (4), for any $\alpha \neq 0$, $H_i(\alpha) > H_i(0)$. As a result, 0 is the unique minimizer of H_i. Then for any $\beta \in \mathbb{R}^p$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho\left(y_i - x_i^T\beta\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho\left(\varepsilon_i - x_i^T(\beta-\beta^*)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}H_i\left(x_i^T(\beta-\beta^*)\right) \geq \frac{1}{n}\sum_{i=1}^{n}H_i(0).\end{aligned}$$

The equality holds iff $x_i^T(\beta-\beta^*) = 0$ for all i since 0 is the unique minimizer of H_i. This implies that

$$X(\beta^*(\rho) - \beta^*) = 0.$$

Since X has full column rank, we conclude that

$$\beta^*(\rho) = \beta^*.$$

□

Proof (Proof of Proposition 2.2) For any $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$, let

$$G(\alpha;\beta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho\left(y_i - \alpha - x_i^T\beta\right).$$

Since α_ρ minimizes $\mathbb{E}\rho(\varepsilon_i - \alpha)$, it holds that

$$G(\alpha;\beta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho\left(\varepsilon_i - \alpha - x_i^T(\beta-\beta^*)\right) \geq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho(\varepsilon_i - \alpha_\rho) = G(\alpha_\rho,\beta^*).$$

Note that α_ρ is the unique minimizer of $\mathbb{E}\rho(\varepsilon_i - \alpha)$, the above equality holds if and only if

$$\alpha + x_i^T(\beta-\beta^*) \equiv \alpha_\rho \Longrightarrow (\mathbf{1}\ X)\begin{pmatrix}\alpha-\alpha_\rho\\ \beta-\beta^*\end{pmatrix} = 0.$$

Since $(\mathbf{1}\ X)$ has full column rank, it must hold that $\alpha = \alpha_\rho$ and $\beta = \beta^*$. □

C-2 Proofs of Corollary 3.1

Proposition C.1 *Suppose that ε_i are i.i.d. such that $\mathbb{E}\rho(\varepsilon_1 - \alpha)$ as a function of α has a unique minimizer α_ρ. Further assume that $X_{J_n^c}$ contains an intercept term, X_{J_n} has full column rank and*

$$\operatorname{span}(\{X_j : j \in J_n\}) \cap \operatorname{span}\left(\left\{X_j : j \in J_n^c\right\}\right) = \{0\} \tag{C-66}$$

Let

$$\beta_{J_n}(\rho) = \arg\min_{\beta_{J_n}} \left\{ \min_{\beta_{J_n^c}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho\left(y_i - x_i^T \beta\right) \right\}.$$

Then $\beta_{J_n}(\rho) = \beta^*_{J_n}$.

Proof let

$$G(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho\left(y_i - x_i^T \beta\right).$$

For any minimizer $\beta(\rho)$ of G, which might not be unique, we prove that $\beta_{J_n}(\rho) = \beta^*_{J_n}$. It follows by the same argument as in Proposition 2.2 that

$$\begin{aligned} x_i^T(\beta(\rho) - \beta^*) \equiv \alpha_0 \Longrightarrow X(\beta(\rho) - \beta^*) = \alpha_0 \mathbf{1} \Longrightarrow X_{J_n}(\beta_{J_n}(\rho)) \\ = -X_{J_n^c}\left(\beta(\rho)_{J_n^c} - \beta^*_{J_n^c}\right) + \alpha_0 \mathbf{1}. \end{aligned}$$

Since $X_{J_n^c}$ contains the intercept term, we have

$$X_{J_n}(\beta_{J_n}(\rho) - \beta^*_{J_n}) \in \operatorname{span}\left(\left\{X_j : j \in J_n^c\right\}\right).$$

It then follows from (C-68) that

$$X_{J_n}\left(\beta_{J_n}(\rho) - \beta^*_{J_n}\right) = 0.$$

Since X_{J_n} has full column rank, we conclude that

$$\beta_{J_n}(\rho) = \beta^*_{J_n}.$$

□

The Proposition C.1 implies that $\beta^*_{J_n}$ is identifiable even when X is not of full column rank. A similar conclusion holds for the estimator $\hat{\beta}_{J_n}$ and the residuals R_i. The following two propositions show that under certain assumptions, $\hat{\beta}_{J_n}$ and R_i are invariant to the choice of $\hat{\beta}$ in the presense of multiple minimizers.

Proposition C.2 *Suppose that* ρ *is convex and twice differentiable with* $\rho''(x) > c > 0$ *for all* $x \in \mathbb{R}$. *Let* $\hat{\beta}$ *be any minimizer, which might not be unique, of*

$$F(\beta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \rho\left(y_i - x_i^T \beta\right)$$

Then $R_i = y_i - x_i \hat{\beta}$ *is independent of the choice of* $\hat{\beta}$ *for any* i.

Proof The conclusion is obvious if $F(\beta)$ has a unique minimizer. Otherwise, let $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ be two different minimizers of F denote by η their difference, i.e. $\eta = \hat{\beta}^{(2)} - \hat{\beta}^{(1)}$. Since F is convex, $\hat{\beta}^{(1)} + v\eta$ is a minimizer of F for all $v \in [0, 1]$. By Taylor expansion,

$$F(\hat{\beta}^{(1)} + v\eta) = F(\hat{\beta}^{(1)}) + v\nabla F(\hat{\beta}^{(1)})\eta + \frac{v^2}{2}\eta^T \nabla^2 F(\hat{\beta}^{(1)})\eta + o(v^2).$$

Since both $\hat{\beta}^{(1)} + v\eta$ and $\hat{\beta}^{(1)}$ are minimizers of F, we have $F(\hat{\beta}^{(1)} + v\eta) = F(\hat{\beta}^{(1)})$ and $\nabla F(\hat{\beta}^{(1)}) = 0$. By letting v tend to 0, we conclude that

$$\eta^T \nabla^2 F(\hat{\beta}^{(1)})\eta = 0.$$

The hessian of F can be written as

$$\nabla^2 F(\hat{\beta}^{(1)}) = \frac{1}{n} X^T \operatorname{diag}\left(\rho''(y_i - x_i^T \hat{\beta}^{(1)})\right) X \succeq \frac{cX^T X}{n}.$$

Thus, η satisfies that

$$\eta^T \frac{cX^T X}{n}\eta = 0 \Longrightarrow X\eta = 0. \tag{C-67}$$

This implies that

$$y - X\hat{\beta}^{(1)} = y - X\hat{\beta}^{(2)}$$

and hence R_i is the same for all i in both cases. □

Proposition C.3 *Suppose that ρ is convex and twice differentiable with $\rho''(x) > c > 0$ for all $x \in \mathbb{R}$. Further assume that X_{J_n} has full column rank and*

$$\operatorname{span}(\{X_j : j \in J_n\}) \cap \operatorname{span}\left(\left\{X_j : j \in J_n^c\right\}\right) = \{0\} \tag{C-68}$$

Let $\hat{\beta}$ be any minimizer, which might not be unique, of

$$F(\beta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - x_i^T \beta\right)$$

Then $\hat{\beta}_{J_n}$ is independent of the choice of $\hat{\beta}$.

Proof As in the proof of Proposition C.2, we conclude that for any minimizers $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$, $X\eta = 0$ where $\eta = \hat{\beta}^{(2)} - \hat{\beta}^{(1)}$. Decompose the term into two parts, we have

$$X_{J_n}\eta_{J_n} = -X_{J_n}^c \eta_{J_n^c} \in \operatorname{span}\left(\left\{X_j : j \in J_n^c\right\}\right).$$

It then follows from (C-68) that $X_{J_n}\eta_{J_n} = 0$. Since X_{J_n} has full column rank, we conclude that $\eta_{J_n} = 0$ and hence $\hat{\beta}_{J_n}^{(1)} = \hat{\beta}_{J_n}^{(2)}$. □

Proof (Proof of Corollary 3.1) Under assumption **A**3*, X_{J_n} must have full column rank. Otherwise there exists $\alpha \in \mathbb{R}^{|J_n|}$ such that $X_{J_n}\alpha$, in which case $\alpha^T X_{J_n}^T (I - H_{J_n^c}) X_{J_n}\alpha = 0$. This violates the assumption that $\tilde{\lambda}_- > 0$. On the other hand, it also guarantees that

$$\operatorname{span}(\{X_j : j \in J_n\}) \cap \operatorname{span}\left(\left\{X_j : j \in J_n^c\right\}\right) = \{0\}.$$

This together with assumption **A**1 and Proposition C.3 implies that $\hat{\beta}_{J_n}$ is independent of the choice of $\hat{\beta}$.

Let $B_1 \in \mathbb{R}^{|J_n^c|\times|J_n|}$, $B_2 \in \mathbb{R}^{|J_n^c|\times|J_n^c|}$ and assume that B_2 is invertible. Let $\tilde{X} \in \mathbb{R}^{n\times p}$ such that

$$\tilde{X}_{J_n} = X_{J_n} - X_{J_n^c} B_1, \quad \tilde{X}_{J_n^c} = X_{J_n^c} B_2.$$

Then $\operatorname{rank}(X) = \operatorname{rank}(\tilde{X})$ and model (1) can be rewritten as

$$y = \tilde{X}\tilde{\beta}^* + \varepsilon$$

where

$$\tilde{\beta}^*_{J_n} = \beta^*_{J_n}, \quad \tilde{\beta}^*_{J_n^c} = B_2^{-1}\beta^*_{J_n^c} + B_1\beta^*_{J_n}.$$

Let $\tilde{\hat{\beta}}$ be an M-estimator, which might not be unique, based on $\tilde{X}$. Then Proposition C.3 shows that $\tilde{\hat{\beta}}_{J_n}$ is independent of the choice of $\tilde{\hat{\beta}}$, and an invariance argument shows that

$$\tilde{\hat{\beta}}_{J_n} = \hat{\beta}_{J_n}.$$

In the rest of proof, we use $\tilde{\cdot}$ to denote the quantity obtained based on $\tilde{X}$. First we show that the assumption **A**4 is not affected by this transformation. In fact, for any $j \in J_n$, by definition we have

$$\operatorname{span}(\tilde{X}_{[j]}) = \operatorname{span}(X_{[j]})$$

and hence the leave-j-th-predictor-out residuals are not changed by Proposition C.2. This implies that $\tilde{h}_{j,0} = h_{j,0}$ and $\tilde{Q}_j = Q_j$. Recall the definition of $h_{j,0}$, the first-order condition of $\hat{\beta}$ entails that $X^T h_{j,0} = 0$. In particular, $X_{J_n^c}^T h_{j,0} = 0$ and this implies that for any $\alpha \in \mathbb{R}^n$,

$$0 = \operatorname{Cov}\left(X_{J_n^c}^T h_{j,0}, \alpha^T h_{j,0}\right) = X_{J_n^c} Q_j \alpha.$$

Thus,

$$\frac{\tilde{X}_j^T \tilde{Q}_j \tilde{X}_j}{\mathrm{tr}(\tilde{Q}_j)} = \frac{\left(X_j - X^c_{J_n}(B_1)_j\right)^T Q_j \left(X_j - X_{J_n^c}(B_1)_j\right)}{\mathrm{tr}(Q_j)} = \frac{X_j^T Q_j X_j}{\mathrm{tr}(Q_j)}.$$

Then we prove that the assumption **A**5 is also not affected by the transformation. The above argument has shown that

$$\frac{\tilde{h}_{j,0}^T \tilde{X}_j}{\|\tilde{h}_{j,0}\|_2} = \frac{h_{j,0}^T X_j}{\|h_{j,0}\|_2}.$$

On the other hand, let $B = \begin{pmatrix} I_{|J_n|} & 0 \\ -B_1 & B_2 \end{pmatrix}$, then B is non-singular and $\tilde{X} = XB$. Let $B_{(j),[j]}$ denote the matrix B after removing j-th row and j-th column. Then $B_{(j),[j]}$ is also non-singular and $\tilde{X}_{[j]} = X_{[j]} B_{(j),[j]}$. Recall the definition of $h_{j,1,i}$, we have

$$\begin{aligned}
\tilde{h}_{j,1,i} &= \left(I - \tilde{D}_{[j]} \tilde{X}_{[j]} \left(\tilde{X}_{[j]}^T \tilde{D}_{[j]} \tilde{X}_j\right)^{-1} \tilde{X}_{[j]}^T\right) e_i \\
&= \left(I - D_{[j]} X_{[j]} B_{(j),[j]} \left(B_{(j),[j]}^T X_{[j]}^T D_{[j]} X_j B_{(j),[j]}\right)^{-1} B_{(j),[j]}^T X_{[j]}\right) e_i \\
&= \left(I - D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_j\right)^{-1} X_{[j]}\right) e_i \\
&= h_{j,1,i}.
\end{aligned}$$

On the other hand, by definition,

$$X_{[j]}^T h_{j,1,i} = X_{[j]}^T \left(I - D_{[j]} X_{[j]} \left(X_{[j]}^T D_{[j]} X_{[j]}\right)^{-1} X_{[j]}^T\right) e_i = 0.$$

Thus,

$$h_{j,1,i}^T \tilde{X}_j = h_{j,1,i}^T \left(X_j - X^c_{J_n}(B_1)_j\right) = h_{j,1,i}^T X_j.$$

In summary, for any $j \in J_n$ and $i \le n$,

$$\frac{\tilde{h}_{j,1,i}^T \tilde{X}_j}{\|\tilde{h}_{j,1,i}\|_2} = \frac{h_{j,1,i}^T X_j}{\|h_{j,1,i}\|_2}.$$

Putting the pieces together we have

$$\tilde{\Delta}_C = \Delta_C.$$

By Theorem 3.1,

$$\max_{j\in J_n} d_{\mathrm{TV}}\left(\mathscr{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

provided that $\tilde{X}$ satisfies the assumption **A**3.

Now let $U\Lambda V$ be the singular value decomposition of $X_{J_n^c}$, where $U \in \mathbb{R}^{n\times p}$, $\Lambda \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{p\times p}$ with $U^TU = V^TV = I_p$ and $\Lambda = \mathrm{diag}(\nu_1, \ldots, \nu_p)$ being the diagonal matrix formed by singular values of $X_{J_n^c}$. First we consider the case where $X_{J_n^c}$ has full column rank, then $\nu_j > 0$ for all $j \le p$. Let $B_1 = (X_{J_n}^T X_{J_n})^- X_{J_n}^T X_{J_n}$ and $B_2 = \sqrt{n/|J_n^c|}V^T\Lambda^{-1}$. Then

$$\frac{\tilde{X}^T\tilde{X}}{n} = \frac{1}{n}\begin{pmatrix} X_{J_n}^T\left(I - X_{J_n^c}\left(X_{J_n^c}^T X_{J_n^c}\right)^{-1}X_{J_n^c}\right)X_{J_n} & 0 \\ 0 & nI \end{pmatrix}.$$

This implies that

$$\lambda_{\max}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \max\left\{\tilde{\lambda}_{\max}, 1\right\}, \quad \lambda_{\min}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \min\left\{\tilde{\lambda}_{\min}, 1\right\}.$$

The assumption **A**3* implies that

$$\lambda_{\max}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = O(\mathrm{polyLog(n)}), \quad \lambda_{\min}\left(\frac{\tilde{X}^T\tilde{X}}{n}\right) = \Omega\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

By Theorem 3.1, we conclude that

Next we consider the case where $X_{J_n}^c$ does not have full column rank. We first remove the redundant columns from $X_{J_n}^c$, i.e. replace $X_{J_n^c}$ by the matrix formed by its maximum linear independent subset. Denote by $\mathbf{X}$ this matrix. Then $\mathrm{span}(X) = \mathrm{span}(\mathbf{X})$ and $\mathrm{span}(\{X_j : j \notin J_n\}) = \mathrm{span}(\{\mathbf{X}_j : j \notin J_n\})$. As a consequence of Propositions C.1 and C.3, neither $\beta^*_{J_n}$ nor $\hat{\beta}_{J_n}$ is affected. Thus, the same reasoning as above applies to this case. □

C-3 Proofs of results in Section 3.3

First we prove two lemmas regarding the behavior of Q_j. These lemmas are needed for justifying Assumption **A**4 in the examples.

Lemma C-1 *Under assumptions A1 and A2,*

$$\|Q_j\|_{\mathrm{op}} \le c_1^2 \frac{K_3^2 K_1}{K_0}, \quad \|Q_j\|_{\mathrm{F}} \le \sqrt{n} c_1^2 \frac{K_3^2 K_1}{K_0}$$

where $Q_j = \mathrm{Cov}(h_{j,0})$ *as defined in section B-1.*

Proof (Proof of Lemma C-1) By definition,

$$\|Q_j\|_{\mathrm{op}} = \sup_{\alpha \in \mathbb{S}^{n-1}} \alpha^T Q_j \alpha$$

where $\mathbb{S}^{n-1}$ is the n-dimensional unit sphere. For given $\alpha \in \mathbb{S}^{n-1}$,

$$\alpha^T Q_j \alpha = \alpha^T \mathrm{Cov}(h_{j,0}) \alpha = \mathrm{Var}(\alpha^T h_{j,0})$$

It has been shown in (B-59) in "Appendix B-6.3" that

$$\frac{\partial r_{i,[j]}}{\partial \varepsilon_k} = e_i^T G_{[j]} e_k,$$

where $G_{[j]} = I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}$. This yields that

$$\begin{aligned}\frac{\partial}{\partial \varepsilon^T}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right) &= \sum_{i=1}^n \alpha_i \psi'(r_{i,[j]}) \cdot \frac{\partial r_{i,[j]}}{\partial \varepsilon} \\ &= \sum_{i=1}^n \alpha_i \psi'(r_{i,[j]}) \cdot e_i^T G_{[j]} = \alpha^T \tilde{D}_{[j]} G_{[j]}.\end{aligned}$$

By standard Poincaré inequality (see Proposition B.2), since $\varepsilon_i = u_i(W_i)$,

$$\begin{aligned}&\mathrm{Var}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right) \le \max_k \|u_k'\|_\infty^2 \cdot \mathbb{E}\left\| \frac{\partial}{\partial \varepsilon^T}\left(\sum_{i=1}^n \alpha_i \psi(r_{i,[j]})\right)\right\|^2 \\ &\le c_1^2 \cdot \mathbb{E}\left(\alpha^T \tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]} \alpha\right) \le c_1^2 \mathbb{E}\|\tilde{D}_{[j]} G_{[j]} G_{[j]}^T \tilde{D}_{[j]}\|_2^2 \\ &\le c_1^2 \mathbb{E}\|\tilde{D}_j\|_{\mathrm{op}}^2 \|G_{[j]}\|_{\mathrm{op}}^2.\end{aligned}$$

We conclude from Lemma B-2 and (B-34) in "Appendix B-2" that

$$\|\tilde{D}_{[j]}\|_{\mathrm{op}} \le K_3, \quad \|G_{[j]}\|_{\mathrm{op}}^2 \le \frac{K_1}{K_0}.$$

Therefore,

$$\|Q_j\|_{\mathrm{op}} = \sup_{\alpha \in \mathbb{S}^{n-1}} \mathrm{Var}\left(\sum_{i=1}^n \alpha_i \psi(R_i)\right) \le c_1^2 \frac{K_3^2 K_1}{K_0}$$

and hence

$$\|Q_j\|_{\mathrm{F}} \le \sqrt{n}\|Q_j\|_{\mathrm{op}} \le \sqrt{n}\cdot c_1^2\frac{K_3^2K_1}{K_0}.$$

□

Lemma C-2 *Under assumptions A1 - A3,*

$$\mathrm{tr}(Q_j) \ge K^* n = \Omega(n\cdot \mathrm{polyLog(n)}),$$

where $K^* = \frac{K_0^4}{K_1^2}\cdot\left(\frac{n-p+1}{n}\right)^2\cdot \min_i \mathrm{Var}(\varepsilon_i)$.

Proof This is a direct consequence of Lemma B-4 in p. 49. □

Throughout the following proofs, we will use several results from the random matrix theory to bound the largest and smallest singular values of Z. The results are shown in "Appendix E". Furthermore, in contrast to other sections, the notation $P(\cdot), \mathbb{E}(\cdot), \mathrm{Var}(\cdot)$ *denotes the probability, the expectation and the variance with respect to both* ε *and* Z *in this section.*

Proof (Proof of Proposition 3.1) By Proposition E.3,

$$\lambda_+ = (1+\sqrt{\kappa})^2 + o_p(1) = O_p(1), \quad \lambda_- = (1-\sqrt{\kappa})^2 - o_p(1) = \Omega_p(1)$$

and thus the assumption **A**3 holds with high probability. By Hanson-Wright inequality ([27,51]; see Proposition E.2), for any given deterministic matrix A,

$$P\left(\left|Z_j^TAZ_j - \mathbb{E}Z_j^TAZ_j\right| \ge t\right) \le 2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4\|A\|_{\mathrm{F}}^2}, \frac{t}{\sigma^2\|A\|_{\mathrm{op}}}\right\}\right]$$

for some universal constant c. Let $A = Q_j$ and conditioning on $Z_{[j]}$, then by Lemma C-1, we know that

$$\|Q_j\|_{\mathrm{op}} \le c_1^2\frac{K_3^2K_1}{K_0}, \quad \|Q_j\|_{\mathrm{F}} \le \sqrt{n}c_1^2\frac{K_3^2K_1}{K_0}$$

and hence

$$\begin{aligned} &P\left(Z_j^TQ_jZ_j - \mathbb{E}(Z_j^TQ_jZ_j|Z_{[j]}) \le -t\middle|Z_{[j]}\right) \\ &\quad\le 2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4\cdot nc_1^4K_3^4K_1^2/K_0^2}, \frac{t}{\sigma^2c_1^2K_3^2K_1/K_0}\right\}\right]. \end{aligned} \tag{C-69}$$

Note that

$$\mathbb{E}\left(Z_j^TQ_jZ_j|Z_{[j]}\right) = \mathrm{tr}\left(\mathbb{E}\left[Z_jZ_j^T|Z_{[j]}\right]Q_j\right) = \mathbb{E}Z_{1j}^2\,\mathrm{tr}(Q_j) = \tau^2\,\mathrm{tr}(Q_j).$$

By Lemma C-2, we conclude that

$$P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \le \tau^2 - \frac{t}{nK^*}\middle| Z_{[j]}\right) \le P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \le \tau^2 - \frac{t}{\mathrm{tr}(Q_j)}\middle| Z_{[j]}\right) \le 2\exp\left[-c\min\left\{\frac{t^2}{\sigma^4 \cdot nc_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{2\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right]. \tag{C-70}$$

Let $t = \frac{1}{2}\tau^2 nK^*$ and take expectation of both sides over $Z_{[j]}$, we obtain that

$$P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \le \frac{\tau^2}{2}\right) \le 2\exp\left[-cn\min\left\{\frac{K^{*2}\tau^4}{4\sigma^4 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{K^*\tau^2}{2\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right]$$

and hence

$$P\left(\min_{j\in J_n}\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \le \frac{\tau^2}{2}\right) \le 2n\exp\left[-cn\min\left\{\frac{K^{*2}\tau^4}{4\sigma^4 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{K^*\tau^2}{2\sigma^2 c_1^2 K_3^2 K_1/K_0}\right\}\right] = o(1). \tag{C-71}$$

This entails that

$$\min_{j\in J_n}\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

Thus, assumption **A**4 is also satisfied with high probability. On the other hand, since Z_j has i.i.d. mean-zero σ^2-sub-gaussian entries, for any deterministic unit vector $\alpha \in \mathbb{R}^n$, $\alpha^T Z_j$ is σ^2-sub-gaussian and mean-zero, and hence

$$P(|\alpha^T Z_j| \ge t) \le 2e^{-\frac{t^2}{2\sigma^2}}.$$

Let $\alpha_{j,i} = h_{j,1,i}/\|h_{j,1,i}\|_2$ and $\alpha_{j,0} = h_{j,0}/\|h_{j,0}\|_2$. Since $h_{j,1,i}$ and $h_{j,0}$ are independent of Z_j, a union bound then gives

$$P\left(\Delta_C \ge t + 2\sigma\sqrt{\log n}\right) \le 2n^2 e^{-\frac{t^2+4\sigma^2\log n}{2\sigma^2}} = 2e^{-\frac{t^2}{2\sigma^2}}.$$

By Fubini's formula ([16], Lemma 2.2.8.),

$$\mathbb{E}\Delta_C^8 = \int_0^\infty 8t^7 P(\Delta_C \ge t)dt \le \int_0^{2\sigma\sqrt{\log n}} 8t^7 dt + \int_{2\sigma\sqrt{\log n}}^\infty 8t^7 P(\Delta_C \ge t)dt$$

$$
\begin{aligned}
&= (2\sigma\sqrt{\log n})^8 + \int_0^\infty 8(t + 2\sigma\sqrt{\log n})^7 P(\Delta_C \geq t + 2\sigma\sqrt{\log n}) dt \\
&\leq (2\sigma\sqrt{\log n})^8 + \int_0^\infty 64(8t^7 + 128\sigma^7 (\log n)^{\frac{7}{2}}) \cdot 2e^{-\frac{t^2}{2\sigma^2}} dt \\
&= O(\sigma^8 \cdot \text{polyLog(n)}) = O\left(\text{polyLog(n)}\right). \qquad \text{(C-72)}
\end{aligned}
$$

This, together with Markov inequality, guarantees that assumption **A**5 is also satisfied with high probability. □

Proof (Proof of Proposition 3.2) It is left to prove that assumption **A**3 holds with high probability. The proof of assumption **A**4 and **A**5 is exactly the same as the proof of Proposition 3.2. By Proposition E.4,

$$\lambda_+ = O_p(1).$$

On the other hand, by Proposition E.7 [37],

$$P\left(\lambda_{\min}\left(\frac{Z^T Z}{n}\right) < c_1\right) \leq e^{-c_2 n}.$$

and thus □

$$\lambda_- = \Omega_p(1).$$

Proof (Proof of Proposition 3.3) Since J_n excludes the intercept term, the proof of assumption **A**4 and **A**5 is still the same as Proposition 3.2. It is left to prove assumption **A**3. Let $R_1, \ldots, R_n$ be i.i.d. Rademacher random variables, i.e. $P(R_i = 1) = P(R_i = -1) = \frac{1}{2}$, and

$$Z^* = \text{diag}(B_1, \ldots, B_n) Z.$$

Then $(Z^*)^T Z^* = Z^T Z$. It is left to show that the assumption **A**3 holds for Z^* with high probability. Note that

$$(Z_i^*)^T = \left(B_i, B_i \tilde{x}_i^T\right).$$

For any $r \in \{1, -1\}$ and borel sets $B_1, \ldots, B_p \subset \mathbb{R}$,

$$
\begin{aligned}
&P(B_i = r, B_i \tilde{Z}_{i1} \in B_1, \ldots, B_i \tilde{Z}_{i(p-1)} \in B_{p-1}) \\
&= P(B_i = r, \tilde{Z}_{i1} \in r B_1, \ldots, \tilde{Z}_{i(p-1)} \in r B_{p-1}) \\
&= P(B_i = r) P(\tilde{Z}_{i1} \in r B_1) \ldots P(\tilde{Z}_{i(p-1)} \in r B_{p-1}) \\
&= P(B_i = r) P(\tilde{Z}_{i1} \in B_1) \ldots P(\tilde{Z}_{i(p-1)} \in B_{p-1}) \\
&= P(B_i = r) P(B_i \tilde{Z}_{i1} \in B_1) \ldots P(B_i \tilde{Z}_{i(p-1)} \in B_{p-1})
\end{aligned}
$$

where the last two lines uses the symmetry of $\tilde{Z}_{ij}$. Then we conclude that Z_i^* has independent entries. Since the rows of Z^* are independent, Z^* has independent entries. Since B_i are symmetric and sub-gaussian with unit variance and $B_i \tilde{Z}_{ij} \stackrel{d}{=} \tilde{Z}_{ij}$, which is also symmetric and sub-gaussian with variance bounded from below, Z^* satisfies the conditions of Propsition 3.2 and hence the assumption **A**3 is satisfied with high probability. □

Proof (Proof of Proposition 3.5 (with Proposition 3.4 being a special case)) Let $Z_* = \Lambda^{-\frac{1}{2}} Z \Sigma^{-\frac{1}{2}}$, then Z_* has i.i.d. standard gaussian entries. By Proposition 3.3, Z_* satisfies assumption **A**3 with high probability. Thus,

$$\begin{aligned}\lambda_+ &= \lambda_{\max}\left(\frac{\Sigma^{\frac{1}{2}} Z_*^T \Lambda Z_* \Sigma^{\frac{1}{2}}}{n}\right) \le \lambda_{\max}(\Sigma)\cdot\lambda_{\max}(\Lambda)\cdot\lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right)\\ &= O_p(\text{polyLog(n)}),\end{aligned}$$

and

$$\begin{aligned}\lambda_- &= \lambda_{\min}\left(\frac{\Sigma^{\frac{1}{2}} Z_*^T \Lambda Z_* \Sigma^{\frac{1}{2}}}{n}\right) \ge \lambda_{\min}(\Sigma)\cdot\lambda_{\min}(\Lambda)\cdot\lambda_{\min}\left(\frac{Z_*^T Z_*}{n}\right)\\ &= \Omega_p\left(\frac{1}{\text{polyLog(n)}}\right).\end{aligned}$$

As for assumption **A**4, the first step is to calculate $\mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]})$. Let $\tilde{Z} = \Lambda^{-\frac{1}{2}} Z$, then $\text{vec}(\tilde{Z}) \sim N(0, I \otimes \Sigma)$. As a consequence,

$$\tilde{Z}_j | \tilde{Z}_{[j]} \sim N\left(\tilde{\mu}_j, \sigma_j^2 I\right)$$

where

$$\tilde{\mu}_j = \tilde{Z}_{[j]} \Sigma_{[j],[j]}^{-1} \Sigma_{[j],j} = \Lambda^{-\frac{1}{2}} Z_{[j]} \Sigma_{[j],[j]}^{-1} \Sigma_{[j],j}.$$

Thus,

$$Z_j | Z_{[j]} \sim N\left(\mu_j, \sigma_j^2 \Lambda\right)$$

where $\mu_j = Z_{[j]} \Sigma_{[j],[j]}^{-1} \Sigma_{[j],j}$. It is easy to see that

$$\lambda_- \le \min_j \sigma_j^2 \le \max_j \sigma_j^2 \le \lambda_+. \tag{C-73}$$

It has been shown that $Q_j \mu_j = 0$ and hence

$$Z_j^T Q_j Z_j = (Z_j - \mu_j)^T Q_j (Z_j - \mu_j).$$

Let $\mathscr{Z}_j = \Lambda^{-\frac{1}{2}}(Z_j - \mu_j)$ and $\tilde{Q}_j = \Lambda^{\frac{1}{2}} Q_j \Lambda^{\frac{1}{2}}$, then $\mathscr{Z}_j \sim N(0, \sigma_j^2 I)$ and

$$Z_j^T Q_j Z_j = \mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j.$$

By Lemma C-1,

$$\|\tilde{Q}_j\|_{\text{op}} \le \|\Lambda\|_{\text{op}} \cdot \|Q_j\|_{\text{op}} \le \lambda_{\max}(\Lambda) \cdot c_1^2 \frac{K_3^2 K_1}{K_0},$$

and hence

$$\|\tilde{Q}_j\|_{\text{F}} \le \sqrt{n}\lambda_{\max}(\Lambda) \cdot c_1^2 \frac{K_3^2 K_1}{K_0}.$$

By Hanson-Wright inequality ([27,51]; see Proposition E.2), we obtain a similar inequality to (C-69) as follows:

$$P\left(|Z_j^T Q_j Z_j - \mathbb{E}(Z_j^T Q_j Z_j | Z_{[j]})| \ge t \middle| Z_{[j]}\right) \le 2\exp\left[-c\min\left\{\frac{t^2}{\sigma_j^4 \cdot n\lambda_{\max}(\Lambda)^2 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{\sigma_j^2 \lambda_{\max}(\Lambda) c_1^2 K_3^2 K_1/K_0}\right\}\right].$$

On the other hand,

$$\mathbb{E}\left(Z_j^T Q_j Z_j | Z_{[j]}\right) = \mathbb{E}\left(\mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j | Z_{[j]}\right) = \sigma_j^2 \operatorname{tr}(\tilde{Q}_j).$$

By definition,

$$\operatorname{tr}(\tilde{Q}_j) = \operatorname{tr}(\Lambda^{\frac{1}{2}} Q_j \Lambda^{\frac{1}{2}}) = \operatorname{tr}(\Sigma Q_j) = \operatorname{tr}\left(Q_j^{\frac{1}{2}} \Lambda Q_j^{\frac{1}{2}}\right) \ge \lambda_{\min}(\Lambda) \operatorname{tr}(Q_j).$$

By Lemma C-2,

$$\operatorname{tr}(\tilde{Q}_j) \ge \lambda_{\min}(\Lambda) \cdot nK^*.$$

Similar to (C-70), we obtain that

$$P\left(\frac{Z_j^T Q_j Z_j}{\operatorname{tr}(Q_j)} \ge \sigma_j^2 - \frac{t}{nK^*} \middle| Z_{[j]}\right) \le 2\exp\left[-c\min\left\{\frac{t^2}{\sigma_j^4 \cdot n\lambda_{\max}(\Lambda)^2 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{t}{\sigma_j^2 \lambda_{\max}(\Lambda) c_1^2 K_3^2 K_1/K_0}\right\}\right].$$

Let $t = \frac{1}{2}\sigma_j^2 n K^*$, we have

$$\begin{aligned} & P\left(\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} \geq \frac{\sigma_j^2}{2}\right) \\ & \leq 2\exp\left[-cn\min\left\{\frac{K^{*2}}{4\lambda_{\max}(\Lambda)^2 c_1^4 K_3^4 K_1^2/K_0^2}, \frac{K^*}{2\lambda_{\max}(\Lambda) c_1^2 K_3^2 K_1/K_0}\right\}\right] \\ & = o\left(\frac{1}{n}\right) \end{aligned}$$

and a union bound together with (C-73) yields that

$$\min_{j\in J_n} \frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} = \Omega_p\left(\min_j \sigma_j^2 \cdot \frac{1}{\mathrm{polyLog(n)}}\right) = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

As for assumption **A**5, let

$$\alpha_{j,0} = \frac{\Lambda^{\frac{1}{2}} h_{j,0}}{\|h_{j,0}\|_2}, \quad \alpha_{j,i} = \frac{\Lambda^{\frac{1}{2}} h_{j,1,i}}{\|h_{j,1,i}\|_2}$$

then for $i = 0, 1, \ldots, p$,

$$\|\alpha_{j,i}\|_2 \leq \sqrt{\lambda_{\max}(\Lambda)}.$$

Note that

$$\frac{h_{j,0}^T Z_j}{\|h_{j,0}\|_2} = \alpha_{j,0}^T Z_j, \quad \frac{h_{j,1,i}^T Z_j}{\|h_{j,1,i}\|_2} = \alpha_{j,i}^T Z_j$$

using the same argument as in (C-72), we obtain that

$$\mathbb{E}\Delta_C^8 = O\left(\lambda_{\max}(\Lambda)^4 \cdot \max_j \sigma_j^8 \cdot \mathrm{polyLog(n)}\right) = O\left(\mathrm{polyLog(n)}\right),$$

and by Markov inequality and (C-73),

$$\mathbb{E}\left(\Delta_C^8 | Z\right) = O_p\left(\mathbb{E}\Delta_C^8\right) = O_p(\mathrm{polyLog(n)}).$$

□

Proof (Proof of Proposition 3.6) The proof that assumptions **A**4 and **A**5 hold with high probability is exactly the same as the proof of Proposition 3.5. It is left to prove

assumption **A**3*; see Corollary 3.1. Let $c = (\min_i |(\Lambda^{-\frac{1}{2}}\mathbf{1})_i|)^{-1}$ and $\mathbf{Z} = (c\mathbf{1}\ \tilde{Z})$. Recall the the definition of $\tilde{\lambda}_+$ and $\tilde{\lambda}_-$, we have

$$\tilde{\lambda}_+ = \lambda_{\max}(\Sigma_{\{1\}}), \quad \tilde{\lambda}_- = \lambda_{\min}(\Sigma_{\{1\}}),$$

where

$$\Sigma_{\{1\}} = \frac{1}{n}\tilde{Z}^T\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}.$$

Rewrite $\Sigma_{\{1\}}$ as

$$\Sigma_{\{1\}} = \frac{1}{n}\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right)^T\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right).$$

It is obvious that

$$\text{span}\left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\tilde{Z}\right) \subset \text{span}(\mathbf{Z}).$$

As a consequence

$$\tilde{\lambda}_+ \le \lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right), \quad \tilde{\lambda}_- \ge \lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right).$$

It remains to prove that

$$\lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = O_p\left(\text{polyLog(n)}\right), \quad \lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \Omega_p\left(\frac{1}{\text{polyLog(n)}}\right).$$

To prove this, we let

$$Z_* = \Lambda^{-\frac{1}{2}}\mathbf{Z}\begin{pmatrix}1 & 0\\ 0 & \Sigma^{-\frac{1}{2}}\end{pmatrix} \triangleq (\nu\ \tilde{Z}_*),$$

where $\nu = c\Lambda^{-\frac{1}{2}}\mathbf{1}$ and $\tilde{Z}_* = \Lambda^{-\frac{1}{2}}\tilde{Z}\Sigma^{-\frac{1}{2}}$. Then

$$\lambda_{\max}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \lambda_{\max}\left(\frac{\Sigma^{\frac{1}{2}}Z_*^T\Lambda Z_*\Sigma^{\frac{1}{2}}}{n}\right) \le \lambda_{\max}(\Sigma)\cdot\lambda_{\max}(\Lambda)\cdot\lambda_{\max}\left(\frac{Z_*^TZ_*}{n}\right),$$

and

$$\lambda_{\min}\left(\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right) = \lambda_{\min}\left(\frac{\Sigma^{\frac{1}{2}}Z_*^T\Lambda Z_*\Sigma^{\frac{1}{2}}}{n}\right) \ge \lambda_{\min}(\Sigma)\cdot\lambda_{\min}(\Lambda)\cdot\lambda_{\min}\left(\frac{Z_*^TZ_*}{n}\right).$$

It is left to show that

$$\lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right) = O_p(\text{polyLog(n)}), \quad \lambda_{\min}\left(\frac{Z_*^T Z_*}{n}\right) = \Omega_p\left(\frac{1}{\text{polyLog(n)}}\right).$$

By definition, $\min_i |\nu_i| = 1$ and $\max_i |\nu_i| = O(\text{polyLog(n)})$, then

$$\lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right) = \lambda_{\max}\left(\frac{\tilde{Z}_*^T \tilde{Z}_*}{n} + \frac{\nu\nu^T}{n}\right) \le \lambda_{\max}\left(\frac{\tilde{Z}_*^T \tilde{Z}_*}{n}\right) + \frac{\|\nu\|_2^2}{n}.$$

Since $\tilde{Z}_*$ has i.i.d. standard gaussian entries, by Proposition E.3,

$$\lambda_{\max}\left(\frac{\tilde{Z}_*^T \tilde{Z}_*}{n}\right) = O_p(1).$$

Moreover, $\|\nu\|_2^2 \le n \max_i |\nu_i|^2 = O(n \cdot \text{polyLog(n)})$ and thus,

$$\lambda_{\max}\left(\frac{Z_*^T Z_*}{n}\right) = O_p(\text{polyLog(n)}).$$

On the other hand, similar to Proposition 3.3,

$$\mathbf{Z}_* = \text{diag}(B_1, \ldots, B_n) Z_*$$

where $B_1, \ldots, B_n$ are i.i.d. Rademacher random variables. The same argument in the proof of Proposition 3.3 implies that $\mathbf{Z}_*$ has independent entries with sub-gaussian norm bounded by $\|\nu\|_\infty^2 \vee 1$ and variance lower bounded by 1. By Proposition E.7, Z_* satisfies assumption **A**3 with high probability. Therefore, **A**3* holds with high probability. □

Proof (Proof of Proposition 3.7) Let $\Lambda = (\lambda_1, \ldots, \lambda_n)$ and $\mathscr{Z}$ be the matrix with entries $\mathscr{Z}_{ij}$, then by Proposition 3.1 or Proposition 3.2, $\mathscr{Z}_{ij}$ satisfies assumption **A**3 with high probability. Notice that

$$\lambda_+ = \lambda_{\max}\left(\frac{\mathscr{Z}^T \Lambda^2 \mathscr{Z}}{n}\right) \le \lambda_{\max}(\Lambda)^2 \cdot \lambda_{\max}\left(\frac{\mathscr{Z}^T \mathscr{Z}}{n}\right) = O_p(\text{polyLog(n)}),$$

and

$$\lambda_- = \lambda_{\min}\left(\frac{\mathscr{Z}^T \Lambda^2 \mathscr{Z}}{n}\right) \ge \lambda_{\min}(\Lambda)^2 \cdot \lambda_{\min}\left(\frac{\mathscr{Z}^T \mathscr{Z}}{n}\right) = \Omega_p\left(\frac{1}{\text{polyLog(n)}}\right).$$

Thus Z satisfies assumption **A**3 with high probability.

Conditioning on any realization of Λ, the law of $\mathscr{Z}_{ij}$ does not change due to the

independence between Λ and $\mathscr{Z}$. Repeating the arguments in the proof of Proposition 3.1 and Proposition 3.2, ow that

$$\frac{\mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j}{\mathrm{tr}(\tilde{Q}_j)} = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right), \quad \text{and}$$
$$\mathbb{E} \max_{i=0,\ldots,n;\, j=1,\ldots,p} \left|\tilde{\alpha}_{j,i}^T \mathscr{Z}_j\right|^8 = O_p(\mathrm{polyLog(n)}), \tag{C-74}$$

where

$$\tilde{Q}_j = \Lambda Q_j \Lambda, \quad \tilde{\alpha}_{j,0} = \frac{\Lambda h_{j,0}}{\|\Lambda h_{j,0}\|_2}, \quad \tilde{\alpha}_{j,1,i} = \frac{\Lambda h_{j,1,i}}{\|\Lambda h_{j,1,i}\|_2}.$$

Then

$$\frac{Z_j^T Q_j Z_j}{\mathrm{tr}(Q_j)} = \frac{\mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j}{\mathrm{tr}(\tilde{Q}_j)} \cdot \frac{\mathrm{tr}(\Lambda Q_j \Lambda)}{\mathrm{tr}(Q_j)} \geq a^2 \cdot \frac{\mathscr{Z}_j^T \tilde{Q}_j \mathscr{Z}_j}{\mathrm{tr}(\tilde{Q}_j)} = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right), \tag{C-75}$$

and

$$\begin{aligned}
\mathbb{E}\Delta_C^8 &= \mathbb{E}\left[\max_{i=0,\ldots,n;\, j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathscr{Z}_j|^8 \cdot \max\left\{\max_j \frac{\|\Lambda h_{j,0}\|_2}{\|h_{j,0}\|_2}, \max_{i,j} \frac{\|\Lambda h_{j,1,i}\|_2}{\|h_{j,1,i}\|_2}\right\}^8\right] \\
&\leq b^8 \mathbb{E}\left[\max_{i=0,\ldots,n;\, j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathscr{Z}_j|^8\right] \\
&= O_p(\mathrm{polyLog(n)}).
\end{aligned} \tag{C-76}$$

By Markov inequality, the assumption **A5** is satisfied with high probability. □

Proof (Proof of Proposition 3.8) The concentration inequality of ζ_i plus a union bound imply that

$$P\left(\max_i \zeta_i > (\log n)^{\frac{2}{\alpha}}\right) \leq n c_1 e^{-c_2 (\log n)^2} = o(1).$$

Thus, with high probability,

$$\lambda_{\max} = \lambda_{\max}\left(\frac{\mathscr{Z}^T \Lambda^2 \mathscr{Z}}{n}\right) \leq (\log n)^{\frac{4}{\alpha}} \cdot \lambda_{\max}\left(\frac{\mathscr{Z}^T \mathscr{Z}}{n}\right) = O_p(\mathrm{polyLog(n)}).$$

Let $n' = \lfloor (1-\delta)n \rfloor$ for some $\delta \in (0, 1-\kappa)$. Then for any subset I of $\{1, \ldots, n\}$ with size n', by Proposition E.6 (Proposition E.7), under the conditions of Proposition 3.1 (Proposition 3.2), there exists constants c_3 and c_4, which only depend on κ, such that

$$P\left(\lambda_{\min}\left(\frac{\mathscr{Z}_I^T \mathscr{Z}_I}{n}\right) < c_3\right) \leq e^{-c_4 n}$$

where $\mathscr{Z}_I$ represents the sub-matrix of $\mathscr{Z}$ formed by $\{\mathscr{Z}_i : i \in I\}$, where $\mathscr{Z}_i$ is the i-th row of $\mathscr{Z}$. Then by a union bound,

$$P\left(\min_{|I|=n'} \lambda_{\min}\left(\frac{\mathscr{Z}_I^T \mathscr{Z}_I}{n}\right) < c_3\right) \le \binom{n}{n'} e^{-c_4 n}.$$

By Stirling's formula, there exists a constant $c_5 > 0$ such that

$$\binom{n}{n'} = \frac{n!}{n'!(n-n')!} \le c_5 \exp\left\{(-\tilde{\delta}\log\tilde{\delta} - (1-\tilde{\delta})\log(1-\tilde{\delta}))n\right\}$$

where $\tilde{\delta} = n'/n$. For sufficiently small δ and sufficiently large n,

$$-\tilde{\delta}\log\tilde{\delta} - (1-\tilde{\delta})\log(1-\tilde{\delta}) < c_4$$

and hence

$$P\left(\min_{|I|=n'} \lambda_{\min}\left(\frac{\mathscr{Z}_I^T \mathscr{Z}_I}{n}\right) < c_3\right) < c_5 e^{-c_6 n} \tag{C-77}$$

for some $c_6 > 0$. By Borel–Cantelli Lemma,

$$\liminf_{n\to\infty} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{\mathscr{Z}_I^T \mathscr{Z}_I}{n}\right) \ge c_3 \quad a.s..$$

On the other hand, since F^{-1} is continuous at δ, then

$$\zeta_{(\lfloor(1-\delta)n\rfloor)} \xrightarrow{a.s.} F^{-1}(\delta) > 0.$$

where $\zeta_{(k)}$ is the k-th largest of $\{\zeta_i : i = 1, \ldots, n\}$. Let I^* be the set of indices corresponding to the largest $\lfloor(1-\delta)n\rfloor$ ζ_i's. Then with probability 1,

$$\begin{aligned}
\liminf_{n\to\infty} \lambda_{\min}\left(\frac{Z^T Z}{n}\right) &= \liminf_{n\to\infty} \lambda_{\min}\left(\frac{\mathscr{Z}^T \Lambda^2 \mathscr{Z}}{n}\right) \\
&\ge \liminf_{n\to\infty} \zeta_{(\lfloor(1-\delta)n\rfloor)} \cdot \liminf_{n\to\infty} \lambda_{\min}\left(\frac{\mathscr{Z}_{I^*}^T \Lambda_{I^*}^2 \mathscr{Z}_{I^*}}{n}\right) \\
&\ge \liminf_{n\to\infty} \zeta_{(\lfloor(1-\delta)n\rfloor)} \cdot \liminf_{n\to\infty} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{\mathscr{Z}_I^T \mathscr{Z}_I}{n}\right) \\
&\ge c_3 F^{-1}(\delta)^2 > 0.
\end{aligned}$$

To prove assumption **A**4, similar to (C-75) in the proof of Proposition 3.7, it is left to show that

$$\min_j \frac{\mathrm{tr}(\Lambda Q_j \Lambda)}{\mathrm{tr}(Q_j)} = \Omega_p\left(\frac{1}{\mathrm{polyLog(n)}}\right).$$

Furthermore, by Lemma C-2, it remains to prove that

$$\min_j \operatorname{tr}(\Lambda Q_j \Lambda) = \Omega_p\left(\frac{n}{\text{polyLog(n)}}\right).$$

Recalling the Eq. (B-60) in the proof of Lemma B-4, we have

$$e_i^T Q_j e_i \geq \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T Z_{[j]}^T \left(Z_{(i),[j]}^T Z_{(i),[j]}\right)^{-1} Z_{[j]} e_i}. \tag{C-78}$$

By Proposition E.5,

$$P\left(\sqrt{\lambda_{\max}\left(\frac{\mathscr{Z}_j^T \mathscr{Z}_j}{n}\right)} > 3C_1\right) \leq 2e^{-C_2 n}.$$

On the other hand, apply (C-77) to $\mathscr{Z}_{(i),[j]}$, we have

$$P\left(\min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathscr{Z}_{(i),[j]})_I^T (\mathscr{Z}_{(i),[j]})_I}{n}\right) < c_3\right) < c_5 e^{-c_6 n}.$$

A union bound indicates that with probability $(c_5 np + 2p)e^{-\min\{C_2,c_6\}n} = o(1)$,

$$\begin{aligned}&\max_j \lambda_{\max}\left(\frac{\mathscr{Z}_{[j]}^T \mathscr{Z}_{[j]}}{n}\right)\\ &\leq 9C_1^2, \quad \min_{i,j} \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathscr{Z}_{(i),[j]})_I^T (\mathscr{Z}_{(i),[j]})_I}{n}\right) \geq c_3.\end{aligned}$$

This implies that for any j,

$$\lambda_{\max}\left(\frac{Z_{[j]}^T Z_{[j]}}{n}\right) = \lambda_{\max}\left(\frac{\mathscr{Z}_{[j]}^T \Lambda^2 \mathscr{Z}_{[j]}}{n}\right) \leq \zeta_{(1)}^2 \cdot 9C_1^2$$

and for any i and j,

$$\begin{aligned}&\lambda_{\min}\left(\frac{Z_{(i),[j]}^T Z_{(i),[j]}}{n}\right) = \lambda_{\min}\left(\frac{\mathscr{Z}_{(i),[j]}^T \zeta_{(i)}^2 \mathscr{Z}_{(i),[j]}}{n}\right)\\ &\geq \min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)} + 1\}^2 \cdot \min_{|I|=\lfloor(1-\delta)n\rfloor} \lambda_{\min}\left(\frac{(\mathscr{Z}_{(i),[j]})_I^T \zeta_{(i)}^2 (\mathscr{Z}_{(i),[j]})_I}{n}\right)\\ &\geq c_3 \min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)} + 1\}^2 > 0.\end{aligned}$$

Moreover, as discussed above,

$$\zeta_{(1)} \le (\log n)^{\frac{2}{\alpha}}, \min\{\zeta_{(\lfloor(1-\delta)n\rfloor)}, \zeta_{(\lfloor(1-\delta)n\rfloor)} + 1\} \to F^{-1}(\delta)$$

almost surely. Thus, it follows from (C-78) that with high probability,

$$\begin{aligned} e_i^T Q_j e_i &\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T Z_{[j]}^T \left(Z_{(i),[j]}^T Z_{(i),[j]}\right)^{-1} Z_{[j]} e_i} \\ &\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + e_i^T \frac{Z_{[j]}^T Z_{[j]}}{n} e_i \cdot c_3(F^{-1}(\delta))^2} \\ &\ge \frac{K_0}{K_1} \cdot \frac{1}{1 + (\log n)^{\frac{4}{\alpha}} \cdot 9C_1^2 \cdot c_3(F^{-1}(\delta))^2}. \end{aligned}$$

The above bound holds for all diagonal elements of Q_j uniformly with high probability. Therefore,

$$\begin{aligned} \mathrm{tr}(\Lambda Q_j \Lambda) &\ge \zeta_{(\lfloor(1-\delta)n\rfloor)}^2 \cdot \lfloor(1-\delta)n\rfloor \cdot \frac{K_0}{K_1} \cdot \frac{1}{1 + (\log n)^{\frac{4}{\alpha}} \cdot 9C_1^2 \cdot c_3(F^{-1}(\delta))^2} \\ &= \Omega_p\left(\frac{n}{\mathrm{polyLog(n)}}\right). \end{aligned}$$

As a result, the assumption **A**4 is satisfied with high probability. Finally, by (C-76), we obtain that

$$\mathbb{E}\Delta_C^8 \le \mathbb{E}\left[\max_{i=0,\ldots,n;\, j=1,\ldots,p} \left|\tilde{\alpha}_{j,i}^T \mathscr{Z}_j\right|^8 \cdot \|\Lambda\|_{\mathrm{op}}^8\right].$$

By Cauchy's inequality,

$$\mathbb{E}\Delta_C^8 \le \sqrt{\mathbb{E}\max_{i=0,\ldots,n;\, j=1,\ldots,p} |\tilde{\alpha}_{j,i}^T \mathscr{Z}_j|^{16}} \cdot \sqrt{\mathbb{E}\max_i \zeta_i^{16}}.$$

Similar to (C-72), we conclude that

$$\mathbb{E}\Delta_C^8 = O\left(\mathrm{polyLog(n)}\right)$$

and by Markov inequality, the assumption **A**5 is satisfied with high probability. □

C-4 More results of least-squares (Section 5)

C-4.1 The relation between $S_j(X)$ and Δ_C

In Sect. 5, we give a sufficient and almost necessary condition for the coordinate-wise asymptotic normality of the least-square estimator $\hat{\beta}^{LS}$; see Theorem 5.1. In this

subsubsection, we show that Δ_C is a generalization of $\max_{j\in J_n} S_j(X)$ for general M-estimators.

Consider the matrix $(X^T DX)^{-1}X^T$, where D is obtain by using general loss functions, then by block matrix inversion formula (see Proposition E.1),

$$\begin{aligned} e_1^T(X^T DX)^{-1}X^T &= e_1^T \begin{pmatrix} X_1^T DX_1 & X_1^T DX_{[1]} \\ X_{[1]}^T DX_1 & X_{[1]}^T DX_{[1]} \end{pmatrix}^{-1} \begin{pmatrix} X_1^T \\ X_{[1]}^T \end{pmatrix} \\ &= \frac{X_1^T \left(I - DX_{[1]} \left(X_{[1]}^T DX_{[1]} \right)^{-1} X_{[1]}^T \right)}{X_1^T \left(D - DX_{[1]} \left(X_{[1]}^T DX_{[1]} \right)^{-1} X_{[1]}^T D \right) X_1} \\ &\approx \frac{X_1^T \left(I - D_{[1]}X_{[1]} \left(X_{[1]}^T D_{[1]}X_{[1]} \right)^{-1} X_{[1]}^T \right)}{X_1^T \left(D - DX_{[1]} \left(X_{[1]}^T DX_{[1]} \right)^{-1} X_{[1]}^T D \right) X_1} \end{aligned}$$

where we use the approximation $D \approx D_{[1]}$. The same result holds for all $j \in J_n$, then

$$\frac{\|e_j^T(X^T DX)^{-1}X^T\|_\infty}{\|e_j^T(X^T DX)^{-1}X^T\|_2} \approx \frac{\left\| X_1^T \left(I - D_{[1]}X_{[1]} \left(X_{[1]}^T D_{[1]}X_{[1]} \right)^{-1} X_{[1]}^T \right) \right\|_\infty}{\left\| X_1^T \left(I - D_{[1]}X_{[1]} \left(X_{[1]}^T D_{[1]}X_{[1]} \right)^{-1} X_{[1]}^T \right) \right\|_2}.$$

Recall that $h_{j,1,i}^T$ is i-th row of $I - D_{[1]}X_{[1]}(X_{[1]}^T D_{[1]}X_{[1]})^{-1}X_{[1]}^T$, we have

$$\max_i \frac{\left| h_{j,1,i}^T X_1 \right|}{\|h_{j,1,i}\|_2} \approx \frac{\left\| e_j^T(X^T DX)^{-1}X^T \right\|_\infty}{\left\| e_j^T(X^T DX)^{-1}X^T \right\|_2}.$$

The right-handed side equals to $S_j(X)$ in the least-square case. Therefore, although of complicated form, assumption **A5** is not an artifact of the proof but is essential for the asymptotic normality.

C-4.2 Additional examples

Benefit from the analytical form of the least-square estimator, we can depart from subgaussinity of the entries. The following proposition shows that a random design matrix Z with i.i.d. entries under appropriate moment conditions satisfies $\max_{j\in J_n} S_j(Z) = o(1)$ with high probability. This implies that, when X is one realization of Z, the conditions Theorem 5.1 are satisfied for X with high probability over Z.

Proposition C.4 *If* $\{Z_{ij} : i \le n,\ j \in J_n\}$ *are independent random variables with*

1. $\max_{i\le n, j\in J_n} (\mathbb{E}|Z_{ij}|^{8+\delta})^{\frac{1}{8+\delta}} \le M$ *for some* $\delta, M > 0$*;*

2. $\min_{i\le n, j\in J_n} \mathrm{Var}(Z_{ij}) > \tau^2$ *for some* $\tau > 0$
3. $P(Z$ *has full column rank*$) = 1 - o(1)$*;*
4. $\mathbb{E}Z_j \in \mathrm{span}\{Z_j : j \in J_n^c\}$ *almost surely for all* $j \in J_n$*;*

where Z_j *is the* j*-th column of* Z*. Then*

$$\max_{j\in J_n} S_j(Z) = O_p\left(\frac{1}{n^{\frac{1}{4}}}\right) = o_p(1).$$

A typical practically interesting example is that Z contains an intercept term, which is not in J_n, and Z_j has i.i.d. entries for $j \in J_n$ with continuous distribution and sufficiently many moments, in which case the first three conditions are easily checked and $\mathbb{E}Z_j$ is a multiple of $(1, \ldots, 1)$, which belongs to $\mathrm{span}\{Z_j : j \in J_n^c\}$.

In fact, the condition 4 allows Proposition C.4 to cover more general cases than the above one. For example, in a census study, a state-specific fix effect might be added into the model, i.e.

$$y_i = \alpha_{s_i} + z_i^T \beta^* + \varepsilon_i$$

where s_i represents the state of subject i. In this case, Z contains a sub-block formed by z_i and a sub-block with ANOVA forms as mentioned in Example 1. The latter is usually incorporated only for adjusting group bias and not the target of inference. Then condition 4 is satisfied if only Z_{ij} has same mean in each group for each j, i.e. $\mathbb{E}Z_{ij} = \mu_{s_i, j}$.

Proof (Proof of Proposition C.4) By Sherman–Morison–Woodbury formula,

$$e_j^T (Z^T Z)^{-1} Z^T = \frac{Z_j^T (I - H_j)}{Z_j^T (I - H_j) Z_j}$$

where $H_j = Z_{[j]}(Z_{[j]}^T Z_{[j]})^{-1} Z_{[j]}^T$ is the projection matrix generated by $Z_{[j]}$. Then

$$S_j(Z) = \frac{\left\| e_j^T (Z^T Z)^{-1} Z^T \right\|_\infty}{\left\| e_j^T (Z^T Z)^{-1} Z^T \right\|_2} = \frac{\left\| Z_j^T (I - H_j) \right\|_\infty}{\sqrt{Z_j^T (I - H_j) Z_j}}. \tag{C-79}$$

Similar to the proofs of other examples, the strategy is to show that the numerator, as a linear contrast of Z_j, and the denominator, as a quadratic form of Z_j, are both concentrated around their means. Specifically, we will show that there exists some constants C_1 and C_2 such that

$$\max_{j\in J_n} \sup_{\substack{A\in\mathbb{R}^{n\times n}, A^2=A,\\ \mathrm{tr}(A)=n-p+1}} \left\{ P\left(\|AZ_j\|_\infty > C_1 n^{\frac{1}{4}} \right) + P\left(Z_j^T A Z_j < C_2 n \right) \right\} = o\left(\frac{1}{n}\right). \tag{C-80}$$

If (C-80) holds, since H_j is independent of Z_j by assumptions, we have

$$
\begin{aligned}
&P\left(S_j(Z) \geq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}\right) = P\left(\frac{\|Z_j^T(I-H_j)\|_\infty}{\sqrt{Z_j^T(I-H_j)Z_j}} \geq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}\right) \\
&\leq P\left(\|(I-H_j)Z_j\|_\infty > C_1 n^{\frac{1}{4}}\right) + P\left(Z_j^T(I-H_j)Z_j < C_2 n\right) \\
&= \mathbb{E}\left[P\left(\|(I-H_j)Z_j\|_\infty > C_1 n^{\frac{1}{4}}\right) \middle| Z_{[j]}\right] \\
&\quad + \mathbb{E}\left[P\left(Z_j^T(I-H_j)Z_j < C_2 n\right) \middle| Z_{[j]}\right] \qquad \text{(C-81)} \\
&\leq \sup_{A\in\mathbb{R}^{n\times n}, A^2=A, \mathrm{tr}(A)=n-p+1} P\left(\|AZ_j\|_\infty > C_1 n^{\frac{1}{4}}\right) + P\left(Z_j^T A Z_j < C_2 n\right) \\
&\leq \max_{j\in J_n}\left\{\sup_{A\in\mathbb{R}^{n\times n}, A^2=A, \mathrm{tr}(A)=n-p+1} P\left(\|AZ_j\|_\infty > C_1 n^{\frac{1}{4}}\right)\right. \\
&\quad \left. + P\left(Z_j^T A Z_j < C_2 n\right)\right\} = o\left(\frac{1}{n}\right). \qquad \text{(C-82)}
\end{aligned}
$$

Thus with probability $1 - o(|J_n|/n) = 1 - o(1)$,

$$\max_{j\in J_n} S_j(Z) \leq \frac{C_1}{\sqrt{C_2}} \cdot n^{-\frac{1}{4}}$$

and hence

$$\max_{j\in J_n} S_j(Z) = O_p\left(\frac{1}{n^{\frac{1}{4}}}\right).$$

Now we prove (C-80). The proof, although looks messy, is essentially the same as the proof for other examples. Instead of relying on the exponential concentration given by the sub-gaussianity, we show the concentration in terms of higher-order moments. In fact, for any idempotent A, the sum square of each row is bounded by 1 since

$$\sum_i A_{ij}^2 = (A^2)_{j,j} \leq \lambda_{\max}(A^2) = 1.$$

By Jensen's inequality,

$$\mathbb{E}Z_{ij}^2 \leq (\mathbb{E}|Z_{ij}|^{8+\delta})^{\frac{2}{8+\delta}}.$$

For any j, by Rosenthal's inequality [48], there exists some universal constant C such that

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^{n} A_{ij}Z_{ij}\right|^{8+\delta} &\le C\left\{\sum_{i=1}^{n}|A_{ij}|^{8+\delta}\mathbb{E}|Z_{ij}|^{8+\delta}+\left(\sum_{i=1}^{n}A_{ij}^2\mathbb{E}Z_{ij}^2\right)^{4+\delta/2}\right\}\\ &\le C\left\{\sum_{i=1}^{n}|A_{ij}|^{2}\mathbb{E}|Z_{ij}|^{8+\delta}+\left(\sum_{i=1}^{n}A_{ij}^2\mathbb{E}Z_{ij}^2\right)^{4+\delta/2}\right\}\\ &\le CM^{8+\delta}\left\{\sum_{i=1}^{n}A_{ij}^2+\left(\sum_{i=1}^{n}A_{ij}^2\right)^{4+\delta/2}\right\}\le 2CM^{8+\delta}.\end{aligned}$$

Let $C_1=(2CM^{8+\delta})^{\frac{1}{8+\delta}}$, then for given i, by Markov inequality,

$$P\left(\left|\sum_{i=1}^{n}A_{ij}Z_{ij}\right|>C_1n^{\frac{1}{4}}\right)\le\frac{1}{n^{2+\delta/4}}$$

and a union bound implies that

$$P\left(\|AZ_j\|_\infty>C_1n^{\frac{1}{4}}\right)\le\frac{1}{n^{1+\delta/4}}=o\left(\frac{1}{n}\right). \tag{C-83}$$

Now we derive a bound for $Z_j^TAZ_j$. Since $p/n\to\kappa\in(0,1)$, there exists $\tilde\kappa\in(0,1-\kappa)$ such that $n-p>\tilde\kappa n$. Then

$$\mathbb{E}Z_j^TAZ_j=\sum_{i=1}^{n}A_{ii}\mathbb{E}Z_{ij}^2>\tau^2\operatorname{tr}(A)=\tau^2(n-p+1)>\tilde\kappa\tau^2n. \tag{C-84}$$

To bound the tail probability, we need the following result: □

Lemma C-3 [2, Lemma 6.2] *Let B be an $n\times n$ nonrandom matrix and $W=(W_1,\dots,W_n)^T$ be a random vector of independent entries. Assume that $\mathbb{E}W_i=0$, $\mathbb{E}W_i^2=1$ and $\mathbb{E}|W_i|^k\le\nu_k$. Then, for any $q\ge1$,*

$$E|W^TBW-\operatorname{tr}(B)|^q\le C_q\left((\nu_4\operatorname{tr}(BB^T))^{\frac{q}{2}}+\nu_{2q}\operatorname{tr}(BB^T)^{\frac{q}{2}}\right),$$

where C_q is a constant depending on q only.

It is easy to extend Lemma C-3 to non-isotropic case by rescaling. In fact, denote σ_i^2 by the variance of W_i, and let $\Sigma=\operatorname{diag}(\sigma_1,\dots,\sigma_n)$, $Y=(W_1/\sigma_1,\dots,W_n/\sigma_n)$. Then

$$W^TBW=Y^T\Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}}Y,$$

with $\mathrm{Cov}(Y) = I$. Let $\tilde{B} = \Sigma^{\frac{1}{2}} B \Sigma^{\frac{1}{2}}$, then

$$\tilde{B}\tilde{B}^T = \Sigma^{\frac{1}{2}} B \Sigma B^T \Sigma^{\frac{1}{2}} \preceq \nu_2 \Sigma^{\frac{1}{2}} B B^T \Sigma^{\frac{1}{2}}.$$

This entails that

$$\mathrm{tr}(\tilde{B}\tilde{B}^T) \le n u_2 \,\mathrm{tr}\left(\Sigma^{\frac{1}{2}} B B^T \Sigma^{\frac{1}{2}}\right) = \nu_2 \,\mathrm{tr}(\Sigma B B^T) \le \nu_2^2 \,\mathrm{tr}(B B^T).$$

On the other hand,

$$\mathrm{tr}(\tilde{B}\tilde{B}^T)^{\frac{q}{2}} \le n \lambda_{\max}(\tilde{B}\tilde{B}^T)^{\frac{q}{2}} = n \nu_2^{\frac{q}{2}} \lambda_{\max}\left(\Sigma^{\frac{1}{2}} B B^T \Sigma^{\frac{1}{2}}\right)^{\frac{q}{2}} \le n \nu_2^q \lambda_{\max}(B B^T)^{\frac{q}{2}}.$$

Thus we obtain the following result

Lemma C-4 *Let B be an $n \times n$ nonrandom matrix and $W = (W_1, \ldots, W_n)^T$ be a random vector of independent mean-zero entries. Suppose $\mathbb{E}|W_i|^k \le \nu_k$, then for any $q \ge 1$,*

$$E|W^T B W - \mathbb{E} W^T B W|^q \le C_q \nu_2^q \left((\nu_4 \,\mathrm{tr}(B B^T))^{\frac{q}{2}} + \nu_{2q} \,\mathrm{tr}(B B^T)^{\frac{q}{2}}\right),$$

where C_q is a constant depending on q only.

Apply Lemma C-4 with $W = Z_j$, $B = A$ and $q = 4 + \delta/2$, we obtain that

$$E\left|Z_j^T A Z_j - \mathbb{E} Z_j^T A Z_j\right|^{4+\delta/2} \le C M^{16+2\delta}\left((\mathrm{tr}(AA^T))^{2+\delta/4} + \mathrm{tr}(AA^T)^{2+\delta/4}\right)$$

for some constant C. Since A is idempotent, all eigenvalues of A is either 1 or 0 and thus $AA^T \preceq I$. This implies that

$$\mathrm{tr}(AA^T) \le n, \quad \mathrm{tr}(AA^T)^{2+\delta/4} \le n$$

and hence

$$E\left|Z_j^T A Z_j - \mathbb{E} Z_j^T A Z_j\right|^{4+\delta/2} \le 2C M^{16+2\delta} n^{2+\delta/4}$$

for some constant C_1, which only depends on M. By Markov inequality,

$$P\left(|Z_j^T A Z_j - \mathbb{E} Z_j^T A Z_j| \ge \frac{\tilde{\kappa}\tau^2 n}{2}\right) \le 2C M^{16+2\delta}\left(\frac{2}{\tilde{\kappa}\tau^2}\right)^{4+\delta/2} \cdot \frac{1}{n^{2+\delta/4}}.$$

Combining with (C-84), we conclude that

$$P\left(Z_j^T A Z_j < C_2 n\right) = O\left(\frac{1}{n^{2+\delta/4}}\right) = o\left(\frac{1}{n}\right) \tag{C-85}$$

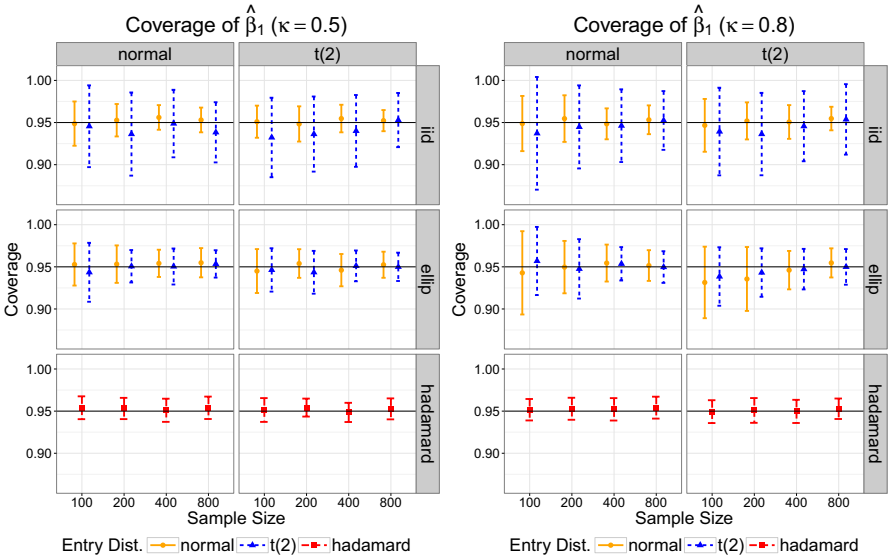

Fig. 5 Empirical 95% coverage of $\hat{\beta}_1$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design (color figure online)

where $C_2 = \frac{\tilde{\kappa}\tau^2}{2}$. Notice that both (C-83) and (C-85) do not depend on j and A. Therefore, (C-80) is proved and hence the Proposition.

D Additional numerical experiments

In this section, we repeat the experiments in Sect. 6 by using L_1 loss, i.e. $\rho(x) = |x|$. L_1-loss is not smooth and does not satisfy our technical conditions. The results are displayed below. It is seen that the performance is quite similar to that with the huber loss (Figs. 5, 6, 7).

E Miscellaneous

In this appendix we state several technical results for the sake of completeness.

Proposition E.1 ([28], formula (0.8.5.6)) *Let $A \in \mathbb{R}^{p\times p}$ be an invertible matrix and write A as a block matrix*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

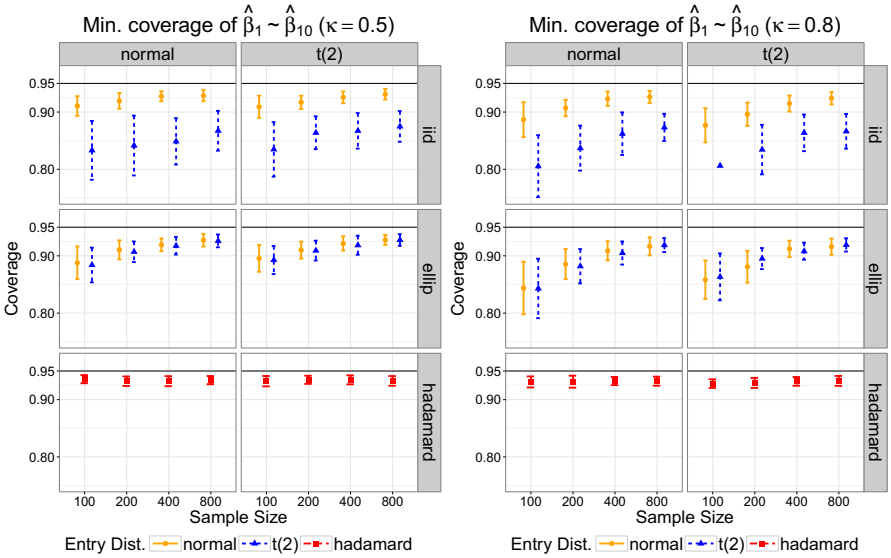

Fig. 6 Mininum empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the minimum empirical 95% coverage. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F =$ Normal; the blue dotted bar corresponds to the case $F = \mathrm{t}_2$; the red dashed bar represents the Hadamard design (color figure online)

with $A_{11} \in \mathbb{R}^{p_1 \times p_1}$, $A_{22} \in \mathbb{R}^{(p-p_1)\times(p-p_1)}$ *being invertible matrices. Then*

$$A^{-1} = \begin{pmatrix} A_{11} + A_{11}^{-1} A_{12} S^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} S^{-1} \\ -S^{-1} A_{21} A_{11}^{-1} & S^{-1} \end{pmatrix}$$

where $S = A_{22} - A_{21} A_{11}^{-1} A_{12}$ *is the Schur's complement.*

Proposition E.2 ([51]; improved version of the original form by [27]) *Let* $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ *be a random vector with independent mean-zero* σ^2*-sub-gaussian components* X_i*. Then, for every* t*,*

$$P\left(|X^T A X - \mathbb{E} X^T A X| > t\right) \leq 2 \exp\left\{-c \min\left(\frac{t^2}{\sigma^4 \|A\|_F^2}, \frac{t}{\sigma^2 \|A\|_{\mathrm{op}}}\right)\right\}$$

Proposition E.3 [3] *If* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are i.i.d. random variables with zero mean, unit variance and finite fourth moment and* $p/n \to \kappa$*, then*

$$\lambda_{\max}\left(\frac{Z^T Z}{n}\right) \xrightarrow{a.s.} (1 + \sqrt{\kappa})^2, \quad \lambda_{\min}\left(\frac{Z^T Z}{n}\right) \xrightarrow{a.s.} (1 - \sqrt{\kappa})^2.$$

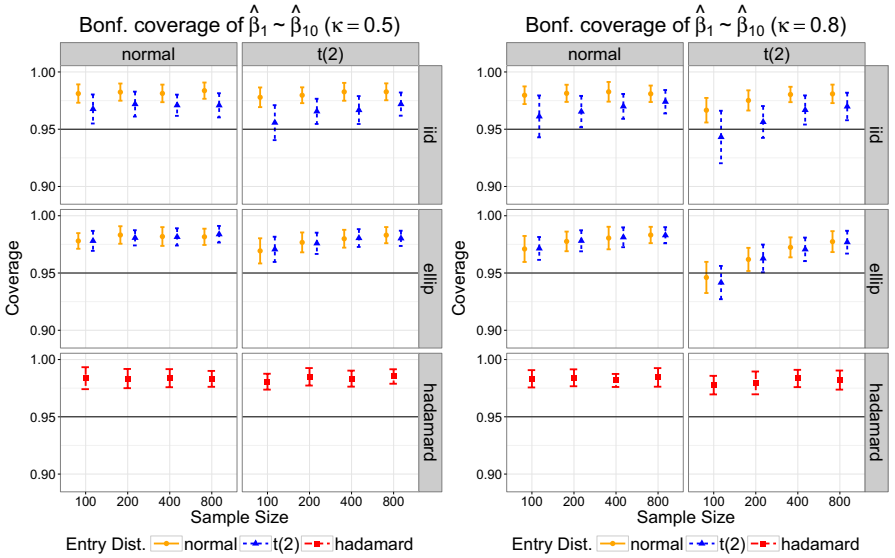

Fig. 7 Empirical 95% coverage of $\hat{\beta}_1 \sim \hat{\beta}_{10}$ after Bonferroni correction with $\kappa = 0.5$ (left) and $\kappa = 0.8$ (right) using L_1 loss. The x-axis corresponds to the sample size, ranging from 100 to 800; the y-axis corresponds to the empirical uniform 95% coverage after Bonferroni correction. Each column represents an error distribution and each row represents a type of design. The orange solid bar corresponds to the case $F = \text{Normal}$; the blue dotted bar corresponds to the case $F = \text{t}_2$; the red dashed bar represents the Hadamard design (color figure online)

Proposition E.4 [35] *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are independent mean-zero random variables with finite fourth moment, then*

$$\mathbb{E}\sqrt{\lambda_{\max}\left(Z^T Z\right)} \le C\left(\max_i \sqrt{\sum_j \mathbb{E}Z_{ij}^2} + \max_j \sqrt{\sum_i \mathbb{E}Z_{ij}^2} + \sqrt[4]{\sum_{i,j} \mathbb{E}Z_{ij}^4}\right)$$

for some universal constant C. *In particular, if* $\mathbb{E}Z_{ij}^4$ *are uniformly bounded, then*

$$\lambda_{\max}\left(\frac{Z^T Z}{n}\right) = O_p\left(1 + \sqrt{\frac{p}{n}}\right).$$

Proposition E.5 [50] *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are independent mean-zero* σ^2*-sub-gaussian random variables. Then there exists a universal constant* $C_1, C_2 > 0$ *such that*

$$P\left(\sqrt{\lambda_{\max}\left(\frac{Z^T Z}{n}\right)} > C\sigma\left(1 + \sqrt{\frac{p}{n}} + t\right)\right) \le 2e^{-C_2 n t^2}.$$

Proposition E.6 [49] *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are i.i.d.* σ^2*-sub-gaussian random variables with zero mean and unit variance, then for* $\varepsilon \geq 0$

$$P\left(\sqrt{\lambda_{\min}\left(\frac{Z^T Z}{n}\right)} \leq \varepsilon(1 - \sqrt{\frac{p-1}{n}})\right) \leq (C\varepsilon)^{n-p+1} + e^{-cn}$$

for some universal constants C *and* c.

Proposition E.7 [37] *Suppose* $\{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, p\}$ *are independent* σ^2*-sub-gaussian random variables such that*

$$Z_{ij} \stackrel{d}{=} -Z_{ij}, \quad \mathrm{Var}(Z_{ij}) > \tau^2$$

for some $\sigma, \tau > 0$*, and* $p/n \to \kappa \in (0, 1)$*, then there exists constants* $c_1, c_2 > 0$*, which only depends on* σ *and* τ*, such that*

$$P\left(\lambda_{\min}\left(\frac{Z^T Z}{n}\right) < c_1\right) \leq e^{-c_2 n}.$$

References

1. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, New York (1962)
2. Bai, Z., Silverstein, J.W.: Spectral Analysis of Large Dimensional Random Matrices, vol. 20. Springer, Berlin (2010)
3. Bai, Z., Yin, Y.: Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann. Probab. **21**(3), 1275–1294 (1993)
4. Baranchik, A.: Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. Ann. Stat. **1**(2), 312–321 (1973)
5. Bean, D., Bickel, P.J., El Karoui, N., Lim, C., Yu, B.: Penalized robust regression in high-dimension. Technical Report 813, Department of Statistics, UC Berkeley (2012)
6. Bean, D., Bickel, P.J., El Karoui, N., Yu, B.: Optimal M-estimation in high-dimensional regression. Proc. Natl. Acad. Sci. **110**(36), 14563–14568 (2013)
7. Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, vol. 117. CRC Press, Boca Raton (2015)
8. Bickel, P.J., Freedman, D.A.: Some asymptotic theory for the bootstrap. Ann. Stat. **9**(6), 1196–1217 (1981)
9. Bickel, P.J., Freedman, D.A.: Bootstrapping regression models with many parameters. Festschrift for Erich L. Lehmann pp. 28–48 (1983)
10. Chatterjee, S.: Fluctuations of eigenvalues and second order Poincaré inequalities. Probab. Theory Relat. Fields **143**(1–2), 1–40 (2009)
11. Chernoff, H.: A note on an inequality involving the normal distribution. Ann. Probab. **9**(3), 533–535 (1981)
12. Cizek, P., Härdle, W.K., Weron, R.: Statistical Tools for Finance and Insurance. Springer, Berlin (2005)
13. Cochran, W.G.: Sampling Techniques. Wiley, Hoboken (1977)
14. David, H.A., Nagaraja, H.N.: Order Statistics. Wiley Online Library, Hoboken (1981)
15. Donoho, D., Montanari, A.: High dimensional robust M-estimation: asymptotic variance via approximate message passing. Probab. Theory Relat. Fields **166**, 935–969 (2016)
16. Durrett, R.: Probability: Theory and Examples. Cambridge University Press, Cambridge (2010)
17. Efron, B.: The Jackknife, the Bootstrap and Other Resampling Plans, vol. 38. SIAM, Philadelphia (1982)

18. El Karoui, N.: Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. Ann. Appl. Probab. **19**(6), 2362–2405 (2009)
19. El Karoui, N.: High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. Ann. Stat. **38**(6), 3487–3566 (2010)
20. El Karoui, N.: Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. arXiv preprint arXiv:1311.2445 (2013)
21. El Karoui, N.: On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. Probab. Theory Relat. Fields, pp. 1–81 (2015)
22. El Karoui, N., Bean, D., Bickel, P.J., Lim, C., Yu, B.: On Robust Regression with High-Dimensional Predictors. Technical Report 811, Department of Statistics, UC Berkeley (2011)
23. El Karoui, N., Bean, D., Bickel, P.J., Lim, C., Yu, B.: On robust regression with high-dimensional predictors. Proc. Natl. Acad. Sci. **110**(36), 14557–14562 (2013)
24. El Karoui, N., Purdom, E.: Can We Trust the Bootstrap in High-Dimension? Technical Report 824. Department of Statistics, UC Berkeley (2015)
25. Esseen, C.G.: Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law. Acta Math. **77**(1), 1–125 (1945)
26. Geman, S.: A limit theorem for the norm of random matrices. Ann. Probab. **8**(2), 252–261 (1980)
27. Hanson, D.L., Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables. Ann. Math. Stat. **42**(3), 1079–1083 (1971)
28. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)
29. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**(1), 73–101 (1964)
30. Huber, P.J.: The 1972 wald lecture robust statistics: a review. Ann. Math. Stat. **43**(4), 1041–1067 (1972)
31. Huber, P.J.: Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. **1**(5), 799–821 (1973)
32. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
33. Johnstone, I.M.: On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat. **29**(2), 295–327 (2001)
34. Jurečkovà, J., Klebanov, L.B.: Inadmissibility of robust estimators with respect to L_1 norm. In: Dodge, Y. (ed.) L_1-Statistical Procedures and Related Topics. Lecture Notes–Monograph Series, vol. 31, pp. 71–78. Institute of Mathematical Statistics, Hayward (1997)
35. Latała, R.: Some estimates of norms of random matrices. Proc. Am. Math. Soc. **133**(5), 1273–1282 (2005)
36. Ledoux, M.: The Concentration of Measure Phenomenon, vol. 89. American Mathematical Society, Providence (2001)
37. Litvak, A.E., Pajor, A., Rudelson, M., Tomczak-Jaegermann, N.: Smallest singular value of random matrices and geometry of random polytopes. Adv. Math. **195**(2), 491–523 (2005)
38. Mallows, C.: A note on asymptotic joint normality. Ann. Math. Stat. **43**(2), 508–515 (1972)
39. Mammen, E.: Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Stat. **17**(1), 382–400 (1989)
40. Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Math. USSR Sbornik **1**(4), 457 (1967)
41. Muirhead, R.J.: Aspects of Multivariate Statistical Theory, vol. 197. Wiley, Hoboken (1982)
42. Portnoy, S.: Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. Ann. Stat. **12**(4), 1298–1309 (1984)
43. Portnoy, S.: Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. Ann. Stat. **13**(4), 1403–1417 (1985)
44. Portnoy, S.: On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. Probab. Theory Relat. Fields **73**(4), 571–583 (1986)
45. Portnoy, S.: A central limit theorem applicable to robust regression estimators. J. Multivar. Anal. **22**(1), 24–50 (1987)
46. Posekany, A., Felsenstein, K., Sykacek, P.: Biological assessment of robust noise models in microarray data analysis. Bioinformatics **27**(6), 807–814 (2011)
47. Relles, D.A.: Robust Regression by Modified Least-Squares. Technical reports, DTIC Document (1967)
48. Rosenthal, H.P.: On the subspaces of $l^p (p > 2)$ spanned by sequences of independent random variables. Isr. J. Math. **8**(3), 273–303 (1970)
49. Rudelson, M., Vershynin, R.: Smallest singular value of a random rectangular matrix. Commun. Pure Appl. Math. **62**(12), 1707–1739 (2009)

50. Rudelson, M., Vershynin, R.: Non-asymptotic theory of random matrices: extreme singular values. arXiv preprint arXiv:1003.2990 (2010)
51. Rudelson, M., Vershynin, R.: Hanson-Wright inequality and sub-gaussian concentration. Electron. Commun. Probab. **18**(82), 1–9 (2013)
52. Scheffe, H.: The Analysis of Variance, vol. 72. Wiley, Hoboken (1999)
53. Silverstein, J.W.: The smallest eigenvalue of a large dimensional Wishart matrix. Ann. Probab. **13**(4), 1364–1368 (1985)
54. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B (Methodolog) **36**(2), 111–147 (1974)
55. Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**(1), 234–251 (1987)
56. Van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)
57. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027 (2010)
58. Wachter, K.W.: Probability plotting points for principal components. In: Ninth Interface Symposium Computer Science and Statistics, pp. 299–308. Prindle, Weber and Schmidt, Boston (1976)
59. Wachter, K.W.: The strong limits of random matrix spectra for sample matrices of independent elements. Ann. Probab. **6**(1), 1–18 (1978)
60. Wasserman, L., Roeder, K.: High dimensional variable selection. Ann. Stat. **37**(5A), 2178 (2009)
61. Yohai, V.J.: Robust M-Estimates for the General Linear Model. Universidad Nacional de la Plata. Departamento de Matematica (1972)
62. Yohai, V.J., Maronna, R.A.: Asymptotic behavior of M-estimators for the linear model. Ann. Stat. **7**(2), 258–268 (1979)