Rate Optimal Estimation and Confidence Intervals for High-dimensional Regression with Missing Covariates

Yining Wang[1], Jialei Wang[3], Sivaraman Balakrishnan[1,2], and Aarti Singh[1]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Statistics, Carnegie Mellon University
[3]Department of Computer Science, University of Chicago

November 6, 2017

Abstract

Although a majority of the theoretical literature in high-dimensional statistics has focused on settings which involve fully-observed data, settings with missing values and corruptions are common in practice. We consider the problems of estimation and of constructing component-wise confidence intervals in a sparse high-dimensional linear regression model when some covariates of the design matrix are missing completely at random. We analyze a variant of the Dantzig selector [9] for estimating the regression model and we use a de-biasing argument to construct component-wise confidence intervals. Our first main result is to establish upper bounds on the estimation error as a function of the model parameters (the sparsity level s, the expected fraction of observed covariates ρ_*, and a measure of the signal strength $\|\beta^*\|_2$). We find that even in an idealized setting where the covariates are assumed to be missing completely at random, somewhat surprisingly and in contrast to the fully-observed setting, there is a dichotomy in the dependence on model parameters and much faster rates are obtained if the covariance matrix of the random design is known. To study this issue further, our second main contribution is to provide lower bounds on the estimation error showing that this discrepancy in rates is unavoidable in a minimax sense. We then consider the problem of high-dimensional inference in the presence of missing data. We construct and analyze confidence intervals using a de-biased estimator. In the presence of missing data, inference is complicated by the fact that the de-biasing matrix is correlated with the pilot estimator and this necessitates the design of a new estimator and a novel analysis. We also complement our mathematical study with extensive simulations on synthetic and semi-synthetic data that show the accuracy of our asymptotic predictions for finite sample sizes.

1 Introduction

High-dimensional statistics concerns the setting where the dimension of the statistical model is comparable to, or even far exceeds, the sample-size. In this context, meaningful statistical estimation is impossible in the absence of additional structure. Accordingly, significant research in high-dimensional statistics (see for instance [10, 15, 16, 17, 35]) has focused on high-dimensional linear regression with sparsity constraints where the goal is estimate or perform inference on a sparse, high-dimensional vector β^* given access to noisy linear measurements.

Modern datasets are frequently afflicted with missing-values and corruptions. As a canonical example consider the gene-expression dataset from Nielsen et al. [30]. This dataset records $p = 5520$ genes for $n = 46$ patients with soft tissue tumors. A total of 6.7% entries are missing; furthermore,

78.6% of the 5520 genes and all of the 46 patients have at least one missing covariate. Motivated by the analysis of corrupted high-dimensional datasets several researchers have considered settings with corrupted covariates: focusing on developing high-dimensional analogues of the classical Expectation-Maximization (EM) algorithm [34], studying their algorithmic convergence properties [2, 38, 39], and understanding statistical rates of convergence for other estimators [5, 13, 25, 26, 27, 32, 33].

Despite extensive past work, several challenging and important open questions remain in establishing the correct dependence of the rates of convergence in missing data problems on model parameters (the sparsity level s, the expected fraction of unobserved covariates ρ_*, and the signal strength $\|\beta^*\|_2$). Understanding these dependencies for the problems of high-dimensional estimation and inference are the focus of this work.

1.1 Preliminaries

We focus on a random design regression model where we observe i.i.d. samples of $y \in \mathbb{R}$, linked to a covariate $X \in \mathbb{R}^p$ through the linear model:

$$y_i = \langle X_i, \beta^* \rangle + \epsilon_i, \tag{1}$$

where ϵ_i is i.i.d. mean zero Gaussian noise, i.e. $\epsilon_i \sim N(0, \sigma_\varepsilon^2)$. Popular estimators include the LASSO [35], the SCAD [17] and the Dantzig selector [9], whose asymptotic rates of convergence and model selection properties are well understood [1, 6, 37, 41]. We further consider the setting where covariates are missing completely at random, i.e. rather than observe the covariates X_i, we observe $\overline{X}_i$ where,

$$\overline{X}_{ij} = \begin{cases} \star \text{ with probability } 1 - \rho_j \\ X_{ij} \text{ otherwise,} \end{cases} \tag{2}$$

where we assume that the probabilities ρ_j are known and define

$$\rho_* = \min_{1 \leqslant i \leqslant p} \rho_j.$$

Our goal is to either estimate or to construct coordinate-wise confidence intervals for the unknown vector β^*. In the high-dimensional setting, the number of observed samples n can be much smaller than p and consistent estimation is impossible without additional structural assumptions. Accordingly, we study sparse models where β^* has at most s nonzero components, where s is allowed to grow with p and n, but satisfies $s \ll n$.

We emphasize that in this model, and indeed in many practical settings (for instance in the dataset of [30]), most samples will have corrupted covariates and as a result complete-case analyses [24] are wasteful. Methods based on data imputation [24] typically require stronger knowledge about the generative process which can be difficult to justify in a high-dimensional setting and taking into account the imputation error in subsequent inference can be challenging.

1.2 Related work

Classical work on statistical estimation and inference in the presence of missing data is extensive (see for instance [11, 19, 24] and references therein), and we focus in this section on closely related works focusing on the sparse high-dimensional setting.

Rosenbaum & Tsybakov [32] proposed the Matrix Uncertainty (MU)-selector for high-dimensional regression under an error-in-variables model, where the design matrix X is observed with deterministic measurement error W that is bounded in the matrix maximum norm. Optimization algorithms

and minimax rates when W is Gaussian white noise are considered in the work [5]. The MU-selector was generalized to handle the missing data setting in the paper [33], and it was found that de-biasing the estimator of the covariance matrix led to improved error bounds. Datta & Zou [14] proposed CoCoLasso, a variant of the LASSO for error-in-variable models where a covariance estimate $\widehat{\Sigma}$ is first projected onto a positive semi-definite cone so that the resulting LASSO problem is convex. Both additive and multiplicative measurement error models were considered in this work and corresponding rates of convergence were derived.

Loh & Wainwright [25] analyzed a gradient descent algorithm for optimizing a non-convex LASSO-type loss function and derived rates of convergence from both statistical and optimization perspectives. Their analysis shows a dependency on $1/\rho_*^4$ for the ℓ_2^2 estimation error. A similar rate of convergence was established in [13] for orthogonal matching pursuit (OMP) type estimators, and Datta & Zou [14], Rosenbaum & Tsybakov [33] for MU-selector and CocoLasso formulations. On the lower bound side, [26] derived lower bounds on the minimax rate, under the assumptions of identity covariance for the design points and bounded signal level $\|\beta^*\|_2$. Their lower bounds depend linearly on $1/\rho_*$. [5] showed that the dependency on $\|\beta^*\|_2$ is necessary for error-in-variable models of high-dimensional regression. However, subtle differences exist between the error-in-variables models considered in [5] and the missing data model consider in this paper, which are reflected in the dependency on the missing rate ρ_* and the interplay between the two terms of σ_ε and $\|\beta^*\|_2$, which exhibit different levels of dependency on ρ_*.

The gap between the upper and lower bounds of prior work on estimation [25, 26] motivate part of this work. We show that in the setting where the design covariance is assumed known a linear dependence on $1/\rho_*$ is achievable, whereas in the case when the covariance matrix is unknown a dependence on $1/\rho_*^2$ is unavoidable. We provide a sharper upper bound than that of Loh & Wainwright [25], and further provide a novel lower bound for the setting with unknown covariance. These results taken together reveal an interesting phenomenon where the rates of estimation depend on whether the covariance matrix of the random design is assumed to be known[1]. From a practical standpoint, when ρ_* is small the difference between estimators that have dependence $1/\rho_*^4$ and those that depend on $1/\rho_*$ can be significant and we investigate these issues further via extensive simulations.

Recent work in high-dimensional statistics has focused on inference for (low-dimensional projections of) β^* [8, 21, 36, 40]. We consider this problem, in the missing completely at random model described in (1), and analyze the performance of a de-biased version of the Dantzig selector. An important distinction between existing de-biasing methods and ours is that the presence of missing data causes the de-biasing matrix to be correlated with the estimator $\widehat{\beta}$. This in turn complicates the analysis and results in a limiting distribution that depends on the missing covariates. We use a variant of the CLIME estimator [7] to resolve this correlation issue and propose a data-driven estimator for the limiting variance of the de-biased estimator.

While we were preparing this manuscript, [3] posted a paper that discusses the similar problem of constructing confidence bands for high-dimensional linear models with measurement errors by considering an estimator based on orthogonal score functions. Though the results of [3] could also be applied to missing data settings, the optimal dependency on the observation rate ρ was not studied.

1.3 Outline

The remainder of the paper is organized as follows. In Section 2 we consider the problem of estimation in the presence of missing data: in particular, Theorem 1 analyzes a variant of the Dantzig selector in both the setting where the covariance of X is taken to be known and in the setting where the

[1]Taking the viewpoint of semi-supervised estimation [12, 22], these results show that, in contrast to linear regression in the uncorrupted setting, unlabeled data, i.e. covariates $\overline{X}_i$ with no associated y_i can be useful in settings with missing data.

covariance is unknown. Under appropriate assumptions, these results show a $1/\rho_*$ dependence in the setting where the covariance is known and a $1/\rho_*^2$ dependence when the covariance is unknown. The dependency over $1/\rho_*$ is better than existing estimators [13, 25] under similar settings, which depend on $1/\rho_*^4$. We turn to lower bounds in Theorems 2 and 3, where we provide in turn minimax lower bounds for the known and unknown covariance settings, showing roughly that the previously obtained dependencies are optimal. In Section 3 we consider the problem of high-dimensional inference in the presence of missing data. In Theorem 4 we derive the limiting distribution of a de-biased Dantzig selector, while in Theorem 5 we provide an estimate of the limiting variance to allow for a practical, data-driven construction of confidence intervals. We provide extensive simulations on synthetic and semi-synthetic data in Section 4, and discuss our results and open problems in Section 5. We provide detailed technical proofs in Section 6 with remaining technical aspects deferred to the Appendix.

1.4 Notation

For a vector x, we use $\|x\|_p := \left(\sum_j |x_j|^p\right)^{1/p}$ to denote the ℓ_p-norm of x. For a matrix A, we use $\|A\|_{L_p}$ to denote the operator p-norm of A; that is, $\|A\|_{L_p} = \sup_{x\neq 0} \|Ax\|_p/\|x\|_p$. We also write $\|A\|_{L_\infty}$ for the maximum norm of a matrix: $\|A\|_{L_\infty} = \max_{j,k} |A_{jk}|$. For a positive semi-definite matrix A, we denote by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ the largest and smallest eigenvalues of A. We use $\mathbb{B}_p(M) = \{x : \|x\|_p \leqslant M\}$ to denote the ℓ_p ball of radius M centered at the origin.

2 Rate-optimal Estimation

In this section we present our main results on estimation in the high-dimensional missing completely at random model. We begin with a description of our estimator which is a modified version of the Dantzig selector. As with the modified LASSO estimator (see [26]) the modified Dantzig selector requires a plug-in estimate of the covariance matrix. In contrast to the modified LASSO, the modified Dantzig selector remains a convex program even if the plug-in covariance matrix is not positive semi-definite and this leads to computational advantages as well as a simpler analysis. We subsequently state the assumptions that underlie our analysis, and then give precise statements of our upper and lower bounds. We defer proofs of these results to Section 6.

2.1 The modified Dantzig selector

We abuse notation slightly and use $\overline{X}$ to denote the observed covariates in (2) with zero-imputation, i.e. with each $\star$ replaced by 0. We denote unbiased estimators of X and its covariance matrix by $\widetilde{X} \in \mathbb{R}^{n\times p}$ and $\widetilde{\Sigma} \in \mathbb{R}^{p\times p}$ which we define as

$$\widetilde{X}_{ij} := \frac{\overline{X}_{ij}}{\rho_j}, \quad \widetilde{\Sigma} := \frac{1}{n}\widetilde{X}^\top \widetilde{X} - D\,\mathrm{diag}\left(\frac{1}{n}\widetilde{X}^\top \widetilde{X}\right), \tag{3}$$

where $D = \mathrm{diag}(1-\rho_1, \cdots, 1-\rho_p)$ is a known $p \times p$ diagonal matrix. It is a simple observation that, conditioned on X, $\mathbb{E}[\widetilde{X}] = X$ and $\mathbb{E}[\widetilde{\Sigma}] = \widehat{\Sigma} = \frac{1}{n}X^\top X$. Our modified Dantzig selector is defined as the solution to the convex program:

$$\widehat{\beta}_n \in \mathrm{argmin}_{\beta\in\mathbb{R}^p} \left\{ \|\beta\|_1 : \left\| \frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\beta \right\|_\infty \leqslant \widetilde{\lambda}_n \right\}, \tag{4}$$

where $\widetilde{\lambda}_n > 0$ is a tuning parameter. Eq. (4) is a variant of the Dantzig selector [9] and is in principle similar to the MU-selector in [32]. We note again that the estimator in (4) is always a convex

optimization problem (regardless of whether $\widetilde{\Sigma}$ is positive semi-definite) and hence can be efficiently computed.

We also consider a variant of the modified Dantzig selector for the idealized scenario where the population covariance $\Sigma_0 = \mathbb{E}[X^\top X]$, for the design matrix is known. In particular, we define $\breve{\beta}_n$ as the solution of

$$\breve{\beta}_n \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 : \left\| \frac{1}{n} \widetilde{X}^\top y - \Sigma_0 \beta \right\|_\infty \leqslant \breve{\lambda}_n \right\}, \tag{5}$$

where we replace the covariance estimate $\widetilde{\Sigma}$ with the known population covariance Σ_0. Noting that the high-dimensional covariance matrix Σ_0 is rarely known in practice, we introduce and analyze this estimator primarily as a theoretical benchmark.

2.2 Assumptions

The analysis in subsequent sections of our paper rely on certain assumptions on the covariates, the noise and the missingness mechanism:

(A1) *Homogenous Gaussian noise*: For each $i \in \{1, \ldots, n\}$, the stochastic noise is independent and identically distributed with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ for some (known) $\sigma_\varepsilon < \infty$.

(A2) *Sub-Gaussian random design*: Each row of X is sampled i.i.d. from some underlying sub-Gaussian distribution with covariance Σ_0 and (known) sub-Gaussian parameter $\sigma_x < \infty$. We further suppose that the population covariance is well-conditioned, i.e. that $0 < \lambda_{\min}(\Sigma_0) \leqslant \lambda_{\max}(\Sigma_0) < \infty$. For notational simplicity we take Σ_0 to be implicit and use $\lambda_{\min}, \lambda_{\max}$ instead in the rest of this paper.

(A3) *Missing completely at random*: Each covariate $j \in \{1, \ldots, p\}$ has entries missing completely at random with probability of observing each entry being equal to ρ_j, and define $\rho_* = \min_{1 \leqslant j \leqslant p} \rho_j > 0$.

(A4) *Sparsity*: The support set $J_0 = \operatorname{supp}(\beta^*) = \{j : |\beta_j^*| \neq 0\}$ satisfies $|J_0| \leqslant s$ for some $s \ll n$.

The assumptions are standard in theoretical work on high-dimensional regression with missing data. We note that assumption (A2) implies (with high probability) a deterministic Restricted Eigenvalue (RE) condition [6] on the sample covariance of X.

2.3 Rates of convergence and minimax lower bounds

We now turn our attention to providing rates of convergence and minimax lower bounds on the estimation error. Theorem 1 establishes upper bounds on the mean square estimation error of β^*. Eq. (6) corresponds to the setting where the population covariance Σ_0 is known and Eq. (7) holds when Σ_0 is unknown.

The following result applies to the modified Dantzig selectors in (4) and (5), where the tuning parameters are chosen as:

$$\breve{\lambda}_n, \widetilde{\lambda}_n \asymp (\sigma_x^2 \|\beta^*\|_2 + \sigma_x \sigma_\varepsilon) \sqrt{\frac{\log p}{\rho_* n}}.$$

Theorem 1. *Assume that (A1) to (A4) are satisfied.*

- **Known Covariance:** *If* $\frac{\log p}{\rho_*^2 n} \to 0$ *then*

$$\|\check{\beta}_n - \beta^*\|_2 = O_{\mathbb{P}}\left\{\frac{\sigma_x^2}{\lambda_{\min}}\left(\|\beta^*\|_2\sqrt{\frac{s\log p}{\rho_* n}} + \frac{\sigma_\varepsilon}{\sigma_x}\sqrt{\frac{s\log p}{\rho_* n}}\right)\right\}. \tag{6}$$

- **Unknown Covariance:** *If* $\max\left\{\frac{\sigma_x^4 s\log(\sigma_x p/\rho_*)}{\rho_*^3\lambda_{\min}^2 n}, \frac{\log p}{\rho_*^4 n}\right\} \to 0$, *then*

$$\|\widehat{\beta}_n - \beta^*\|_2 = O_{\mathbb{P}}\left\{\frac{\sigma_x^2}{\lambda_{\min}}\left(\|\beta^*\|_2\sqrt{\frac{s\log p}{\rho_*^2 n}} + \frac{\sigma_\varepsilon}{\sigma_x}\sqrt{\frac{s\log p}{\rho_* n}}\right)\right\}. \tag{7}$$

Remarks:

1. The two results show that at least from the perspective of upper bounds there is a gap in the rates achieved by the modified Dantzig selector in the known and unknown covariance settings. In particular, the squared estimation error where Σ_0 is known scales as $1/\rho_*$ while in the setting where Σ_0 is unknown scales as $1/\rho_*^2$.

2. Compared to Loh & Wainwright [25] our bounds are better by an $O(1/\rho^*)$ factor for $\widehat{\beta}_n$ when Σ_0 is unknown and an $O(1/\rho_*^{3/2})$ factor better when Σ_0 is known. Our bounds are not directly comparable to the work of Rosenbaum & Tsybakov [32] which considers a fixed-design setting with no stochastic model assumed over X. We however remark that error bounds in Rosenbaum & Tsybakov [32] depend on $\|\beta^*\|_1$, which could be a factor of $\sqrt{s}$ worse than $\|\beta^*\|_2$. The dependency on $\|\beta^*\|_1$ of MU-selector type estimators was later improved by [4] by considering an additional ℓ_∞ norm regularization. The latter paper however considers the general error-in-variable models, and dependency on ρ^* in a missing data model is not explicitly stated.

3. The conditions between n and other model parameters that we require for the error bounds to hold arise from the use of Bernstein-type concentration inequalities. In the missing data setting, controlling the deviation of the empirical and true covariance matrix of X (for instance) requires a careful analysis of moments of the observed matrix $\overline{X}$ and a subsequent application of Bernstein-type concentration inequalities. This leads to two distinct tail behaviours, the more typical sub-Gaussian tail behaviour depending on the variance of the summands when n is sufficiently large and the small-sample sub-exponential tail behaviour. To ease readability, we focus on the sub-Gaussian behaviour by assuming the sample size is sufficiently large. We discuss this further in Section 5.

4. We also note that in contrast to bounds for regression without missing data the upper bounds here, somewhat counterintuitively, deteriorate as $\|\beta^*\|_2$ gets larger. This has been observed in prior work [2, 25] and is roughly due to the fact that as $\|\beta^*\|_2$ grows (keeping ρ_* fixed) more information is missing in each sample.

5. We note that bounds on the ℓ_1 estimation error follow in a straightforward way using the relationships that under the conditions of the theorem with high-probability we have that, $\|\widehat{\beta}_n - \beta^*\|_1 \leqslant 2\sqrt{s}\|\widehat{\beta}_n - \beta^*\|_2$ and $\|\check{\beta}_n - \beta^*\|_1 \leqslant 2\sqrt{s}\|\check{\beta}_n - \beta^*\|_2$.

We now turn our attention to minimax lower bounds for the estimation error. We focus first on the case when the covariance matrix Σ_0 is assumed to be known. In this setting, we follow a similar argument to that of prior work [26] but we maintain the dependence on the various model parameters (particularly, σ_ε and $\|\beta^*\|_2$) in the lower bound.

Theorem 2. Known Covariance: *Suppose* $4 \leqslant s < 4p/5$, $\frac{s\log(p/s)}{\rho_* n} \to 0$ *and* $\Sigma_0 = I$. *Then there exists a universal constant* $C_0 > 0$ *and an arbitrary constant* $c > 0$ *such that,*

$$\inf_{\widehat{\beta}_n} \sup_{\beta^* \in \mathbb{B}_2(M) \cap \mathbb{B}_0(s)} \mathbb{E}\|\widehat{\beta}_n - \beta^*\|_2^2$$
$$\geqslant C_0 \cdot \min\left\{\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2, e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\right\} \cdot \min\left\{\sqrt{\frac{s\log(p/s)}{(1-\rho_*)^2 n}}, \frac{s\log(p/s)}{\rho_* n}\right\}. \quad (8)$$

Remarks:

1. In the setting when $\frac{(1-\rho_*)^2 s\log(p/s)}{\rho_*^2 n} \to 0$ the lower bound can be simplified to:

$$C_0 \cdot \min\left\{\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2, e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\right\} \frac{s\log(p/s)}{\rho_* n}.$$

   Furthermore, if the missing rate $(1-\rho_*)$ is at least a constant and the sparsity level s or the noise level σ_ε is not too small, the term $e^{c^2(1-\rho_*)s}\sigma_\varepsilon^2$ is negligible because it increases exponentially with s (and thus does not contribute to the minimum). In this case, noting that in our lower bound both $\lambda_{\min}$ and $\sigma_x = 1$, we see that the lower bound matches the upper bound in (6) upto a universal constant.

2. We note that the second term in the lower bound arises from an interesting aspect of the missing data problem, roughly $n/\exp((1-\rho_*)s)$ samples obtained from the model are uncorrupted. In this case, as indicated by our lower bound a complete-case analysis (simply throwing away the samples with missing covariates) will lead to a matching upper bound, i.e. an upper bound that does not depend on $\|\beta^*\|_2$.

In the case when Σ_0 is unknown, our primary goal is to show that the $1/\rho_*^2$ dependence in the upper bound is unavoidable. To accomplish this we need to consider packing sets of the parameters where both the covariance matrix Σ_0 and the unknown regression vector β^* are varied. This calculation is quite technical, and as we discuss further in Section 5, we are unable to prove a sharp lower bound on the mean-squared estimation error. Instead we consider lower bounding the minimax estimation error for estimating a single coordinate of the vector β^*, and show that this task already requires a sample-size that scales as $1/\rho_*^2$. Formally, we fix a small positive constant $\gamma_0 \in (0, 1/2)$ and define,

$$\Lambda(\gamma_0) = \{\Sigma_0 \in \mathbb{S}_+^p : 1 - \gamma_0 \leqslant \lambda_{\min}(\Sigma_0) \leqslant \lambda_{\max}(\Sigma_0) \leqslant 1 + \gamma_0\},$$

where $\mathbb{S}_+^p$ is the class of all positive definite $p \times p$ matrices. We have the following result:

Theorem 3. *Suppose that* $s \geqslant 4$, $\max\{\frac{\sigma_\varepsilon^2}{M^2\rho_* n}, \frac{1}{\gamma_0\rho_*^2 n}\} \to 0$. *Then for any fixed* $j \in \{1, \ldots, p\}$ *there is a universal constant* $C_1 > 0$ *and an arbitrary constant* $c > 0$ *such that,*

$$\inf_{\widehat{\beta}_n} \sup_{\substack{\beta^* \in \mathbb{B}_2(M) \cap \mathbb{B}_0(s) \\ \Sigma_0 \in \Lambda(\gamma_0)}} \mathbb{E}|\widehat{\beta}_{nj} - \beta_j^*|^2 \geqslant C_1 \cdot \max\left\{\frac{\sigma_\varepsilon^2}{\rho_* n}, \min\left(\frac{1-\rho_*}{1+2c}M^2, e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\right)\frac{1}{\rho_*^2 n}\right\}.$$

Remarks:

1. Once again for simplicity considering the case when the sparsity level s is not too small, the lower bound scales as roughly $\|\beta^*\|_2^2/(\rho_*^2 n)$, indicating that the $1/\rho_*^2$ dependence obtained in the upper bound is unavoidable in general.

2. Our lower bound is for the error of estimating a single co-ordinate of β^*, and is derived from a careful perturbation of the covariance matrix Σ_0 and regression vector β^* for which we are able to analyze the KL divergence quite precisely. Extending our lower bound to obtain an $s\log(p/s)$ scaling seems to be a challenging but important avenue for further investigation and we discuss this issue further in Section 5.

3 Confidence intervals for regression coefficients

In this section we turn our attention to the problem of constructing confidence intervals for coordinates of β^*. We describe a method that builds confidence intervals for β^* by de-biasing the modified Dantzig selector. The de-biasing method builds on recent work [36] and requires a sufficiently accurate estimate of the precision matrix Σ_0^{-1}. This in turn requires the following additional assumption:

(A5) There exist known constants $b_0, b_1 < \infty$ such that each row (and column) of Σ_0^{-1} belongs to $\mathbb{B}_0(b_0) \cap \mathbb{B}_1(b_1)$, i.e. each row of Σ_0^{-1} is b_0-sparse and $\|\Sigma_0^{-1}\|_{L_1} \leqslant b_1$.

Condition (A5) allows us to use CLIME [7] or the node-wise LASSO [28] to estimate an approximate inverse of Σ_0 that asymptotically de-biases the estimate $\widehat{\beta}_n$ from (4). Similar conditions for high-dimensional inference were studied in [36]. We discuss potential settings where (A5) could be relaxed in Section 5.

3.1 The de-biased modified Dantzig selector

In this section, we first introduce our de-biased estimator and then analyze its asymptotic distribution. In the next section we provide a data-driven method to estimate the limiting variance of the de-biased estimator. The de-biasing procedure uses an estimate of the precision matrix which we obtain by solving the CLIME optimization program from [7]. Formally, we choose a tuning parameter

$$\widetilde{\nu}_n \asymp \sigma_x^2 b_1 \sqrt{\frac{\log p}{\rho_*^2 n}}.$$

Recalling, the matrix $\widetilde{\Sigma}$ in (3) we define $\widehat{\Theta}$ to be the $p \times p$ matrix:

$$\widehat{\Theta} \in \operatorname{argmin}_{\Theta \in \mathbb{R}^{p\times p}} \left\{ \|\Theta\|_1 : \|\widetilde{\Sigma}\Theta - I_{p\times p}\|_\infty \leqslant \widetilde{\nu}_n \text{ and } \|\Theta\widetilde{\Sigma} - I_{p\times p}\|_\infty \leqslant \widetilde{\nu}_n \right\}. \tag{9}$$

The analysis of this estimator is standard. For completeness we include a proof of the following result in the supplementary materials:

Lemma 1. *Under (A1), (A3) and (A5), suppose* $\frac{\log p}{\rho_*^2 n} \to 0$. *Then with probability* $1 - o(1)$ *it holds that* $\max\{\|\widehat{\Theta}\|_{L_1}, \|\widehat{\Theta}\|_{L_\infty}\} \leqslant b_1$ *and that*

$$\max\{\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_1}, \|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_\infty}\} \leqslant 2\widetilde{\nu}_n b_0 b_1.$$

We refer to $\widehat{\Theta}$ as the modified CLIME estimator. Given the modified Dantzig estimator $\widehat{\beta}_n$ in (4) and the modified CLIME estimator we construct the de-biased estimator $\widehat{\beta}_n^u$:

$$\widehat{\beta}_n^u = \widehat{\beta}_n + \widehat{\Theta}\left(\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\widehat{\beta}_n\right). \tag{10}$$

Our next main result derives the limiting distribution of the de-biased estimator. Define the matrix $\widehat{\Upsilon}$ as:

$$\widehat{\Upsilon}_{jk} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}\sum_{t\neq j}\frac{1-\rho_t}{\rho_j\rho_t}X_{ij}^2X_{it}^2[\beta_t^*]^2, & j=k; \\ \frac{1}{n}\sum_{i=1}^{n}\sum_{t\neq j,k}\frac{1-\rho_t}{\rho_t}X_{ij}X_{ik}X_{it}^2[\beta_t^*]^2, & j\neq k, \end{cases}$$

and the matrix $\widehat{\Gamma}\in\mathbb{R}^{p\times p}$ as

$$\widehat{\Gamma} = \frac{\sigma_\varepsilon^2}{n}X^\top X + \frac{\sigma_\varepsilon^2}{n}\widetilde{D}\mathrm{diag}(X^\top X) + \widehat{\Upsilon},$$

where $\widetilde{D} = \mathrm{diag}(\frac{1}{\rho_1}-1,\cdots,\frac{1}{\rho_p}-1)$. With these definitions in place we have the following result:

Theorem 4. *Suppose that,*

$$\sigma_x^4 b_0 b_1^2\sqrt{\frac{\log^2 p}{\rho_*^4 n}}\left(\frac{\sigma_\varepsilon\sqrt{\rho_*}}{\sigma_x}+\|\beta^*\|_2\right)\left(1+\frac{s}{\lambda_{\min}b_0b_1}\right)\to 0. \tag{11}$$

then for any variable subset $S\subseteq[p]$ with constant size it holds that with probability $1-o(1)$ over the random design X,

$$\sqrt{n}\left(\widehat{\beta}_n^u-\beta^*\right)_S \xrightarrow{d} N\left(0,\left[\Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1}\right]_{SS}\right) \quad \textit{conditioned on } X.$$

Remarks:

1. We obtain the above result as a special case of a more general result. In particular, the initial estimator $\widehat{\beta}_n$ only needs to satisfy the condition that,

$$\sigma_x^2 b_0 b_1\widetilde{\nu}_n\left(\frac{\sigma_\varepsilon}{\sigma_x}\sqrt{\frac{\log p}{\rho_*}}+\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2}}+\frac{\sqrt{n}\|\widehat{\beta}_n-\beta^*\|_1}{\sigma_x^2b_0b_1}\right)\xrightarrow{p}0, \tag{12}$$

   for the conclusion of the theorem to hold.

2. It is possible to demonstrate the rate optimality of the above theorem in a certain regime. In more details, consider the case when $\Sigma_0 = I$ and the observation rates $\rho_1=\rho_2=\cdots=\rho_p=\rho_*$. Fix a single coordinate j and let $V_j := \mathrm{Var}(\sqrt{n}(\widehat{\beta}_n^u-\beta^*)_j)$ denote the rescaled mean-squared error of the j-th coordinate. By Theorem 4, when n is sufficiently large

$$V_j \xrightarrow{p}\widehat{\Gamma}_{jj}\xrightarrow{p}\frac{\sigma_\varepsilon^2}{\rho_*}+\frac{1-\rho_*}{\rho_*^2}\sum_{t\neq j}[\beta_t^*]^2\leqslant\frac{\sigma_\varepsilon^2}{\rho_*}+\frac{1-\rho_*}{\rho_*^2}\|\beta^*\|_2^2. \tag{13}$$

   Comparing this with Theorem 3, we observe that the variance V_j matches the minimax rates of coordinate-wise estimation up to a universal constant. Formally, under the additional assumption $\sigma_\varepsilon^2\gg e^{-0.5c^2(1-\rho_*)s}\|\beta^*\|_2^2$ that σ_ε is not exponentially small, we have that

$$\limsup_{p,n\to\infty}\frac{V_j^2}{\inf_{\widehat{\beta}_n}\sup_{\beta\in\mathbb{B}_2(\|\beta^*\|_2)\cap\mathbb{B}_0(s),\Sigma\in\Lambda(\gamma_0)}n\mathbb{E}|\widehat{\beta}_{nj}-\beta_j|^2}\leqslant 2C_1^{-1}(1+2c),$$

   where $C_1>0$ is the universal constant in Theorem 3.

3. Although the de-biased estimator we propose is inspired by prior work [8, 21, 36, 40] the analysis in the missing data case is complicated by the fact that estimates of both $\widehat{\Theta}$ and $\widehat{\beta}_n$ depend on the randomness induced by the missing entries. To circumvent this issue we rely on a careful argument that relates $\widehat{\Theta}$ to its deterministic counterpart Σ_0^{-1}.

4. Finally, we note that the limiting covariance depends on several unobserved quantities, most problematically the true regression vector β^* and unobserved entries of the design matrix X. We overcome these issues and provide and analyze a data-driven estimate of the limiting covariance matrix in the next section.

3.2 Data-driven approximation of the limiting covariance

To aid in the practical construction of confidence intervals we propose an estimate of the asymptotic variance and study its rates of convergence. Our estimates are constructed by replacing the unobserved design matrix X with $\widetilde{X}$ defined in (3) and the true regression vector β^* with the modified Dantzig estimate $\widehat{\beta}_n$. Formally, we define

$$\widetilde{\Gamma} = \frac{\sigma_\varepsilon^2}{n}\widetilde{X}^\top \widetilde{X} + \widetilde{\Upsilon},$$

where

$$\widetilde{\Upsilon}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t\neq j,k}(1-\rho_t)\widetilde{X}_{ij}\widetilde{X}_{ik}\widetilde{X}_{it}^2\widehat{\beta}_{nt}^2,$$

for $j, k \in \{1, \cdots, p\}$. The following theorem shows that $\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top$ is a good approximation of $\Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1}$ when n is sufficiently large:

Theorem 5. *Suppose the conclusion in Lemma 1 holds,* $\frac{\log p}{\rho_*^4 n} \to 0$ *and* $\|\widehat{\beta}_n - \beta^*\|_2 \xrightarrow{p} 0$. *Then*

$$\left\|\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top - \Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1}\right\|_\infty = O_{\mathbb{P}}\left(\frac{\sigma_x^4 b_1^2 \log^2 p}{\rho_*^2}\left\{\left(\|\beta^*\|_2^2 + \frac{\rho_*\sigma_\varepsilon^2}{\sigma_x^2}\right)\left(b_0\widetilde{\nu}_n + \sqrt{\frac{\log p}{\rho_* n}}\right) + \|\beta^*\|_2\|\widehat{\beta}_n - \beta^*\|_1\right\}\right).$$

Remark: Based on Theorems 4 and 5, an asymptotic $(1-\alpha)$ confidence interval of β_j^* can be computed as

$$\mathrm{CI}_j(\alpha) = \left[\widehat{\beta}_{nj}^u - \frac{\Phi^{-1}(1-\alpha/2)\sqrt{(\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top)_{jj}}}{\sqrt{n}}, \widehat{\beta}_{nj}^u + \frac{\Phi^{-1}(1-\alpha/2)\sqrt{(\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top)_{jj}}}{\sqrt{n}}\right], \tag{14}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the CDF of the standard Gaussian distribution. We now turn our attention to studying the finite-sample behaviour of the modified Dantzig selector and its associated confidence intervals in a variety of simulations.

4 Simulation results

In this section, we report a variety of simulation results on synthetic and semi-synthetic data aimed at assessing the modified Dantzig selector, the limiting behaviour of the de-biased estimator and the coverage of the confidence interval proposed in (14).

4.1 Synthetic data

We fix $\sigma_\varepsilon = 0.1$ and set $\Sigma_0 = \Omega^{-1}$ where Ω is chosen to be the following banded matrix:

$$\Omega_{ij} = \begin{cases} 0.5^{|i-j|} & \text{if } |i-j| \leqslant 5 \\ 0 & \text{otherwise} \end{cases}.$$

We assume a uniform observation rate $\rho_1 = \cdots = \rho_p = \rho_*$, which ranges from 0.5 to 0.9. The support set $J_0 \subset [p]$ of β^* is selected uniformly at random, with $|J_0| = 10$. β^* is then generated as $\beta_j^* \sim \text{Bernoulli}\{+1, -1\}$ independently for $j \in J_0$ and $\beta_j^* = 0$ for $j \notin J_0$. Both the modified Dantzig selector (4) and the modified CLIME estimator (9) are computed using the alternating direction method of multipliers (ADMM) algorithm.

4.1.1 Verification of asymptotic normality

We run 1000 independent realizations of our experiments and study the distributions of $\sqrt{n}(\widehat{\beta}_n^u - \beta^*)$. We plot the empirical distribution of

$$\widehat{\delta}_j = \frac{\sqrt{n}(\widehat{\beta}_{nj}^u - \beta_j^*)}{\sqrt{(\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top)_{jj}}}$$

together with the standard normal distribution. Figure 1 shows that the empirical distribution of $\widehat{\delta}_j$ agrees quite well with that of the standard normal distribution. In addition, we find that more samples are required to ensure asymptotic normality when observation rates are low (e.g., $\rho_* = 0.5$).

4.1.2 Average CI coverage and length

We calculate the average coverage and length of the constructed confidence intervals from T independent realizations, defined as

$$\text{Avgcov}(j) = \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}(\beta_{0j} \in \text{CI}_j^{(i)}(\alpha)), \quad \text{and} \quad \text{Avglen}(j) = \frac{1}{T}\sum_{i=1}^{T} \text{length}(\text{CI}_j^{(i)}(\alpha)),$$

where $\text{CI}_j(\alpha)$ is defined in (14). We also report the average coverage and length of coordinate-wise confidence intervals across a coordinate subset $J \subseteq [p]$, defined as

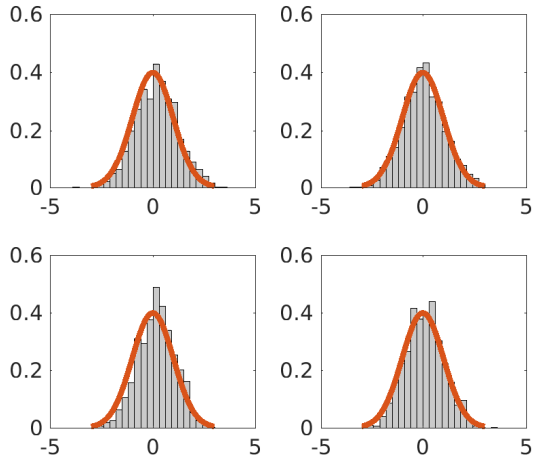
$$\text{Avgcov}(J) = \frac{1}{|J|}\sum_{j \in J} \text{Avgcov(j)} \qquad \text{and} \qquad \text{Avglen}(J) = \frac{1}{|J|}\sum_{j \in J} \text{Avglen(j)}.$$

Tables 1 summarize the results for various (n, p, ρ_*) settings.

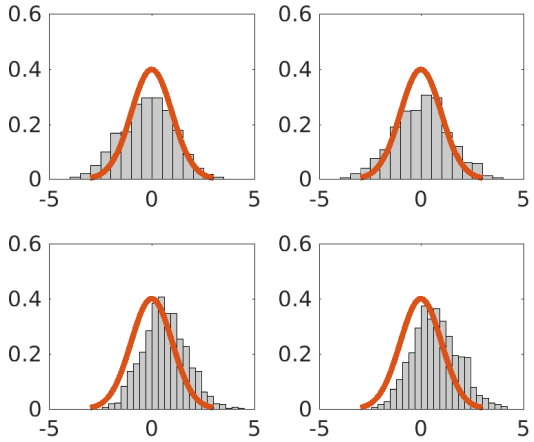
4.2 Semi-synthetic data

In this section we conduct experiments on two datasets: DNA and Madelon[2], where the distribution of the design matrices are not necessarily sub-Gaussian. The DNA data contains 2000 instances and 180 covariates, while Madelon contains 2000 data points and 500 covariates. For these two datasets, we only use their data matrix X and construct the response y according to a sparse linear regression model. Following the simulation study, we randomly remove observed covariates with

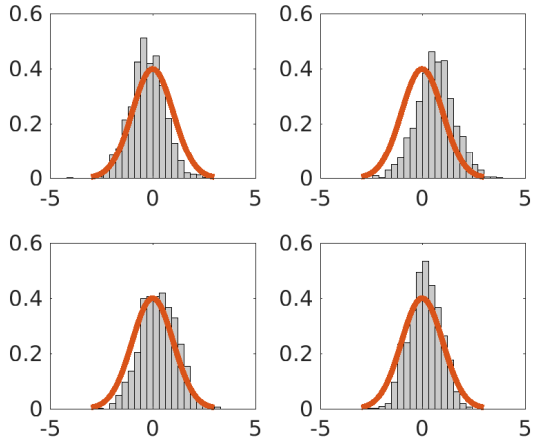
[2]Available from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



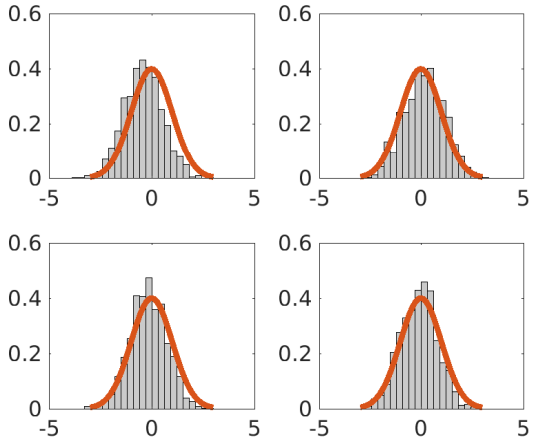
$n = 1500, p = 500, \rho_* = 0.9$



$n = 5000, p = 500, \rho_* = 0.7$



$n = 8000, p = 500, \rho_* = 0.5$



$n = 12000, p = 500, \rho_* = 0.5$

Figure 1: Empirical distribution and density of $\widehat{\delta}_j = \frac{\sqrt{n}(\widehat{\beta}^u_{nj} - \beta^*_j)}{\sqrt{(\widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top)_{jj}}}$ of 1000 independent realizations. The top row in each subfigure corresponds to two coordinates randomly chosen from J_0, and the bottom row in each subfigure corresponds to two coordinates randomly chosen from J_0^c. The red curve in each case denotes the density of the standard normal distribution.

Table 1: 95% confidence intervals for high-dimensional regression with missing data when $\rho_* \in [0.7, 0.9]$.

(n, p, ρ_*)	Random $j \in J_0$		Random $j \notin J_0$		J_0		J_0^c	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
(1000,200,0.9)	0.941	0.182	0.951	0.192	0.938	0.208	0.966	0.187
(1000,200,0.8)	0.945	0.318	0.948	0.329	0.944	0.334	0.979	0.331
(1000,200,0.7)	0.952	0.494	0.983	0.540	0.949	0.547	0.989	0.529
(1500,500,0.9)	0.931	0.155	0.966	0.170	0.945	0.183	0.971	0.158
(1500,500,0.8)	0.927	0.278	0.982	0.294	0.937	0.308	0.985	0.284
(1500,500,0.7)	0.963	0.415	0.994	0.469	0.971	0.497	0.995	0.450
(2000,1000,0.9)	0.947	0.144	0.974	0.144	0.949	0.160	0.975	0.139
(2000,1000,0.8)	0.967	0.249	0.987	0.264	0.939	0.281	0.990	0.254
(2000,1000,0.7)	0.952	0.378	0.995	0.422	0.930	0.451	0.997	0.409
(3000,2000,0.9)	0.958	0.116	0.954	0.118	0.951	0.133	0.981	0.115
(3000,2000,0.8)	0.919	0.202	0.979	0.220	0.948	0.236	0.993	0.212
(3000,2000,0.7)	0.891	0.315	0.998	0.349	0.950	0.372	0.998	0.348

Table 2: 95% confidence intervals for regression with missing data when $\rho_* = 0.5$.

(n, p, ρ_*)	Random $j \in J_0$		Random $j \notin J_0$		J_0		J_0^c	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
(1000,200,0.5)	0.928	1.051	0.998	1.223	0.942	1.384	0.999	1.194
(2000,200,0.5)	0.971	0.715	0.997	0.849	0.971	0.799	0.995	0.813
(3000,200,0.5)	0.956	0.574	0.976	0.644	0.961	0.668	0.989	0.640
(4000,200,0.5)	0.936	0.468	0.984	0.541	0.943	0.527	0.986	0.534
(1500,500,0.5)	0.986	0.795	0.978	0.911	0.756	0.954	1.000	0.896
(3000,500,0.5)	0.849	0.510	0.899	0.575	0.479	0.634	0.998	0.572
(8000,500,0.5)	0.972	0.352	0.978	0.408	0.908	0.417	0.988	0.403
(12000,500,0.5)	0.941	0.272	0.965	0.315	0.936	0.328	0.976	0.309

Table 3: 95% confidence intervals for regression with missing data on real world datasets.

$(\text{dataset}, \rho_*)$	Random $j \in J_0$		Random $j \notin J_0$		J_0		J_0^c	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
(DNA,0.9)	0.924	0.120	0.956	0.128	0.937	0.128	0.957	0.129
(DNA,0.8)	0.908	0.195	0.959	0.216	0.926	0.212	0.965	0.218
(DNA,0.7)	0.888	0.286	0.967	0.318	0.925	0.314	0.973	0.317
(DNA,0.5)	0.713	0.464	0.964	0.516	0.745	0.512	0.976	0.519
(Madelon,0.9)	0.943	0.095	0.963	0.101	0.949	0.098	0.945	0.105
(Madelon,0.8)	0.966	0.167	0.976	0.174	0.961	0.181	0.971	0.223
(Madelon,0.7)	0.962	0.229	0.977	0.236	0.956	0.253	0.977	0.261
(Madelon,0.5)	0.663	0.334	0.977	0.357	0.682	0.377	0.965	0.356

probability $1-\rho_*$, and then perform statistical inference based on the datasets with missing covariates. The performance of the constructed confidence intervals are reported in Table 3. We see that the proposed procedure produces roughly normal estimates for the parameters of interest when ρ_* is not too small, demonstrating that the estimators and confidence intervals can be robust to violations of the assumptions on the design matrix.

5 Discussion

In this paper, we studied the problems of estimation of and constructing confidence intervals for a high-dimensional regression vector when covariates are missing completely at random. In the context of estimation, in contrast to the situation in regression without missing data, we find a discrepancy between bounds obtained when Σ_0 is taken to be known and when it is unknown. We sharpen existing analyses in both these settings and develop minimax lower bounds to show that this discrepancy is unavoidable. Finally, we provide a method to construct confidence intervals in the presence of missing data through de-biasing, and study its length and coverage properties. Several important questions remain open and discuss some of these here.

Theorem 3 shows that if the population covariance Σ_0 of the design matrix X is unknown, then the mean square estimation error of a fixed component in β^* must depend quadratically on the observation ratio ρ_*. We conjecture that such results also hold for the estimation error of the entire regression model β^* as well. More specifically, we conjecture that under suitable finite-sample conditions,

$$\inf_{\widehat{\beta}_n} \sup_{\substack{\beta^* \in \mathbb{B}_2(M) \cap \mathbb{B}_0(s) \\ \Sigma_0 \in \Lambda(\gamma_0)}} \mathbb{E}\|\widehat{\beta}_n - \beta^*\|_2^2 \geqslant C_1' \cdot \max\left\{ \frac{\sigma_\varepsilon^2 s \log p}{\rho_* n}, \min\left(\frac{1-\rho_*}{1+2c} M^2, e^{c^2(1-\rho_*)s} \sigma_\varepsilon^2 \right) \frac{s \log p}{\rho_*^2 n} \right\}.$$

Establishing such a bound however requires a generalization of our lower bound construction in a novel fashion. In particular, our current construction relies on a carefully designed packing set of covariance matrices that do not "leak information" unless both X_1 and X_j (for a fixed j) are observed, and extending this construction more generally appears to be challenging.

Our upper bounds for both estimation and inference focus on a large-sample regime when the Bernstein-type inequalities we use result in sub-Gaussian behaviour. In problems with missing data, the natural plug-in estimators, of the covariance matrix for instance, exhibit different rates of convergence in the small-sample regime. Understanding the tightness of our bounds in this small-sample regime would be interesting.

For inference we use sparsity assumptions that ensure that the precision matrix Σ_0^{-1} is estimable, which are restrictive as the precision matrix is a nuisance parameter. In the fully observed setting weaker assumptions are used for instance in [21] at the cost of asymptotic efficiency of the average length of the resulting confidence interval. In the missing data setting however the dependence between the estimates $\widehat{\Theta}$ and $\widehat{\beta}_n$ caused due to the missingness is challenging to deal with directly. Instead, we use arguments that relate $\widehat{\Theta}$ to its deterministic population counterpart Σ_0^{-1}. Understanding the extent to which this dependence can be circumvented, and weakening the assumptions required on the nuisance parameter Σ_0^{-1} remains an open question.

6 Proofs

In this section, we turn to the proofs of our main theorems. We include in the main text the main body of the proofs deferring more technical aspects to the supplementary material.

6.1 Additional notation

We use the matrix R to denote the missingness pattern, i.e. define:

$$R_{ij} = \begin{cases} 0, & \text{if } X_{ij} = \star, \\ 1, & \text{otherwise.} \end{cases}$$

In order to compactly derive and state concentration bounds for the case when Σ_0 is known and unknown we will use the following additional notation.

Definition 1. *Let A, B be random or deterministic square matrices of the same size and ε be a random vector of i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ components. Let $\varphi_{u,v}(A, B; \log N)$, $\varphi_{u,\infty}(A, B; \log N)$, $\varphi_{\varepsilon,\infty}(A)$ be terms such that, with probability $1 - o(1)$ as $n \to \infty$, for all subset $\mathcal{S}$ of vectors with $|\mathcal{S}| \leqslant N$, the following hold for all $u, v \in \mathcal{S}$:*

$$\begin{aligned} \left|u^\top (A - B) v\right| &\leqslant \varphi_{u,v}(A, B; \log N) \cdot \|u\|_2 \|v\|_2; \\ \left\|A^\top \varepsilon\right\|_\infty &\leqslant \varphi_{\varepsilon,\infty}(A) \cdot \sigma_\varepsilon. \end{aligned}$$

Note that $\varphi_{u,v}(\cdot, \cdot)$ is symmetric and satisfies the triangle inequality. Also, infinity norms like $\|A - B\|_\infty$ or $\|(A - B)u\|_\infty$ for a fixed u can be upper bounded by $\varphi_{u,v}(A, B; O(\log \dim(A)))$, by considering the set of unit vectors $\{e_1, \cdots, e_{\dim(A)}\}$.

6.2 Proof of Theorem 1

We need the following two concentration lemmas, which are proved in the supplementary material.

Lemma 2. *Denote random matrices $A^{(\ell)}$, $\ell \in \{0, 1, 2\}$ as $A^{(0)} = \widehat{\Sigma}$, $A^{(1)} = \frac{1}{n}\widetilde{X}^\top X$ and $A^{(2)} = \widetilde{\Sigma}$, respectively. Then for $\ell \in \{0, 1, 2\}$:*

$$\varphi_{u,v}\left(A^{(\ell)}, \Sigma_0; \log N\right) \leqslant O\left(\sigma_x^2 \max\left\{\frac{\log N}{\rho_*^{1.5\ell} n}, \sqrt{\frac{\log N}{\rho_*^{\ell} n}}\right\}\right).$$

Lemma 3. *If $\frac{\log p}{\rho_* n} \to 0$ then $\varphi_{\varepsilon,\infty}(\frac{1}{n}\widetilde{X}) \leqslant O(\sigma_x \sqrt{\frac{\log p}{\rho_* n}})$.*

We present the following lemma. Its proof is given in the supplementary material.

Lemma 4. *Suppose* $\frac{\log p}{\rho_*^4 n} \to 0$ *for* $\widehat{\beta}_n$ *or* $\frac{\log p}{\rho_*^2 n} \to 0$ *for* $\check{\beta}_n$*, and let* $J_0 = \mathrm{supp}(\beta^*)$ *be the support of* β^**. If* $\widetilde{\lambda}_n \geqslant \Omega\{\sigma_x\sqrt{\frac{\log p}{n}}(\frac{\sigma_x\|\beta^*\|_2}{\rho_*} + \frac{\sigma_\varepsilon}{\sqrt{\rho_*}})\}$ *and* $\check{\lambda}_n \geqslant \Omega\{\sigma_x\sqrt{\frac{\log p}{\rho_* n}}(\sigma_x\|\beta^*\|_2 + \sigma_\varepsilon)\}$*, then with probability* $1 - o(1)$ *we have that*

1. $\|(\widehat{\beta}_n - \beta^*)_{J_0^c}\|_1 \leqslant \|(\widehat{\beta}_n - \beta^*)_{J_0}\|_1$;
2. $\|(\check{\beta}_n - \beta^*)_{J_0^c}\|_1 \leqslant \|(\check{\beta}_n - \beta^*)_{J_0}\|_1$.

Definition 2 (Restricted eigenvalue condition)**.** *A* $p \times p$ *matrix* A *is said to satisfy* $\mathrm{RE}(s, \phi_{\min})$ *if for all* $J \subseteq [p]$*,* $|J| \leqslant s$ *the following holds:*

$$\inf_{h \neq 0, \|h_{J^c}\|_1 \leqslant \|h_J\|_1} \frac{h^\top A h}{h^\top h} \;\geqslant\; \phi_{\min}.$$

The following lemma is proved in the supplementary material.

Lemma 5. *Suppose* $\frac{\sigma_x^4 s \log(\sigma_x \log p/\rho_*)}{\rho_*^3 \lambda_{\min}^2 n} \to 0$*. Then with probability* $1 - o(1)$*, the sample covariance for the missing data problem* $\widetilde{\Sigma}$ *satisfies* $\mathrm{RE}(s, (1 - o(1))\lambda_{\min}(\Sigma_0))$.

We are now ready to prove Theorem 1 that establishes the rate of convergence of the modified Dantzig selector estimators. We consider $\widehat{\beta}_n$ first. Define $\widetilde{\lambda}_n\mu = \frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\widehat{\beta}_n$. By $y = X\beta^* + \varepsilon$, we have that

$$\widetilde{\Sigma}(\widehat{\beta}_n - \beta^*) = \left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^* + \left(\Sigma_0 - \widetilde{\Sigma}\right)\beta^* - \widetilde{\lambda}_n\mu + \frac{1}{n}\widetilde{X}^\top\varepsilon.$$

Multiply both sides by $(\widehat{\beta}_n - \beta^*)$ and apply Hölder's inequality:

$$\begin{aligned}
&(\widehat{\beta}_n - \beta^*)^\top\widetilde{\Sigma}(\widehat{\beta}_n - \beta^*) \\
&\leqslant \|\widehat{\beta}_n - \beta^*\|_1 \left\{\left\|\left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^*\right\|_\infty + \left\|\left(\Sigma_0 - \widetilde{\Sigma}\right)\beta^*\right\|_\infty + \widetilde{\lambda}_n\|\mu\|_\infty + \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty\right\} \\
&\leqslant \|\widehat{\beta}_n - \beta^*\|_1 \cdot O_{\mathbb{P}}\left\{\varphi_{u,v}\left(\frac{1}{n}\widetilde{X}^\top X, \Sigma_0; \log p\right)\|\beta^*\|_2 + \varphi_{u,v}\left(\widetilde{\Sigma}, \Sigma_0; \log p\right)\|\beta^*\|_2 + \widetilde{\lambda}_n + \varphi_{\varepsilon,\infty}\left(\frac{1}{n}\widetilde{X}\right)\sigma_\varepsilon\right\} \\
&\leqslant \|\widehat{\beta}_n - \beta^*\|_1 \cdot O_{\mathbb{P}}\left\{\sigma_x^2\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2 n}} + \sigma_x\sigma_\varepsilon\sqrt{\frac{\log p}{\rho_* n}} + \widetilde{\lambda}_n\right\}.
\end{aligned}$$

Here the last inequality is due to Lemmas 2 and 3. Suppose $\frac{\sigma_x^4 s \log(\sigma_x p/\rho_*)}{\rho_*^4 \lambda_{\min}^2 n} \to 0$ and $\widetilde{\lambda}_n$ is appropriately set as in Lemma 4. We then have

$$\|\widehat{\beta}_n - \beta^*\|_1 \leqslant 2\|(\widehat{\beta}_n - \beta^*)_{J_0}\|_1 \leqslant 2\sqrt{s}\|\widehat{\beta}_n - \beta^*\|_2 \tag{15}$$

by Lemma 4 and

$$(\widehat{\beta}_n - \beta^*)^\top\widetilde{\Sigma}(\widehat{\beta}_n - \beta^*) \geqslant (1 - o(1))\lambda_{\min}\|\widehat{\beta}_n - \beta^*\|_2^2$$

by Lemma 5. Chaining all inequalities we get

$$\begin{aligned}
\|\widehat{\beta}_n - \beta^*\|_2 &\leqslant O_{\mathbb{P}}\left(\frac{\sqrt{s}}{\lambda_{\min}}\left\{\sigma_x^2\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2 n}} + \sigma_x\sigma_\varepsilon\sqrt{\frac{\log p}{\rho_* n}} + \widetilde{\lambda}_n\right\}\right) \\
&\leqslant O_{\mathbb{P}}\left(\frac{\sqrt{s}}{\lambda_{\min}}\left\{\sigma_x^2\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2 n}} + \sigma_x\sigma_\varepsilon\sqrt{\frac{\log p}{\rho_* n}}\right\}\right).
\end{aligned}$$

The ℓ_1 norm error bound $\|\widehat{\beta}_n - \beta^*\|_1$ can be easily obtained by the fact that $\|\widehat{\beta}_n - \beta^*\|_1 \leqslant 2\sqrt{s}\|\widehat{\beta}_n - \beta^*\|_2$ as shown in Eq. (15).

Finally, consider $\check{\mu}_n$ and define $\check{\lambda}_n\check{\mu} = \frac{1}{n}\widetilde{X}^\top y - \Sigma_0\check{\beta}_n$. Note that $\|\check{\delta}\|_\infty \leqslant 1$ and

$$\Sigma_0(\check{\beta}_n - \beta^*) = \left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^* - \check{\lambda}_n\check{\mu} + \frac{1}{n}\widetilde{X}^\top\varepsilon.$$

Note in addition that $(\check{\beta}_n - \beta^*)^\top\Sigma_0(\check{\beta}_n - \beta^*) \geqslant \lambda_{\min}\|\check{\beta}_n - \beta^*\|_2^2$ by Assumption (A2). Subsequently, the same line of argument for $\widehat{\beta}_n$ yields

$$\begin{aligned}\|\check{\beta}_n - \beta^*\|_2 &\leqslant \frac{2\sqrt{s}}{\lambda_{\min}} \cdot O_{\mathbb{P}}\left\{\varphi_{u,v}\left(\frac{1}{n}\widetilde{X}^\top X, \Sigma_0; \log p\right)\|\beta^*\|_2 + \widetilde{\lambda}_n + \varphi_{\varepsilon,\infty}\left(\frac{1}{n}\widetilde{X}\right)\sigma_\varepsilon\right\} \\ &\leqslant O_{\mathbb{P}}\left\{\left(\sigma_x^2\|\beta^*\|_2 + \sigma_x\sigma_\varepsilon\right)\sqrt{\frac{s\log p}{\lambda_{\min}^2\rho_* n}}\right\}.\end{aligned}$$

6.3 Proof of Theorem 2

We consider the worst case with equal observation rates across covariates: $\rho_1 = \cdots = \rho_p = \rho_*$ and use Fano's inequality (Lemma 12) to establish the minimax lower bound in Theorem 2. Without loss of generality we shall restrain ourselves to even p and $s/2$ scenarios. Construct hypothesis β as

$$\beta = (\underbrace{a, \cdots, a}_{\text{repeat } s/2 \text{ times}}, \underbrace{0, \pm\delta, 0, \cdots, \pm\delta, 0}_{\text{exactly } s/2 \text{ copies of } \delta}), \tag{16}$$

where $\delta \to 0$ is some parameter to be chosen later and $a = \sqrt{\frac{2M^2}{s} - \delta^2}$ is carefully chosen so that $\|\beta\|_2 = M$. Clearly $\beta \in \mathbb{B}_2(M) \cap \mathbb{B}_0(s)$. Let $d_H(\beta, \beta') = \sum_{j=1}^p I[\beta_j \neq \beta_j']$ be the Hamming distance between β and β'. The following lemma shows that it is possible to construct a large hypothesis classes where any two models in the hypothesis class are far away under the Hamming distance:

Lemma 6 ([31], Lemma 4)**.** *Define $\mathcal{H} = \{z \in \{-1, 0, +1\}^p : \|z\|_0 = s\}$. For p, s even and $s < 2p/3$, there exists a subset $\widetilde{\mathcal{H}} \subseteq \mathcal{H}$ with cardinality $|\widetilde{\mathcal{H}}| \geqslant \exp\{\frac{s}{2}\log\frac{p-s}{s/2}\}$ such that $\rho_H(z, z') \geqslant s/2$ for all dinstinct $s, s' \in \widetilde{\mathcal{H}}$.*

This does not affect the minimax lower bound to be proved. Using the above lemma and under the condition that $s \leqslant 4p/5$, one can construct Θ consisting of hypothesis of the form in Eq. (16) such that $\log|\Theta| \asymp s\log(p/s)$ and $\|\beta - \beta'\|_2 \geqslant \sqrt{s/4}\delta$ for all distinct $\beta, \beta' \in \Theta$. It remains to evaluate the KL divergence between P_β and $P_{\beta'}$.

Let x_{obs} and x_{mis} denote the observed and missing covariates of a particular data point and let β_{obs}, β_{mis} be the corresponding partition of coordinates of β. The likelihood of x_{obs} and y can be obtained by integrating out x_{mis} (assuming there are q coordinates that are observed):

$$\begin{aligned}p(y, x_{\text{obs}}; \beta) &= \rho_*^q(1-\rho_*)^{p-q}\int \mathcal{N}_p(x_{\text{obs}}, x_{\text{mis}}; 0, I)\mathcal{N}(y - (x_{\text{obs}}^\top\beta_{\text{obs}} - x_{\text{mis}})^\top\beta_{\text{mis}}; 0, \sigma_\varepsilon^2)\mathrm{d}x_{\text{mis}} \\ &= p(x_{\text{obs}}) \cdot \frac{1}{\sqrt{2\pi(\sigma_\varepsilon^2 + \|\beta_{\text{mis}}\|_2^2)}}\exp\left\{-\frac{(y - x_{\text{obs}}^\top\beta_{\text{obs}})^2}{2(\sigma_\varepsilon^2 + \|\beta_{\text{mis}}\|_2^2)}\right\}.\end{aligned}$$

Here $\mathcal{N}$ and $\mathcal{N}_p$ denote the univariate and multivariate Normal distributions. Note that $p(x_{\rm obs})$ does not depend on β. Subsequently,

$$\begin{aligned}
\mathrm{KL}(P_\beta \| P_{\beta'}) &= \mathbb{E}_{\beta,\rho_*} \log \frac{p(y, x_{\rm obs}; \beta')}{p(y, x_{\rm obs}; \beta)} \\
&= \mathbb{E}_{\beta,\rho_*} \left\{ \frac{1}{2} \log \frac{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2}{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2} + \frac{1}{2} \left[\frac{(y - x_{\rm obs}^\top \beta'_{\rm obs})^2}{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2} - \frac{(y - x_{\rm obs}^\top \beta_{\rm obs})^2}{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2} \right] \right\} \\
&= \mathbb{E}_{\rho_*} \left\{ \frac{1}{2} \log \frac{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2}{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2} + \frac{1}{2} \left[\frac{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2 + \|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2}{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2} - 1 \right] \right\} \\
&\overset{(a)}{\leqslant} \mathbb{E}_{\rho_*} \left\{ \frac{1}{2} \left[\frac{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2}{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2} + \frac{\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2}{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2} \right] - 1 + \frac{1}{2} \frac{\|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2}{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2} \right\} \\
&= \mathbb{E}_{\rho_*} \left\{ \frac{1}{2} \frac{\left(\|\beta'_{\rm mis}\|_2^2 - \|\beta_{\rm mis}\|_2^2\right)^2}{(\sigma_\varepsilon^2 + \|\beta_{\rm mis}\|_2^2)(\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2)} + \frac{1}{2} \frac{\|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2}{\sigma_\varepsilon^2 + \|\beta'_{\rm mis}\|_2^2} \right\}. \qquad (17)
\end{aligned}$$

Here for (a) we apply the inequality that $\log(1+x) \leqslant x$ for all $x > 0$. For some constant $c \in (0, 1/2)$, define $\mathcal{E}(c)$ as the event that at least $\frac{1-\rho_*}{1+2c}$ portion of the first $s/2$ coordinates in x are missing. By Chernoff bound, [3] $\Pr[\mathcal{E}(c)] \geqslant 1 - e^{-c^2(1-\rho_*)s}$. Note that under $\mathcal{E}(c)$, $\|\beta_{\rm mis}\|_2^2, \|\beta'_{\rm mis}\|_2^2 \geqslant \frac{(1-\rho_*)s}{2(1+2c)} a^2$ almost surely. Subsequently,

$$\begin{aligned}
\mathrm{KL}(P_\beta \| P_{\beta'}) &\leqslant \frac{1}{2} \frac{\mathbb{E}_{\rho_*|\mathcal{E}(c)} \left(\|\beta'_{\rm mis}\|_2^2 - \|\beta_{\rm mis}\|_2^2\right)^2}{\left(\sigma_\varepsilon^2 + \frac{(1-\rho_*)s}{2(1+2c)} a^2\right)^2} + \frac{1}{2} \frac{\mathbb{E}_{\rho_*|\mathcal{E}(c)} \|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2}{\sigma_\varepsilon^2 + \frac{(1-\rho_*)s}{2(1+2c)} a^2} \\
&+ e^{-c^2(1-\rho_*)s} \left[\frac{1}{2} \frac{\mathbb{E}_{\rho_*|\overline{\mathcal{E}}(c)} \left(\|\beta'_{\rm mis}\|_2^2 - \|\beta_{\rm mis}\|_2^2\right)^2}{\sigma_\varepsilon^4} + \frac{1}{2} \frac{\mathbb{E}_{\rho_*|\overline{\mathcal{E}}(c)} \|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2}{\sigma_\varepsilon^2} \right].
\end{aligned}$$

Because β and β' are identical in the first $s/2$ coordinates, both $\|\beta'_{\rm mis}\|_2^2 - \|\beta_{\rm mis}\|_2^2$ and $\|\beta_{\rm obs} - \beta'_{\rm obs}\|_2^2$ are independent of $\mathcal{E}(c)$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\rho_*} \left(\|\beta'_{\rm mis}\|_2^2 - \|\beta_{\rm mis}\|_2^2\right)^2 = \mathbb{E}_{\rho_*} \left(\|\beta'_{{\rm mis},>s/2}\|_2^2 - \|\beta_{{\rm mis},>s/2}\|_2^2\right)^2 &\leqslant 4(1-\rho_*)^2 s^2 \delta^4; \\
\mathbb{E}_{\rho_*} \|\beta'_{\rm obs} - \beta_{\rm obs}\|_2^2 = \mathbb{E}_{\rho_*} \|\beta'_{{\rm obs},>s/2} - \beta_{{\rm obs},>s/2}\|_2^2 &\leqslant 2\rho_* s \delta^2.
\end{aligned}$$

Here $\beta_{\cdot,>s/2}$ denote the $\beta_\cdot$ vector without its first $s/2$ coordinates, and in both inequalities we note by construction that $\|\beta_{>s/2}\|_0, \|\beta'_{>s/2}\|_0 \leqslant s/2$. Because $a^2 = \frac{2M^2}{s} - \delta^2$, we have that $\frac{(1-\rho_*)s}{2(1+2c)} a^2 = \frac{1-\rho_*}{1+2c} M^2 - \frac{(1-\rho_*)}{2(1+2c)} s\delta^2$. For now assume that $\frac{1-\rho_*}{1+2c} s\delta^2 \ll \sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c} M^2$, which then implies $\sigma_\varepsilon^2 + \frac{(1-\rho_*)s}{2(1+2c)} a^2 \geqslant \frac{1}{2} \left(\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c} M^2\right)$. We will justify this assumption at the end of this proof. Combining all inequalities we have

$$\mathrm{KL}(P_\beta \| P_{\beta'}) \leqslant \frac{8(1-\rho_*)^2 s^2 \delta^4}{\left(\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c} M^2\right)^2} + \frac{2\rho_* s \delta^2}{\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c} M^2} + e^{-c^2(1-\rho_*)s} \left[\frac{2(1-\rho_*)^2 s^2 \delta^4}{\sigma_\varepsilon^4} + \frac{\rho_* s \delta^2}{\sigma_\varepsilon^2} \right].$$

Let P_β^n and $P_{\beta'}^n$ be the distribution of n i.i.d. samples parameterized by β and β', respectively. Because the samples are i.i.d., we have that $\mathrm{KL}(P_\beta^n \| P_{\beta'}^n) = n\mathrm{KL}(P_\beta \| P_{\beta'})$. On the other hand,

[3] If $X_1, \cdots, X_n$ are i.i.d. random variables taking values in $\{0, 1\}$ then $\Pr[\frac{1}{n} \sum_{i=1}^n X_i < (1-\delta)\mu] \leqslant \exp\{-\frac{\delta^2 \mu}{2}\}$ for $0 < \delta < 1$, where $\mu = \mathbb{E}X$.

because $\log|\Theta| \asymp s\log(p/s)$, to ensure $1 - \frac{\mathrm{KL}(P_\beta^n \| P_{\beta'}^n) + \log 1/2}{\log|\Theta|} \geqslant \Omega(1)$ we only need to show $\mathrm{KL}(P_\beta^n \| P_{\beta'}^n) \asymp s\log(p/s)$, which is implied by

$$\begin{aligned}
\frac{(1-\rho_*)^2 s^2\delta^4}{\left(\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2\right)^2} \asymp \frac{s\log(p/s)}{n} \quad &\Longleftarrow \quad \delta^2 \asymp \left(\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2\right)\sqrt{\frac{\log(p/s)}{(1-\rho_*)^2 sn}};\\
\frac{\rho_* s\delta^2}{\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2} \asymp \frac{s\log(p/s)}{n} \quad &\Longleftarrow \quad \delta^2 \asymp \left(\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2\right)\frac{\log(p/s)}{\rho_* n};\\
e^{-c^2(1-\rho_*)s}\frac{(1-\rho_*)^2 s^2\delta^4}{\sigma_\varepsilon^4} \asymp \frac{s\log(p/s)}{n} \quad &\Longleftarrow \quad \delta^2 \asymp e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\sqrt{\frac{\log(p/s)}{(1-\rho_*)^2 sn}};\\
e^{-c^2(1-\rho_*)s}\frac{\rho_* s\delta^2}{\sigma_\varepsilon^2} \quad &\Longleftarrow \quad \delta^2 \asymp e^{c^2(1-\rho_*)s}\sigma_\varepsilon^2\frac{\log(p/s)}{\rho_* n}.
\end{aligned}$$

Combining all terms we have that

$$\delta^2 \asymp \min\left\{\sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2, e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\right\} \cdot \min\left\{\sqrt{\frac{\log(p/s)}{(1-\rho_*)^2 sn}}, \frac{\log(p/s)}{\rho_* n}\right\}. \tag{18}$$

The bound for $\|\beta - \beta'\|_2^2$ can then be obtained by $\|\beta - \beta'\|_2^2 \geqslant \frac{s}{4}\delta^2$.

The final part of the proof is to justify the assumption that $\frac{1-\rho_*}{1+2c}s\delta^2 \ll \sigma_\varepsilon^2 + \frac{1-\rho_*}{1+2c}M^2$. Invoking Eq. (18), the assumption is valid if $\frac{1-\rho_*}{1+2c}\max\left\{\sqrt{\frac{s\log(p/s)}{(1-\rho_*)^2 n}}, \frac{s\log(p/s)}{\rho_* n}\right\} \to 0$, which holds if $\frac{s\log(p/s)}{\rho_* n} \to 0$.

6.4 Proof of Theorem 3

We again take $\rho_1 = \cdots = \rho_p = \rho_*$. The first term $\frac{\sigma_\varepsilon^2}{\rho_* n}$ in the minimax lower bound is trivial to establish: consider $\beta^* = \delta e_j$ and $\beta_1 = -\delta e_j$ with $\Sigma_0 = \Sigma_1 = I$. By Eq. (17), we have that

$$\mathrm{KL}(P_{\beta^*}^n \| P_{\beta_1}^n) = n \cdot \mathrm{KL}(P_{\beta^*} \| P_{\beta_1}) \leqslant \frac{2\rho_* n\delta^2}{\sigma_\varepsilon^2}.$$

Equating $\mathrm{KL}(P_{\beta^*}^n \| P_{\beta_1}^n) \asymp O(1)$ we have that $\delta^2 \asymp \frac{\sigma_\varepsilon^2}{\rho_* n}$. Because $\frac{\sigma_\varepsilon^2}{M^2\rho_* n} \to 0$, we know that $\beta^*, \beta_1 \in \mathbb{B}_2(M) \cap \mathbb{B}_0(1)$ when n is sufficiently large. Invoking Le Cam's method (Lemma 13) with $|\beta_{0j} - \beta_{1j}|^2 = 4\delta^2 \asymp \frac{\sigma_\varepsilon^2}{\rho_* n}$ we prove the desired minimax lower bound of $\frac{\sigma_\varepsilon^2}{\rho_* n}$.

We next focus on the second term in the minimax lower bound that involves $1/\rho_*^2 n$. Without loss of generality assume $j > s-1$. Construct two hypothesis (β^*, Σ_0) and (β_1, Σ_1) as follows:

$$\begin{aligned}
\beta^* &= (\underbrace{\frac{\widetilde{a}}{\sqrt{s-2}}, \cdots, \frac{\widetilde{a}}{\sqrt{s-2}}}_{\text{repeat } s-2 \text{ times}}, \widetilde{a}, \underbrace{0, \cdots, 0, \widetilde{a}\gamma, 0, \cdots, 0}_{\beta_{0j} = \widetilde{a}\gamma}), \qquad \Sigma_0 = I_{p\times p} - \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top);\\
\beta_1 &= (\underbrace{\frac{\widetilde{a}}{\sqrt{s-2}}, \cdots, \frac{\widetilde{a}}{\sqrt{s-2}}}_{\text{repeat } s-2 \text{ times}}, \widetilde{a}, \underbrace{0, \cdots, 0, -\widetilde{a}\gamma, 0, \cdots, 0}_{\beta_{0j} = -\widetilde{a}\gamma}), \qquad \Sigma_1 = I_{p\times p} + \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top).
\end{aligned}$$

Here $\gamma \to 0$ is some parameter to be determined later and $\widetilde{a}$ is set to $\widetilde{a} = \sqrt{\frac{M^2}{2+\gamma^2}}$ to ensure that $\|\beta^*\|_2 = \|\beta_1\|_2 = M$. It is immediate by definition that $\beta^*, \beta_1 \in \mathbb{B}_2(M) \cap \mathbb{B}_0(s)$. In addition, by

Gershgorin circle theorem all eigenvalues of Σ_0 and Σ_1 lie in $[1-\gamma, 1+\gamma]$. As $\gamma \to 0$, it holds that $\Sigma_0, \Sigma_1 \in \Lambda(\gamma_0)$ for any constant $\gamma_0 \in (0, 1/2)$ when n is sufficiently large. A finite-sample statement of this fact is given at the end of the proof.

Unlike the identity covariance case, the likelihood $p(y, x_{\text{obs}}; \beta, \Sigma)$ for incomplete observations are complicated when Σ has non-zero off-diagonal elements. The following lemma gives a general characterization of the likelihood when $\beta \neq 0$. Its proof is given in the supplementary material.

Lemma 7. *Partition the covariance* Σ *as* $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, *where* Σ_{11} *corresponds to* x_{obs} *and* Σ_{22} *corresponds to* x_{mis}. *Define* $\Sigma_{22:1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. *Let* $q = \dim(\Sigma_{11})$ *be the number of observed covariates. Then*

$$\begin{aligned} p(y, x_{\text{obs}}; \beta, \Sigma) \quad = \quad & \rho_*^q (1-\rho_*)^{p-q} \cdot \frac{1}{\sqrt{(2\pi)^q |\Sigma_{11}|}} \exp\left\{ -\frac{1}{2} x_{\text{obs}}^\top \Sigma_{11}^{-1} x_{\text{obs}} \right\} \\ & \cdot \frac{1}{\sqrt{2\pi(\sigma_\varepsilon^2 + \beta_{\text{mis}}^\top \Sigma_{22:1} \beta_{\text{mis}})}} \exp\left\{ -\frac{(y - x_{\text{obs}}^\top \beta_{\text{obs}} - \beta_{\text{mis}}^\top \Sigma_{21} \Sigma_{11}^{-1} x_{\text{obs}})^2}{2(\sigma_\varepsilon^2 + \beta_{\text{mis}}^\top \Sigma_{22:1} \beta_{\text{mis}})} \right\}. \end{aligned}$$

We now present the following lemma, which is key to establish the $1/\rho_*^2$ rate in the minimax lower bound. Its proof is given in the supplementary material.

Lemma 8. $p(y, x_{\text{obs}}; \beta^*, \Sigma_0) = p(y, x_{\text{obs}}; \beta_1, \Sigma_1)$ *unless* both x_{s-1} *and* x_j *are observed.*

Let P_0 and P_1 denote the distributions parameterized by (β^*, Σ_0) and (β_1, Σ_1), respectively. Let $\mathcal{A}$ denote the event that both x_{s-1} and x_j are observed. By Lemma 8, we have that

$$\text{KL}(P_0 \| P_1) = \Pr[\mathcal{A}] \mathbb{E}_0 \left[\log \frac{p(y, x_{\text{obs}}; \beta^*, \Sigma_0)}{p(y, x_{\text{obs}}; \beta_1, \Sigma_1)} \middle| \mathcal{A} \right] = \rho_*^2 \mathbb{E}_0 \left[\log \frac{p(y, x_{\text{obs}}; \beta^*, \Sigma_0)}{p(y, x_{\text{obs}}; \beta_1, \Sigma_1)} \middle| \mathcal{A} \right].$$

Suppose $\Sigma_0 = [\Sigma_{011}\ \Sigma_{012}; \Sigma_{021}\ \Sigma_{022}]$ and $\Sigma_1 = [\Sigma_{111}\ \Sigma_{112}; \Sigma_{121}\ \Sigma_{122}]$ are partitioned in the same way as in Lemma 7. Conditioned on the event $\mathcal{A}$, we have that

$$\begin{aligned} \Sigma_{022} &= \Sigma_{122} = I_{(p-q)\times(p-q)}, \\ \Sigma_{012} &= \Sigma_{021}^\top = \Sigma_{112} = \Sigma_{121}^\top = 0_{q\times(p-q)}, \\ \Sigma_{011} &= I_{q\times q} - \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top), \\ \Sigma_{111} &= I_{q\times q} + \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top), \end{aligned}$$

and by Lemma 14, we have that

$$\Sigma_{011}^{-1} = I + \frac{\gamma^2}{1-\gamma^2}(e_{s-1}e_{s-1}^\top + e_j e_j^\top) + \frac{\gamma}{1-\gamma^2}(e_{s-1}e_j^\top + e_j e_{s-1}^\top)$$

and

$$\Sigma_{111}^{-1} = I + \frac{\gamma^2}{1-\gamma^2}(e_{s-1}e_{s-1}^\top + e_j e_j^\top) - \frac{\gamma}{1-\gamma^2}(e_{s-1}e_j^\top + e_j e_{s-1}^\top).$$

In addition, $\det(\Sigma_{011}) = \det(\Sigma_{111}) = 1-\gamma^2$. Note also that $\Sigma_{022:1} = \Sigma_{122:1} = I_{(p-q)\times(p-q)}$ and hence $\beta_{0\text{mis}}^\top \Sigma_{022:1} \beta_{0\text{mis}} = \beta_{1\text{mis}}^\top \Sigma_{122:1} \beta_{1\text{mis}}$ because $\|\beta_{0\text{mis}}\|_2^2 = \|\beta_{1\text{mis}}\|_2^2$ regardless of which

covariates are missing. Define $x_{\text{obs},<s} = \{x_j : x_j \text{ is observed}, j < s\}$ and $\beta_{\text{obs},<s} = \{\beta_j : x_j \text{ is observed}, j < s\}$. Subsequently, invoking Lemma 7 we get

$$\begin{aligned}
\mathbb{E}_{0|\mathcal{A}}\left[\log\frac{P_0}{P_1}\right] &= -\frac{2\gamma}{1-\gamma^2}\mathbb{E}_0[x_{s-1}x_j] - \mathbb{E}_{0|\mathcal{A}}\left\{\frac{1}{2}\frac{(y-x_{\text{obs}}^\top\beta_{0\text{obs}})^2-(y-x_{\text{obs}}^\top\beta_{1\text{obs}})^2}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \\
&\overset{(a)}{=} \frac{2\gamma^2}{1-\gamma^2} + \mathbb{E}_{0|\mathcal{A}}\left\{\frac{x_j(\beta_{0j}-\beta_{1j})(y-x_{\text{obs},<s}^\top\beta_{0\text{obs},<s})}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \\
&\overset{(b)}{=} \frac{2\gamma^2}{1-\gamma^2} + \mathbb{E}_{0|\mathcal{A}}\left\{\frac{x_j(\beta_{0j}-\beta_{1j})(x_{\text{mis},<s}^\top\beta_{0\text{mis},<s}+x_j\beta_{0j}+\varepsilon)}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \\
&\overset{(c)}{=} \frac{2\gamma^2}{1-\gamma^2} + \mathbb{E}_{R|\mathcal{A}}\left\{\frac{\beta_{0j}(\beta_{0j}-\beta_{1j})\mathbb{E}_0[x_j^2]+(\beta_{0j}-\beta_{1j})\mathbb{E}_{0|R}[x_j(x_{\text{mis},<s-1}^\top\beta_{0\text{mis},<s-1}+\varepsilon)]}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \\
&= \frac{2\gamma^2}{1-\gamma^2} + \mathbb{E}_{R|\mathcal{A}}\left\{\frac{\beta_{0j}(\beta_{0j}-\beta_{1j})\mathbb{E}_0[x_j^2]}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \\
&= \frac{2\gamma^2}{1-\gamma^2} + \mathbb{E}_{R|\mathcal{A}}\left\{\frac{2\widetilde{a}^2\gamma^2}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\}.
\end{aligned}$$

Here (a) is due to $\beta_{0\text{obs},<s} = \beta_{1\text{obs},<s}$ and $\beta_{0j}^2 = \beta_{1j}^2$, and (b) is because $\beta_{0k} = 0$ for all $k \geqslant s$ except for $k = j$. Note also that under $\mathcal{A}$, x_j is observed and hence β_{0j} always belongs to $\beta_{0\text{obs}}$. For (c), note that x_{s-1} is observed under $\mathcal{A}$ and x_j is independent of $x_{<s-1}$ and ε conditioned on R, thanks to the missing completely at random assumption (A3). For any constant $c \in (0, 1/2)$ define $\mathcal{E}'(c)$ as the event that at least $\frac{1-\rho_*}{1+2c}$ portion of the first $(s-2)$ coordinates in x are missing. Note that $\|\beta_{0\text{mis}}\|_2^2 \geqslant \frac{1-\rho_*}{1+2c}\widetilde{a}^2$ almost surely under $\mathcal{A} \cap \mathcal{E}'(C)$ and by Chernoff bound $\Pr[\mathcal{A}] \geqslant 1 - e^{-c^2(1-\rho_*)(s-2)} \geqslant 1 - e^{-0.5c^2(1-\rho_*)s}$ for $s \geqslant 4$. Subsequently, by law of total expectation

$$\mathbb{E}_{R|\mathcal{A}}\left\{\frac{2\widetilde{a}^2\gamma^2}{\sigma_\varepsilon^2+\|\beta_{0\text{mis}}\|_2^2}\right\} \leqslant \frac{2\widetilde{a}^2\gamma^2}{\sigma_\varepsilon^2+\frac{1-\rho_*}{1+2c}\widetilde{a}^2} + e^{-0.5c^2(1-\rho_*)s}\frac{2\widetilde{a}^2\gamma^2}{\sigma_\varepsilon^2}.$$

Replace $\widetilde{a}^2 = \frac{M^2}{2+\gamma^2}$. We then have that

$$\begin{aligned}
\mathrm{KL}(P_0^n\|P_1^n) &\leqslant n\rho_*^2\left[\frac{2\gamma^2}{1-\gamma^2} + \frac{2M^2\gamma^2}{(2+\gamma^2)\sigma_\varepsilon^2+\frac{1-\rho_*}{1+2c}M^2} + e^{-0.5c^2(1-\rho_*)s}\frac{2M^2\gamma^2}{(2+\gamma^2)\sigma_\varepsilon^2}\right] \\
&\leqslant n\rho_*^2\left[\frac{2\gamma^2}{1-\gamma^2} + \frac{2(1+2c)\gamma^2}{1-\rho_*} + e^{-0.5c^2(1-\rho_*)s}\frac{M^2\gamma^2}{\sigma_\varepsilon^2}\right].
\end{aligned}$$

Equating $\mathrm{KL}(P_0^n\|P_1^n) \asymp O(1)$ and applying the condition that $\gamma^2 \to 0$, we have that

$$\gamma^2 \asymp \min\left\{\frac{1-\rho_*}{2(1+2c)}, e^{0.5c^2(1-\rho_*)s}\frac{\sigma_\varepsilon^2}{M^2}\right\}\frac{1}{\rho_*^2 n}. \tag{19}$$

Subsequently,

$$\left|\beta_{0j}-\beta_{1j}\right|^2 = 4\widetilde{a}^2\gamma^2 \asymp \min\left\{\frac{1-\rho_*}{2(1+2c)}M^2, e^{0.5c^2(1-\rho_*)s}\sigma_\varepsilon^2\right\}\frac{1}{\rho_*^2 n}.$$

Invoking Lemma 13 we finish the proof of the minimax lower bound.

Finally, we justify the conditions $\gamma^2 \to 0$ and $\gamma < \gamma_0$ that are used in the proof. Eq. (19) yields $\gamma^2 \leqslant O(\frac{1}{\rho_*^2 n})$. So $\gamma^2 \to 0$ and $\gamma < \gamma_0$ is implied by $\frac{1}{\gamma_0^2\rho_*^2 n} \to 0$.

6.5 Proof of Theorem 4

Using $y = X\beta^* + \varepsilon$ we have that

$$\widetilde{\Sigma}(\widehat{\beta}_n - \beta^*) + \left(\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\widehat{\beta}_n\right) = \Bigg(\underbrace{\frac{1}{n}\widetilde{X}^\top X - \widetilde{\Sigma}}_{\Delta_n}\Bigg)\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon. \tag{20}$$

Define $\Delta_n = \frac{1}{n}\widetilde{X}^\top X - \widetilde{\Sigma}$. Recall that $\widehat{\beta}_n^u = \widehat{\beta}_n + \widehat{\Theta}\left(\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\widehat{\beta}_n\right)$. Subsequently, multiplying both sides of Eq. (20) with $\sqrt{n}\widehat{\Theta}$ and re-organizing terms we have

$$\begin{aligned}
&\sqrt{n}(\widehat{\beta}_n^u - \beta^*) = \sqrt{n}\widehat{\Theta}\left(\Delta_n\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon\right) - \sqrt{n}(\widehat{\Theta}\widetilde{\Sigma} - I)(\widehat{\beta}_n - \beta^*) \\
&= \sqrt{n}\Sigma_0^{-1}\left(\Delta_n\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon\right) - \underbrace{\sqrt{n}(\widehat{\Theta}\widetilde{\Sigma} - I)(\widehat{\beta}_n - \beta^*)}_{r_n} + \underbrace{\sqrt{n}(\widehat{\Theta} - \Sigma_0^{-1})\left(\Delta_n\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon\right)}_{\widetilde{r}_n}.
\end{aligned}$$

Define $r_n = \sqrt{n}(\widehat{\Theta}\widetilde{\Sigma} - I)(\widehat{\beta}_n - \beta^*)$ and $\widetilde{r}_n = \sqrt{n}(\widehat{\Theta} - \Sigma_0^{-1})\left(\Delta_n\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon\right)$

Lemma 9. *Suppose* $\frac{\log p}{\rho_*^4 n} \to 0$ *and the conclusion in Lemma 1 holds. Then* $\|r_n\|_\infty \leqslant O_{\mathbb{P}}(\sqrt{n}\widetilde{\nu}_n\|\widehat{\beta}_n - \beta^*\|_1)$ *and* $\|\widetilde{r}_n\|_\infty \leqslant O_{\mathbb{P}}(\sigma_x b_0 b_1 \widetilde{\nu}_n(\sigma_\varepsilon\sqrt{\frac{\log p}{\rho_*}} + \sigma_x\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2}}))$.

Lemma 9 based on Hölder's inequality and is proved in the supplementary materials. If the condition in Eq. (12) holds, Lemma 9 implies that $\max\{\|r_n\|_\infty, \|\widetilde{r}_n\|_\infty\} \xrightarrow{p} 0$, which means both terms r_n and $\widetilde{r}_n$ are asymptotically negligible in the infinity norm sense. It then suffices to analyze the limiting distribution (conditioned on X) of $a_n = \sqrt{n}\Sigma_0^{-1}\left(\Delta_n\beta^* + \frac{1}{n}\widetilde{X}^\top\varepsilon\right)$. By Assumptions (A1) and (A3), $\mathbb{E}\Delta_n|X = 0$, $\mathbb{E}\varepsilon|\widetilde{X} = 0$ and hence $\mathbb{E}a_n|X = 0$. We next analyze the conditional covariance $\mathbb{V}a_n|X$. Recall that $\Delta_n = \frac{1}{n}\widetilde{X}^\top X - \widetilde{\Sigma}$. By definition, for any $j, k \in \{1, \cdots, p\}$

$$[\Delta_n]_{jk} = \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{R_{ij}}{\rho_j}\left(1 - \frac{R_{ik}}{\rho_k}\right)X_{ij}X_{ik}, & j \neq k; \\ 0, & j = k. \end{cases}$$

Here $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ otherwise. Subsequently, $a_n = \Sigma_0^{-1}\widetilde{a}_n$ where

$$[\widetilde{a}_n]_j = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Bigg(\underbrace{\frac{R_{ij}X_{ij}}{\rho_j}\varepsilon_i + \sum_{k\neq j}\frac{R_{ij}}{\rho_j}\left(1 - \frac{R_{ik}}{\rho_k}\right)X_{ij}X_{ik}\beta_{0k}}_{T_{ij}}\Bigg).$$

Because $R \perp\!\!\!\perp X, \varepsilon$ and $\varepsilon \perp\!\!\!\perp X$, we have that $\mathbb{E}T_{ij}|X = 0$. Therefore, for any $j \in \{1, \cdots, p\}$

$$\mathbb{V}T_{ij}|X = \mathbb{E}\left[|T_{ij}|^2|X\right] = \frac{\sigma_\varepsilon^2 X_{ij}^2}{\rho_j} + \sum_{t\neq j}\frac{1-\rho_t}{\rho_j\rho_t}X_{ij}^2X_{it}^2\beta_{0t}^2$$

and for $j \neq k$,

$$\mathrm{cov}(T_{ij}, T_{ik}|X) = \mathbb{E}\left[T_{ij}T_{ik}|X\right] = \sigma_\varepsilon^2 X_{ij}X_{ik} + \sum_{t\neq j,k}\frac{1-\rho_t}{\rho_t}X_{ij}X_{ik}X_{it}^2\beta_{0t}^2.$$

Because $\{T_{ij}\}_{i=1}^n$ are i.i.d. random variables, by central limiting theorem, for any subset $S \subseteq [p]$ with constant size

$$[a_n]_{SS} \xrightarrow{d} \mathcal{N}_{|S|}\left(0, \mathrm{cov}_{SS}(a_n|X)\right) \xrightarrow{d} \mathcal{N}_{|S|}\left(0, \left[\Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1}\right]_{SS}\right),$$

where all randomness is conditioned on X.

6.6 Proof of Theorem 5

By triangle inequality and Hölder's inequality,

$$\begin{aligned}
&\|\Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1} - \widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top\|_\infty \\
&\leqslant \|(\Sigma_0^{-1} - \widehat{\Theta})\widehat{\Gamma}\Sigma_0^{-1}\|_\infty + \|\widehat{\Theta}\widehat{\Gamma}(\Sigma_0^{-1} - \widehat{\Theta}^\top)\|_\infty + \|\widehat{\Theta}(\widehat{\Gamma} - \widetilde{\Gamma})\widehat{\Theta}^\top\|_\infty \\
&\leqslant 2\max\left\{\|\Sigma_0^{-1}\|_{L_1}, \|\widehat{\Theta}\|_{L_1}, \|\widehat{\Theta}\|_{L_\infty}\right\} \max\left\{\|\Sigma_0^{-1} - \widehat{\Theta}\|_{L_1}, \|\Sigma_0^{-1} - \widehat{\Theta}\|_{L_\infty}\right\} \|\widehat{\Gamma}\|_\infty + \|\widehat{\Theta}\|_{L_1}^2\|\widehat{\Gamma} - \widetilde{\Gamma}\|_\infty.
\end{aligned}$$

With Lemma 1, the bound can be simplified to (with probability $1 - o(1)$)

$$\|\Sigma_0^{-1}\widehat{\Gamma}\Sigma_0^{-1} - \widehat{\Theta}\widetilde{\Gamma}\widehat{\Theta}^\top\|_\infty \leqslant 4b_0b_1^2\widetilde{\nu}_n\|\widehat{\Gamma}\|_\infty + b_1^2\|\widehat{\Gamma} - \widetilde{\Gamma}\|_\infty. \tag{21}$$

Note that by standard concentration inequalities of supreme of sub-Gaussian random variables, $\|X\|_\infty \leqslant O_{\mathbb{P}}(\sigma_x\sqrt{\log p})$. Also, by Hölder's inequality $\|\widehat{\Upsilon}\|_\infty \leqslant \rho_*^{-2}\|X\|_\infty^4\|\beta^{*2}\|_1$. Subsequently,

$$\|\widehat{\Gamma}\|_\infty \leqslant \frac{\sigma_\varepsilon^2}{\rho_*}\|X\|_\infty^2 + \frac{\|X\|_\infty^4\|\beta^*\|_2^2}{\rho_*^2} \leqslant O_{\mathbb{P}}\left\{\sigma_x^4\log^2 p\left(\frac{\sigma_\varepsilon^2}{\sigma_x^2\rho_*} + \frac{\|\beta^*\|_2^2}{\rho_*^2}\right)\right\}. \tag{22}$$

It remains to upper bound $\|\widehat{\Gamma} - \widetilde{\Gamma}\|_\infty$. Decompose the difference as

$$\|\widehat{\Gamma} - \widetilde{\Gamma}\|_\infty \leqslant \sigma_\varepsilon^2\left\|\frac{1}{n}\widetilde{X}^\top\widetilde{X} - \frac{1}{n}X^\top X - \widetilde{D}\mathrm{diag}\left(\frac{1}{n}X^\top X\right)\right\|_\infty + \|\widehat{\Upsilon} - \widetilde{\Upsilon}\|_\infty.$$

We first focus on the first term. Recall that $D = \mathrm{diag}(1-\rho_1, \cdots, 1-\rho_p)$, $\widetilde{D} = (\frac{1}{\rho_1} - 1, \cdots, \frac{1}{\rho_p} - 1)$ and therefore $\|\widetilde{D}\|_\infty \leqslant 1 - 1/\rho_*$ and $\frac{1}{n}\widetilde{X}^\top\widetilde{X} = \widetilde{\Sigma} + D\mathrm{diag}(\frac{1}{n}\widetilde{X}^\top\widetilde{X})$. Subsequently, the first infinity norm term is upper bounded by

$$\|\widetilde{\Sigma} - \Sigma_0\|_\infty + \left\|D\mathrm{diag}\left(\frac{1}{n}\widetilde{X}^\top\widetilde{X}\right) - \widetilde{D}\mathrm{diag}(\Sigma_0)\right\|_\infty + \frac{1}{\rho_*}\|\widehat{\Sigma} - \Sigma_0\|_\infty.$$

By Lemma 2, if $\frac{\log p}{\rho_*^4 n} \to 0$ then $\|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant O_{\mathbb{P}}(\sigma_x^2\sqrt{\frac{\log p}{\rho_*^2 n}})$ and $\|\widehat{\Sigma} - \Sigma_0\|_\infty \leqslant O_{\mathbb{P}}(\sigma_x^2\sqrt{\frac{\log p}{n}})$. For the remaining term, we invoke the following lemma that is proved in the supplementary materials:

Lemma 10. *If* $\frac{\log p}{\rho_* n} \to 0$ *then* $\|D\mathrm{diag}(\frac{1}{n}\widetilde{X}^\top\widetilde{X}) - \widetilde{D}\mathrm{diag}(\Sigma_0)\|_\infty \leqslant O_{\mathbb{P}}(\sigma_x^2\sqrt{\frac{\log p}{\rho_*^3 n}})$.

Consequently,

$$\sigma_\varepsilon^2\left\|\frac{1}{n}\widetilde{X}^\top\widetilde{X} - \frac{1}{n}X^\top X - \widetilde{D}\mathrm{diag}\left(\frac{1}{n}X^\top X\right)\right\|_\infty \leqslant O_{\mathbb{P}}\left\{\sigma_\varepsilon^2\sigma_x^2\sqrt{\frac{\log p}{\rho_*^3 n}}\right\}. \tag{23}$$

Finally, we derive the upper bound for $\|\widehat{\Upsilon} - \widetilde{\Upsilon}\|_\infty$. We first construct a $p \times p$ matrix $\overline{\Upsilon}$ as an "intermediate" quantity defined as

$$\overline{\Upsilon}_{jk} = \frac{1}{n}\sum_{i=1}^n\sum_{t\neq j,k}(1-\rho_t)\widetilde{X}_{ij}\widetilde{X}_{ik}\widetilde{X}_{it}^2\beta_{0t}^2 \quad \text{for} \quad j,k \in \{1,\cdots,p\}.$$

Note that $\overline{\Upsilon}$ involves the missing design $\widetilde{X}$ and the true model β^*. Further define $\widetilde{\Upsilon}_{jkt}$ and Υ_{jkt} for $j, k, t \in \{1, \cdots, p\}$ as

$$\widetilde{\Upsilon}_{jkt} = \frac{1}{n}\sum_{i=1}^{n}(1-\rho_t)\widetilde{X}_{ij}\widetilde{X}_{ik}\widetilde{X}_{it}^2, \qquad \Upsilon_{jkt} = \mathbb{E}\widetilde{\Upsilon}_{jkt}|X.$$

We next state the following concentration results on $\widetilde{\Upsilon}_{jkt}$ and Υ_{jkt}, which will be proved in the supplementary material.

Lemma 11. *Fix $j, k \in [p]$ and suppose $\frac{\log p}{\rho_*^3 n} \to 0$. We then have that*

$$\max_{j,k\in[p]}\max_{t\neq j,k}\left|\Upsilon_{jkt}\right| \leqslant O_{\mathbb{P}}\left(\frac{\sigma_x^4\log^2 p}{\rho_*^2}\right)$$

and

$$\max_{j,k\in[p]}\max_{t\neq j,k}\left|\widetilde{\Upsilon}_{jkt}-\Upsilon_{jkt}\right| \leqslant O_{\mathbb{P}}\left(\sigma_x^4\log^2 p\sqrt{\frac{\log p}{\rho_*^5 n}}\right).$$

We then upper bound $\|\widetilde{\Upsilon}-\widehat{\Upsilon}\|_\infty$ by bounding $\|\widehat{\Upsilon}-\overline{\Upsilon}\|_\infty$ and $\|\widetilde{\Upsilon}-\overline{\Upsilon}\|_\infty$ separately.

Upper bound for $\|\widetilde{\Upsilon}-\overline{\Upsilon}\|_\infty$ By definition, $\widetilde{\Upsilon}_{jk} = \sum_{t\neq j,k}\widetilde{\Upsilon}_{jkt}\widehat{\beta}_{nt}^2$ and $\overline{\Upsilon}_{jkt} = \sum_{t\neq j,k}\widetilde{\Upsilon}_{jkt}\beta_{0t}^2$. Hölder's inequality then yields

$$\|\widetilde{\Upsilon}-\overline{\Upsilon}\|_\infty \leqslant \max_{j,k\in[p]}\max_{t\neq j,k}\left|\widetilde{\Upsilon}_{jkt}\right|\cdot\|\widehat{\beta}_n^2-\beta^{*2}\|_1.$$

Under the condition that $\frac{\log p}{\rho_*^3 n}\to 0$, it holds that $\max_{j,k}\max_{t\neq j,k}|\widetilde{\Upsilon}_{jkt}| \leqslant O_{\mathbb{P}}(1)\cdot\max_{j,k}\max_{t\neq j,k}|\Upsilon_{jkt}|$. Furthermore, $\|\widehat{\beta}_n^2-\beta^{*2}\|_1 \leqslant \|\widehat{\beta}_n+\beta^*\|_\infty\|\widehat{\beta}_n-\beta^*\|_1 \leqslant (\|\beta^*\|_2+\|\widehat{\beta}_n-\beta^*\|_2)\|\widehat{\beta}_n-\beta^*\|_1$. Invoking Lemma 11 and the condition that $\|\widehat{\beta}_n-\beta^*\|_2 \xrightarrow{p} 0$ we get

$$\|\widetilde{\Upsilon}-\overline{\Upsilon}\|_\infty \leqslant O_{\mathbb{P}}\left\{\frac{\sigma_x^4\log^2 p}{\rho_*^2}\|\beta^*\|_2\|\widehat{\beta}_n-\beta^*\|_1\right\}. \tag{24}$$

Upper bound for $\|\widehat{\Upsilon}-\overline{\Upsilon}\|_\infty$ Note that $\widehat{\Upsilon}_{jk} = \sum_{t\neq j,k}\Upsilon_{jkt}\beta_{0t}^2$ and $\overline{\Upsilon}_{jk} = \sum_{t\neq j,k}\widetilde{\Upsilon}_{jkt}\beta_{0t}^2$. By Hölder's inequality,

$$\|\overline{\Upsilon}-\widehat{\Upsilon}\|_\infty \leqslant \max_{j,k\in[p]}\max_{t\neq j,k}\left|\widetilde{\Upsilon}_{jkt}-\Upsilon_{jkt}\right|\cdot\|\beta^{*2}\|_1.$$

Invoking Lemma 11 we then have

$$\|\overline{\Upsilon}-\widehat{\Upsilon}\|_\infty \leqslant O_{\mathbb{P}}\left\{\sigma_x^4\log^2 p\|\beta^*\|_2^2\sqrt{\frac{\log p}{\rho_*^5 n}}\right\}. \tag{25}$$

Finally, combining Eqs. (21,22,23,24,25) we complete the proof of Theorem 5.

References

[1] Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*(Jun), 1179–1225.

[2] Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, *45*(1), 77–120.

[3] Belloni, A., Chernozhukov, V., & Kaul, A. (2017). Confidence bands for coefficients in high dimensional linear models with error-in-variables. *arXiv preprint arXiv:1703.00469*.

[4] Belloni, A., Rosenbaum, M., & Tsybakov, A. B. (2016). An $(\ell_1, \ell_2, \ell_\infty)$-regularization approach to high-dimensional errors-in-variables models. *Electronic Journal of Statistics*, *10*(2), 1729–1750.

[5] Belloni, A., Rosenbaum, M., & Tsybakov, A. B. (2016). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[6] Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, (pp. 1705–1732).

[7] Cai, T., Liu, W., & Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607.

[8] Cai, T. T., Liang, T., & Rakhlin, A. (2014). Geometric inference for general high-dimensional linear inverse problems. *arXiv preprint arXiv:1404.4408*.

[9] Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, (pp. 2313–2351).

[10] Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, *52*(2), 489–509.

[11] Carroll, R., Ruppert, D., & Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[12] Chapelle, O., Scholkopf, B., & Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press.

[13] Chen, Y., & Caramanis, C. (2013). Noisy and missing data regression: Distribution-oblivious support recovery. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[14] Datta, A., & Zou, H. (2015). Cocolasso for high-dimensional error-in-variables regression. *arXiv preprint arXiv:1510.07123*.

[15] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, *52*(4), 1289–1306.

[16] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.

[17] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

[18] Hsu, D., Kakade, S. M., & Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, *17*(52), 1–6.

[19] Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the u.s. department of energy. *Journal of the American Statistical Association*, *81*(395), 680–688.

[20] Ibragimov, I. A., & Has' minskii, R. Z. (2013). *Statistical estimation: asymptotic theory*, vol. 16. Springer Science & Business Media.

[21] Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*(1), 2869–2909.

[22] Lafferty, J. D., & Wasserman, L. A. (2007). Statistical analysis of semi-supervised regression. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

[23] Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

[24] Little, R. J. A., & Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

[25] Loh, P.-L., & Wainwright, M. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, *40*(3), 1637–1664.

[26] Loh, P.-L., & Wainwright, M. J. (2012). Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*.

[27] Loh, P.-L., & Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, *16*, 559–616.

[28] Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

[29] Miller, K. S. (1981). On the inverse of the sum of matrices. *Mathematics Magazine*, *54*(2), 67–72.

[30] Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., et al. (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *The Lancet*, *359*(9314), 1301–1307.

[31] Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over-balls. *IEEE Transactions on Information Theory*, *57*(10), 6976–6994.

[32] Rosenbaum, M., & Tsybakov, A. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics*, *38*(5), 2620–2651.

[33] Rosenbaum, M., & Tsybakov, A. (2013). Improved matrix uncertainty selector.

[34] Städler, N., Stekhoven, D. J., & Bühlmann, P. (2014). Pattern alternating maximization algorithm for missing data in high-dimensional problems. *Journal of Machine Learning Research*, *15*, 1903–1928.

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

[36] van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *32*(3), 1166–1202.

[37] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, *55*(5), 2183–2202.

[38] Wang, Z., Gu, Q., Ning, Y., & Liu, H. (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

[39] Yi, X., & Caramanis, C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

[40] Zhang, C.-H., & Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *76*, 217–242.

[41] Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Supplementary Material for: Rate Optimal Estimation and Confidence Intervals for High-dimensional Regression with Missing Covariates

Yining Wang, Jialei Wang, Sivaraman Balakrishnan and Aarti Singh

This supplementary material provides detailed proofs for technical lemmas whose proofs are omitted in the main text.

A Technical Lemmas

Lemma 12 (Generalized Fano's inequality, [S20]). *Let Θ be a parameter set and $d : \Theta \times \Theta \to \mathbb{R}_{\geqslant 0}$ be a semimetric. Let P_θ be the distribution induced by θ and P_θ^n be the distribution of n i.i.d. observations from P_θ. If $d(\theta, \theta') \geqslant \alpha$ and $\mathrm{KL}(P_\theta \| P_{\theta'}) \leqslant \beta$ for all distinct $\theta, \theta' \in \Theta$, then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{P_{\theta^n}} \left[d(\widehat{\theta}, \theta) \right] \geqslant \frac{\alpha}{2} \left(1 - \frac{n\beta + \log 2}{\log |\Theta|} \right).$$

Lemma 13 (Le Cam's method, [S23]). *Suppose P_{θ_0} and P_{θ_1} are distributions induced by θ_0 and θ_1. Let $P_{\theta_0}^n$ and $P_{\theta_1}^n$ be distributions of n i.i.d. observations from P_{θ_0} and P_{θ_1}, respectively. Then for any estimator $\widehat{\theta}$ it holds that*

$$\frac{1}{2} \left[P_{\theta_0}^n(\widehat{\theta} \neq \theta_0) + P_{\theta_1}^n(\widehat{\theta} \neq \theta_1) \right] \geqslant \frac{1}{2} - \frac{1}{2} \| P_{\theta_0}^n - P_{\theta_1}^n \|_{\mathrm{TV}} \geqslant \frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{n \mathrm{KL}(P_{\theta_0} \| P_{\theta_1})}.$$

Lemma 14 (Miller [S29], Eq. (13)). *Suppose H is a matrix of rank at most 2 and $(I + H)$ is invertible. Then*

$$(I + H)^{-1} = I - \frac{aH - H^2}{a + b},$$

where $a = 1 + \mathrm{tr}(H)$ and $2b = [\mathrm{tr}(H)]^2 + \mathrm{tr}(H^2)$.

B Proofs of concentration bounds

B.1 Proof of Lemma 2

Fix arbitrary $u, v \in \mathcal{S}$. For $j, k \in [p]$ and $\ell \in \{0, 1, 2\}$, define

$$\xi_{jk}^{(0)}(R_i, \rho) = 1, \quad \xi_{jk}^{(1)}(R_i, \rho) = \frac{R_{ij}}{\rho_j}, \quad \xi_{jk}^{(2)}(R_i, \rho) = \begin{cases} \frac{R_{ij}}{\rho_j}, & j = k; \\ \frac{R_{ij} R_{ik}}{\rho_j \rho_k}, & j \neq k. \end{cases}$$

Also let $T_i^{(\ell)} = \sum_{j,k=1}^p \xi_{jk}^{(\ell)}(R_i, \rho) X_{ij} X_{ik} u_j v_k$. We then have that

$$\left| u^\top (\widehat{\Sigma} - \Sigma_0) v \right| = \left| \frac{1}{n} \sum_{i=1}^n T_i^{(0)} - \mathbb{E} T_i^{(0)} \right|, \tag{S1}$$

$$\left| u^\top (\frac{1}{n} \widetilde{X}^\top X - \Sigma_0) v \right| = \left| \frac{1}{n} \sum_{i=1}^n T_i^{(1)} - \mathbb{E} T_i^{(2)} \right|, \tag{S2}$$

$$\left| u^\top (\widetilde{\Sigma} - \Sigma_0) v \right| = \left| \frac{1}{n} \sum_{i=1}^n T_i^{(2)} - \mathbb{E} T_i^{(2)} \right|. \tag{S3}$$

The main idea is to use Berstein inequality with moment conditions (Lemma 23) to establish concentration bounds and achieve optimal dependency over ρ. Define $V^{(\ell)} = \mathbb{E}\left[|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^2\right]$. We then have that

$$V^{(\ell)} \leqslant \mathbb{E}|T_i^{(\ell)}|^2 = \sum_{j,k,j',k'=1}^{p} \mathbb{E}\left\{\xi_{jk}^{(\ell)}\xi_{j'k'}^{(\ell)}\right\}\mathbb{E}\left\{X_{ij}X_{ik}X_{ij'}X_{ik'}u_jv_ku_{j'}v_{k'}\right\}.$$

It is then of essential importance to evaluate $\mathbb{E}\left\{\xi_{jk}^{(\ell)}\xi_{j'k'}^{(\ell)}\right\}$. For $\ell = 0$ the expectation trivially equals 1. For $\ell = 1$ and $\ell = 2$, we apply the following proposition, which is easily proved by definition.

Proposition 1. $\mathbb{E}\left\{\xi_{jk}^{(1)}\xi_{j'k'}^{(1)}\right\} = 1 + I[j = j'](\frac{1}{\rho_j} - 1)$ *and* $\mathbb{E}\left\{\xi_{jk}^{(2)}\xi_{j'k'}^{(2)}\right\} = 1 + I[j = j'](\frac{1}{\rho_j} - 1) + I[k = k'](\frac{1}{\rho_k} - 1) + I[j = j' \wedge k = k'](\frac{1}{\rho_j} - 1)(\frac{1}{\rho_k} - 1) + I[j = j' = k = k'](1 - \frac{1}{\rho_j})\frac{1}{\rho_j}$. *Here* $I[\cdot]$ *is the indicator function.*

We are now ready to derive $\mathbb{E}|T_i^{(\ell)}|^2$.

$$\begin{aligned}
\mathbb{E}|T_i^{(0)}|^2 &= \mathbb{E}\left\{|X_i^\top u|^2|X_i^\top v|^2\right\};\\
\mathbb{E}|T_i^{(1)}|^2 &= \mathbb{E}\left\{|X_i^\top u|^2|X_i^\top v|^2\right\} + \sum_{j=1}^{p}\left(\frac{1}{\rho_j} - 1\right)u_j^2\mathbb{E}\left\{X_{ij}^2|X_i^\top v|^2\right\}\\
&\leqslant \mathbb{E}\left\{|X_i^\top u|^2|X_i^\top v|^2\right\} + \frac{1}{\rho_*}\sum_{j=1}^{p}u_j^2\mathbb{E}\left\{|X_i^\top e_j|^2|X_i^\top v|^2\right\};\\
\mathbb{E}|T_i^{(2)}|^2 &= \mathbb{E}\left\{|X_i^\top u|^2|X_i^\top v|^2\right\} + \sum_{j=1}^{p}\left(\frac{1}{\rho_j} - 1\right)(u_j^2 + v_j^2)\mathbb{E}\left\{X_{ij}^2|X_i^\top v|^2\right\}\\
&\quad + \sum_{k=1}^{p}\left(\frac{1}{\rho_k} - 1\right)v_j^2\mathbb{E}\left\{X_{ik}^2|X_i^\top u|^2\right\}\\
&\quad + \sum_{j,k=1}^{p}\left(\frac{1}{\rho_j} - 1\right)\left(\frac{1}{\rho_k} - 1\right)u_j^2v_k^2\mathbb{E}\left\{X_{ij}^2X_{ik}^2\right\}\\
&\quad + \sum_{j=1}^{p}\left(1 - \frac{1}{\rho_j}\right)\frac{1}{\rho_j}u_j^2v_j^2\mathbb{E}X_{ij}^4\\
&\leqslant \mathbb{E}\left\{|X_i^\top u|^2|X_i^\top v|^2\right\} + \frac{1}{\rho_*}\sum_{j=1}^{p}u_j^2\mathbb{E}\left\{|X_i^\top e_j|^2|X_i^\top v|^2\right\} + \frac{1}{\rho_*^2}\sum_{j,k=1}^{p}u_j^2v_k^2\mathbb{E}\left\{|X_i^\top e_j|^2|X_i^\top e_k|^2\right\}.
\end{aligned}$$

By Cauchy-Schwartz inequality and moment upper bounds of sub-Gaussian random variables (Lemma 19), we have that

$$\mathbb{E}\left\{|X_i^\top a|^2|X_i^\top b|^2\right\} \leqslant \sqrt{\mathbb{E}|X_i^\top a|^4}\sqrt{\mathbb{E}|X_i^\top b|^4} \leqslant 16\sigma_x^4\|a\|_2^2\|b\|_2^2.$$

Consequently, there exists universal constant $c_2 > 0$ such that

$$\mathbb{E}|T_i^{(0)}|^2 \leqslant c_2\sigma_x^4\|u\|_2^2\|v\|_2^2, \quad \mathbb{E}|T_i^{(1)}|^2 \leqslant \frac{c_2}{\rho_*}\sigma_x^4\|u\|_2^2\|v\|_2^2, \quad \mathbb{E}|T_i^{(2)}|^2 \leqslant \frac{c_2}{\rho_*^2}\sigma_x^4\|u\|_2^2\|v\|_2^2.$$

We next find an $L > 0$ so that the moment condition in Lemma 23 is satisfied, namely $\mathbb{E}|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^k \leqslant \frac{1}{2}V^{(\ell)}L^{k-2}k!$ for all $k > 1$. Note that for all $\ell \in \{0, 1, 2\}$, there exist functions $\xi_j^{(\ell)}$ and $\overline{\xi}_j^{(\ell)}$

only depending on j such that $\xi_{jk}^{(\ell)} = \xi_j^{(\ell)}\xi_k^{(\ell)} + I[j=k]\cdot\overline{\xi}_j^{(\ell)}$ and furthermore $\max_j|\xi_j^{(\ell)}| \leqslant 1/\rho_*$, $\overline{\xi}_j^{(0)} = \overline{\xi}_j^{(1)} = 0$ and $\max_j|\overline{\xi}_j^{(2)}| \leqslant 1/\rho_*^2$. Subsequently,

$$\begin{aligned}\mathbb{E}|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^k &= \mathbb{E}\left|\sum_{j,k=1}^p\left(\xi_j^{(\ell)}\xi_k^{(\ell)} + I[j=k]\cdot\overline{\xi}_j^{(\ell)} - 1\right)X_{ij}X_{ik}u_jv_k\right|^k \\ &\leqslant 3^k\left(\mathbb{E}\left|\sum_{j,k=1}^p\xi_j^{(\ell)}\xi_k^{(\ell)}X_{ij}X_{ik}u_jv_k\right|^k + \mathbb{E}\left|\sum_{j=1}^p\overline{\xi}_j^{(\ell)}X_{ij}^2u_jv_j\right|^k + \mathbb{E}\left|\sum_{j,k=1}^p X_{ij}X_{ik}u_jv_k\right|^k\right).\end{aligned}$$

Here the second line is a consequence of the following inequality: for all $a,b,c \geqslant 0$ we have that $(a+b+c)^k \leqslant (3\max\{a,b,c\})^k \leqslant 3^k\max\{a^k,b^k,v^k\} \leqslant 3^k(a^k+b^k+c^k)$. Define $\widetilde{u}_j = u_j\xi_j^{(\ell)}$, $\widetilde{v}_k = v_k\xi_k^{(\ell)}$, $\overline{u}_j = u_j\sqrt{|\overline{\xi}_j^{(\ell)}|}$ and $\overline{v}_j = v_j\sqrt{|\overline{\xi}_j^{(\ell)}|}$. Apply Lemma 24 with $|\sum_{j=1}^p\overline{\xi}_j^{(\ell)}X_{ij}^2u_jv_j| \leqslant X_i^\top AX_i$, $A = \mathrm{diag}(|\overline{u}_1\overline{v}_1|,\cdots,|\overline{u}_p\overline{v}_p|)$ and note that $\mathrm{tr}(A) \leqslant |\overline{u}|^\top|\overline{v}| \leqslant \|\overline{u}\|_2\|\overline{v}\|_2$ and $\|A\|_{\mathrm{op}} = \max_{1\leqslant j\leqslant p}|\overline{u}_j\overline{v}_j| \leqslant \|\overline{u}\|_2\|\overline{v}\|_2$. Subsequently, for all $t>0$

$$\Pr\left[X_i^\top AX_i > 3\sigma_x^2\|\overline{u}\|_2\|\overline{v}\|_2(1+t)\right] \leqslant e^{-t}. \tag{S4}$$

Let $F(x) = \Pr[X_i^\top AX_i \leqslant x], x\geqslant 0$ be the CDF of $X_i^\top AX_i$ and $G(x) = 1-F(x)$. Using integration by parts, we have that

$$\mathbb{E}|X_i^\top AX_i|^k = \int_0^\infty x^k\mathrm{d}F(x) = -\int_0^\infty x^k\mathrm{d}G(x) = \int_0^\infty kx^{k-1}G(x)\mathrm{d}x.$$

Here in the last equality we use the fact that $\lim_{x\to\infty}x^kG(x) = 0$ for any fixed $k\in\mathbb{N}$, because $G(x)\leqslant\exp\{1-\frac{x}{M}\}$ by Eq. (S4), where $M = 3\sigma_x^2\|\overline{u}\|_2\|\overline{v}\|_2$. Consequently,

$$\begin{aligned}\mathbb{E}|X_i^\top AX_i|^k &= \int_0^M kx^{k-1}G(x)\mathrm{d}x + k\int_M^\infty x^{k-1}G(x)\mathrm{d}x \\ &\leqslant M^k + k\int_0^\infty M^{k-1}(1+z)^{k-1}e^{-z}\cdot M\mathrm{d}z \\ &= M^k + kM^k\int_0^\infty(1+z)^{k-1}e^{-z}\mathrm{d}z \\ &\leqslant M^k + kM^k\cdot k! \leqslant (k+1)!M^k.\end{aligned}$$

Here in the second line we apply change-of-variable $x = M(1+z)$ and the fact that $G(M(1+z)) \leqslant e^{-z}$ in the integration term. Because $2^k \geqslant k+1$ for all $k\geqslant 1$, we conclude that

$$\mathbb{E}\left|\sum_{j=1}^p\overline{\xi}_j^{(\ell)}X_{ij}^2u_jv_j\right|^k \leqslant 6^k\sigma_x^{2k}k!\mathbb{E}\|\overline{u}\|_2^k\|\overline{v}\|_2^k, \quad \forall k\geqslant 1.$$

Subsequently, applying Cauchy-Schwartz inequality together with moment bounds for sub-Gaussian random variables (Lemma 19) we obtain

$$\begin{aligned}&\mathbb{E}|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^k \\ &\leqslant 3^k\left(\sqrt{\mathbb{E}|X_i^\top\widetilde{u}|^{2k}}\sqrt{\mathbb{E}|X_i^\top\widetilde{v}|^{2k}} + 6^k\sigma_x^{2k}k!\mathbb{E}\|\overline{u}\|_2^k\|\overline{v}\|_2^k + \sqrt{\mathbb{E}|X_i^\top u|^{2k}}\sqrt{\mathbb{E}|X_i^\top v|^{2k}}\right) \\ &\leqslant 3^k\cdot 2k\cdot 6^k\Gamma(k)\sigma_x^{2k}\cdot\left(\sqrt{\mathbb{E}\|\widetilde{u}\|_2^{2k}}\sqrt{\mathbb{E}\|\widetilde{v}\|_2^{2k}} + \sqrt{\mathbb{E}\|\overline{u}\|_2^{2k}}\sqrt{\mathbb{E}\|\overline{v}\|_2^{2k}} + \|u\|_2^k\|v\|_2^k\right) \\ &\leqslant \rho_*^{\ell/2}\left(\frac{C'\|u\|_2\|v\|_2\sigma_x^2}{\rho_*^\ell}\right)^k k!,\end{aligned}$$

where $C' < \infty$ is some absolute constant. Compare the bound of $\mathbb{E}|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^k$ with the variance $\mathbb{E}|T_i^{(\ell)}|^2$ we obtained earlier, we have that $L = \sigma_x^2\|u\|_2\|v\|_2 \cdot C'^3/\rho_*^{1.5\ell}$ is sufficient to guarantee $\mathbb{E}|T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}|^k \leqslant \frac{1}{2}V^{(\ell)}L^{k-2}k!$ for all [4] $k > 2$. Applying Bernstein inequality with moment conditions (Lemma 23) and union bound over all $u, v \in \mathcal{S}$, we have that

$$\Pr\left[\forall u, v \in \mathcal{S}, \left|\frac{1}{n}\sum_{i=1}^n T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}\right| > \|u\|_2\|v\|_2\epsilon\right] \leqslant 2N^2 \exp\left\{-\frac{n\epsilon^2}{2(\widetilde{V}^{(\ell)} + \widetilde{L}\epsilon)}\right\}$$

for all $\epsilon > 0$, where $\widetilde{V}^{(\ell)} = \frac{V^{(\ell)}}{\|u\|_2^2\|v\|_2^2}$ and $\widetilde{L} = \frac{L}{\|u\|_2\|v\|_2}$. Subsequently,

$$\begin{aligned}\sup_{u,v\in\mathcal{S}}\left|\frac{1}{n}\sum_{i=1}^n T_i^{(\ell)} - \mathbb{E}T_i^{(\ell)}\right| &\leqslant O_{\mathbb{P}}\left(\|u\|_2\|v\|_2\max\left\{\frac{\widetilde{L}\log N}{n}, \sqrt{\frac{\widetilde{V}^{(\ell)}\log N}{n}}\right\}\right) \\ &\leqslant O_{\mathbb{P}}\left(\sigma_x^2\|u\|_2\|v\|_2\max\left\{\frac{\log N}{\rho_*^{1.5\ell}n}, \sqrt{\frac{\log N}{\rho_*^{\ell}n}}\right\}\right),\end{aligned}$$

as desired.

B.2 Proof of Lemma 3

Define $\delta_j = \frac{1}{n}\sum_{i=1}^n Z_{ij}$ where $Z_{ij} = \widetilde{X}_{ij}\varepsilon_i$. Because $\mathbb{E}\varepsilon_i|X = 0$, we have that $\mathbb{E}Z_{ij} = 0$. In addition,

$$\mathbb{E}|Z_{ij}|^2 = \frac{\sigma_\varepsilon^2\sigma_x^2}{\rho_j} \leqslant \frac{\sigma_\varepsilon^2\sigma_x^2}{\rho_*} =: V$$

and for $k > 2$,

$$\begin{aligned}\mathbb{E}|Z_{ij}|^k &= \rho_j \cdot \frac{1}{\rho_j^k} \cdot \mathbb{E}\varepsilon_i^k \cdot \mathbb{E}|X_{ij}|^k \\ &\leqslant \frac{1}{\rho_*^{k-1}} \cdot k^2 2^k \sigma_x^k \sigma_\varepsilon^k \Gamma\left(\frac{k}{2}\right)^2 \\ &\leqslant \frac{k^2(2\sigma_x\sigma_\varepsilon)^k}{\rho_*^{k-1}}k! \\ &\leqslant \rho_*\left(\frac{8\sigma_x\sigma_\varepsilon}{\rho_*}\right)^k k!.\end{aligned}$$

By setting $L = 64\sigma_x\sigma_\varepsilon/\rho_*$ we have that $\mathbb{E}|Z_{ij}|^k \leqslant \frac{1}{2}VL^{k-2}k!$ for all $k > 1$. Subsequently, applying Bernstein inequality with moment conditions (Lemma 23) and union bound over $j = 1, \cdots, p$ we have that

$$\Pr\left[\|\delta\|_\infty > \epsilon\right] \leqslant 2p\exp\left\{-\frac{n\epsilon^2}{2(V + L\epsilon)}\right\}$$

for any $\epsilon > 0$. Suppose $\frac{\epsilon L}{V} \to 0$. We then have that

$$\|\delta\|_\infty \leqslant O_{\mathbb{P}}\left(\sigma_\varepsilon\sigma_x\sqrt{\frac{\log p}{\rho_* n}}\right).$$

The condition $\frac{\epsilon L}{V} \to 0$ is satisfied with $\frac{\log p}{\rho_* n} \to 0$.

[4] The case of $k = 2$ is trivially true.

B.3 Proof of Lemma 10

Fix arbitrary $j \in \{1, \cdots, p\}$ and consider

$$T_{ij} = (1-\rho_j)\widetilde{X}_{ij}^2 = \frac{(1-\rho_j)R_{ij}X_{ij}^2}{\rho_j^2}.$$

It is easy to verify that $[D\mathrm{diag}(\frac{1}{n}\widetilde{X}^\top\widetilde{X})]_{jj} = \frac{1}{n}\sum_{i=1}^n T_{ij}$ and $[\widetilde{D}\mathrm{diag}(\Sigma_0)]_{jj} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}T_{ij} = \frac{(1-\rho_j)\Sigma_{0jj}}{\rho_j}$. We use moment based Bernstein's inequality (Lemma 23) to bound the perturbation $|\frac{1}{n}\sum_{i=1}^n T_{ij} - \mathbb{E}T_{ij}|$. Define $V_j = \mathbb{E}|T_{ij} - \mathbb{E}T_{ij}|^2$. We then have

$$V_j \leqslant \mathbb{E}|T_{ij}|^2 = \frac{(1-\rho_j)^2\mathbb{E}X_{ij}^4}{\rho_j^3} \leqslant \frac{3\sigma_x^4}{\rho_*^3}$$

and for all $k \geqslant 3$,

$$\mathbb{E}|T_{ij} - \mathbb{E}T_{ij}|^k \leqslant 2^k\left(\mathbb{E}|T_{ij}|^k + |\mathbb{E}T_{ij}|^k\right) \leqslant \frac{4^{k+1}}{\rho_*^{2k-1}}\sigma_x^{2k}k!.$$

It can then be verified that $\mathbb{E}|T_{ij} - \mathbb{E}T_{ij}|^k \leqslant \frac{1}{2}V_jL^{k-2}k!$ for all $k \geqslant 2$ if $L = \frac{512\sigma_x^2}{\rho_*^2}$. By Lemma 23 and a union bound over all $j \in \{1, \cdots, p\}$, we have that

$$\Pr\left[\forall j, \left|\frac{1}{n}\sum_{i=1}^n T_{ij} - \mathbb{E}T_{ij}\right| > \epsilon\right] \leqslant 2p\exp\left\{-\frac{n\epsilon^2}{2(V+L\epsilon)}\right\}$$

for all $\epsilon > 0$, where $V = \frac{3\sigma_x^4}{\rho_*^3}$ and $L = \frac{512\sigma_x^2}{\rho_*^2}$. Under the assumption that $\frac{\epsilon L}{V} \to 0$, we have that

$$\left\|\widetilde{D}\mathrm{diag}\left(\frac{1}{n}\widetilde{X}^\top\widetilde{X}\right) - D\mathrm{diag}(\Sigma_0)\right\|_\infty = \sup_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n T_{ij} - \mathbb{E}T_{ij}\right| \leqslant O_{\mathbb{P}}\left(\sigma_x^2\sqrt{\frac{\log p}{\rho_*^3 n}}\right).$$

The condition $\frac{\epsilon L}{V} \to 0$ is then satisfied with $\frac{\log p}{\rho_* n} \to 0$.

B.4 Proof of Lemma 11

By definition and the missing data model,

$$\Upsilon_{jkt} = \mathbb{E}\widetilde{\Upsilon}_{jkt}|X = \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{1-\rho_t}{\rho_j\rho_k}X_{ij}^2X_{it}^2, & j = k; \\ \frac{1}{n}\sum_{i=1}^n \frac{1-\rho_t}{\rho_k}X_{ij}X_{ik}X_{it}^2, & j \neq k. \end{cases}$$

Subsequently,

$$\max_{j,k\in[p]}\max_{t\neq j,k}\left|\Upsilon_{jkt}\right| \leqslant \frac{\|X\|_\infty^4}{\rho_*^2} \leqslant O_{\mathbb{P}}\left(\frac{\sigma_x^4\log^2 p}{\rho_*^2}\right).$$

To prove the second part of this lemma, we first fix arbitrary $j, k \in [p]$ and $t \neq j, k$. Define

$$T_{ijkt} = \left(\xi_{jkt}(R_i, \rho) - \mathbb{E}\xi_{jkt}(R_i, \rho)\right)X_{ij}X_{ik}X_{it}^2,$$

where $\xi_{jkt}(R_i, \rho) = \frac{(1-\rho_t)R_{ij}R_{ik}R_{it}}{\rho_j\rho_k\rho_t^2}$. It is easy to verify that $\widetilde{\Upsilon}_{jkt} - \Upsilon_{jkt} = \frac{1}{n}\sum_{i=1}^n T_{ijkt}$ and $\mathbb{E}T_{ijkt}|X = 0$. We then use Bernstein inequality with support conditions (Lemma 22) to bound the

concentration of $\frac{1}{n}\sum_{i=1}^{n} T_{ijkt}$ towards zero. Define $A = \max_{i,j,k,t} |T_{ijkt}|$ and $V = \max_{i,j,k,t} \mathbb{E}|T_{ijkt}|^2$. By Hölder's inequality we have that

$$A \leqslant \frac{\|X\|_\infty^4}{\rho_*^4} \leqslant O_{\mathbb{P}}\left(\frac{\sigma_x^4 \log^2 p}{\rho_*^4}\right).$$

Here in the $O_{\mathbb{P}}(\cdot)$ notation the randomness is on the generating process of X and is *independent* of the randomness of missing patterns R. In addition, note that

$$\mathbb{E}\left| \left(\xi_{jkt} - \mathbb{E}\xi_{jkt}\right)\left(\xi_{jkt'} - \mathbb{E}\xi_{jkt'}\right) \right| \leqslant \frac{1}{\rho_*^5}$$

for all $j, k, t, t' \in \{1, \cdots, p\}$ and $t, t' \neq j, k$. Subsequently,

$$V = \max_{i,j,k,t} \mathbb{E}|T_{ijkt}|^2 \leqslant \frac{1}{\rho_*^5} X_{ij}^2 X_{ik}^2 X_{it}^4 \leqslant \frac{\|X\|_\infty^8}{\rho_*^5} \leqslant O_{\mathbb{P}}\left(\frac{\sigma_x^8 \log^4 p}{\rho_*^5}\right).$$

Applying Lemma 22 conditioned on $\|X\|_\infty \leqslant O(\frac{\sigma_x^4 \log^2 p}{\rho_*^2})$, we have that with probability $1 - O(\delta)$ for some $\delta = o(1)$ the following holds:

$$\left|\frac{1}{n}\sum_{i=1}^{n} T_{ijkt}\right| \leqslant O\left(\sigma_x^4 \log^2 p \sqrt{\frac{\log(1/\delta)}{\rho_*^5 n}}\right) =: \epsilon,$$

provided that $\frac{\epsilon A}{V} \to 0$. Applying union bound over all $j, k \in [p]$ and $t \in [p] \backslash \{j, k\}$ we get

$$\max_{j,k\in[p]} \max_{t\neq j,k} \left|\frac{1}{n}\sum_{i=1}^{n} T_{ijkt}\right| \leqslant O_{\mathbb{P}}\left(\sigma_x^4 \log^2 p \sqrt{\frac{\log p}{\rho_*^5 n}}\right),$$

The condition $\frac{\epsilon A}{V} \to 0$ is satisfied with $\frac{\log p}{\rho_*^3 n} \to 0$.

C Proof of restricted eigenvalue conditions

In this section we review the standard analysis that establishes restricted eigenvalue conditions for sample covariance and adapt it to our missing data setting by invoking Lemma 2.

Lemma 15. *Suppose A, B are $p \times p$ random matrices with $\Pr[\|A - B\|_\infty \leqslant M] \geqslant 1 - o(1)$ for some $M < \infty$. If A satisfies $\mathrm{RE}(s, \phi_{\min})$ and B satisfies $\mathrm{RE}(s, \phi'_{\min})$, then with probability $1 - o(1)$ we have that*

$$\phi'_{\min} \geqslant \phi_{\min} - \{O(1) \cdot \varphi_{u,v}(A, B; O(s \log(Mp))) + O(1/n)\}.$$

Proof. For any $h \in \mathbb{R}^p$ it holds that

$$\frac{h^\top B h}{h^\top h} \geqslant \frac{h^\top A h}{h^\top h} - \frac{h^\top (B - A) h}{h^\top h}.$$

With appropriate scalings, it suffices to bound

$$\sup_{h:\|h_{J^c}\|_1 \leqslant \|h_J\|_1, \|h\|_2 \leqslant 1} \left|h^\top (B - A) h\right|$$

for all $J \subseteq [p]$, $|J| \leqslant s$ as the largest possible gap between $\phi_{\min}$ and $\phi'_{\min}$.

Define $\mathbb{B}_p(r) = \{x \in \mathbb{R}^p : \|x\|_p \leqslant r\}$ as the p-norm ball of radius r. Because $\|h_{J^c}\|_1 \leqslant \|h_J\|_1$ implies $\|h\|_1 \leqslant 2\|h_J\|_1 \leqslant 2\sqrt{s}\|h\|_2$, we have that

$$\sup_{h:\|h_{J^c}\|_1 \leqslant \|h_J\|_1, \|h\|_2 \leqslant 1} \left|h^\top (B-A)h\right| \leqslant \sup_{h \in \mathbb{B}_2(1) \cap \mathbb{B}_1(2\sqrt{s})} \left|h^\top (B-A)h\right|.$$

By Lemma 11 in the supplementary material of [S25], we have that

$$\begin{aligned} \mathbb{B}_2(1) \cap \mathbb{B}_1(2\sqrt{s}) &\subseteq 3\mathrm{conv}\left\{\mathbb{B}_0(4s) \cap \mathbb{B}_2(1)\right\} \\ &\subseteq \mathrm{conv}\{\underbrace{\mathbb{B}_0(4s) \cap \mathbb{B}_2(3)}_{K(4s)}\}. \end{aligned}$$

Here $\mathrm{conv}(A)$ denotes the convex hull of set A. Let $K(4s) = \mathbb{B}_0(4s) \cap \mathbb{B}_2(3)$ and denote $N_{\epsilon,\|\cdot\|_2}(K(4s))$ as the *covering number* of $K(4s)$ with respect to the Euclidean norm $\|\cdot\|_2$. That is, $N_{\epsilon,\|\cdot\|_2}(K(4s))$ is the size of the smallest *covering set* $H \subseteq K(4s)$ such that $\sup_{h \in K(4s)} \inf_{h' \in H} \|h - h'\|_2 \leqslant \epsilon$. By definition of the concentration bounds, we have that with probability $1 - o(1)$

$$\sup_{h \in H} \left|h^\top (A-B)h\right| \leqslant \varphi_{u,u}(A, B; \log |H|) \sup_{h \in H} \|h\|_2^2 \leqslant 9\varphi_{u,u}(A, B; \log N_{\epsilon,\|\cdot\|_2}(K(4s))).$$

Subsequently, for any $\epsilon \in (0,1)$ with probability $1 - o(1)$

$$\begin{aligned} \sup_{h \in \mathbb{B}_2(1) \cap \mathbb{B}_1(2\sqrt{s})} \left|h^\top (B-A)h\right| &\leqslant \sup_{h \in \mathrm{conv}\{K(4s)\}} \left|h^\top (A-B)h\right| \\ &\leqslant \sup_{\substack{\xi_1,\cdots,\xi_T \geqslant 0, \\ \xi_1+\cdots+\xi_T=1, \\ h_1,\cdots,h_T \in K(4s)}} \sum_{i,j=1}^{T} \xi_i \xi_j \left|h_i^\top (A-B)h_j\right| \\ &\leqslant \sup_{h,h' \in K(4s)} \left|h^\top (A-B)h'\right| \\ &\leqslant \sup_{h,h' \in H_{\epsilon,\|\cdot\|_2}[K(4s)]} \left|h^\top (A-B)h'\right| + (6\epsilon + 3\epsilon^2)\|A-B\|_{L_2} \\ &\leqslant 36\left\{\varphi_{u,u}(A, B; \log N_{\epsilon,\|\cdot\|_2}(K(4s))) + \epsilon p M\right\}. \end{aligned}$$

Here the last inequality is implied by the condition that $\|A-B\|_\infty \leqslant M$ with probability $1 - O(n^{-\alpha})$. Taking $\epsilon = O(1/(p^2 M))$ we have that $\epsilon p M = O(1/p) = O(1/n)$.

The final part of the proof is to establish upper bounds for the covering number $N_{\epsilon,\|\cdot\|_2}(K(4s))$. First note that by definition

$$K(4s) = \bigcup_{J \subseteq [p] : |J| \leqslant 4s} \{h : \mathrm{supp}(h) = J \wedge \|h\|_2 \leqslant 3\}.$$

The covering number of a union of subsets can be upper bounded by the following proposition:

Proposition 2. *Let $K = K_1 \cup \cdots \cup K_m$. Then $N_{\epsilon,\|\cdot\|_2}(K) \leqslant \sum_{i=1}^m N_{\epsilon,\|\cdot\|_2}(K_i)$.*

Proof. Let $H_i \subseteq K_i$ be covering sets of subset K_i. Define $H = H_1 \cup \cdots \cup H_m$. Clearly $|H| \leqslant \sum_{i=1}^m |H_i| \leqslant \sum_{i=1}^m N_{\epsilon,\|\cdot\|_2}(K_i)$. It remains to prove that H is a valid ϵ-covering set of K. Take arbitrary $h \in K$. By definition, there exists $i \in [m]$ such that $h \in K_i$. Subsequently, there exists $h^* \in H_i \subseteq H$ such that $\|h - h^*\|_2 \leqslant \epsilon$. Therefore, H is a valid ϵ-covering set of K. □

Define $K_J(r) = \{h : \mathrm{supp}(h) = J \wedge \|h\|_2 \leqslant r\}$. The covering number of K_J is established in the following proposition:

Proposition 3. $N_{\epsilon,\|\cdot\|_2}(K_J(r)) \leqslant \left(\frac{4r+\epsilon}{\epsilon}\right)^{|J|}$.

Proof. $K_J(r)$ is nothing but a centered $|J|$-dimensional ball of radius r, locating at the coordinates indexed by J. The covering number result of high-dimensional ball is due to Lemma 2.5 of [S44]. □

Combining the three propositions, we obtain

$$\log N_{\epsilon,\|\cdot\|_2}(K(4s)) \leqslant \log\left(\sum_{j=0}^{4s}\binom{p}{j}\right) + \log\left\{\left(\frac{12+\epsilon/2}{\epsilon/2}\right)^{4s}\right\} \leqslant O\left(s\log(p/\epsilon)\right).$$

With the configuration of $\epsilon = O(1/(p^2M))$, we have that

$$\log N_{\epsilon,\|\cdot\|_2}(K(4s)) \leqslant O(s\log(pM)).$$

□

We are now ready to prove Lemma 5.

Proof of Lemma 5. Consider $A = \widetilde{\Sigma}$ and $B = \Sigma_0$ in Lemma 15. Lemma 2 yields

$$\varphi_{u,v}(\widetilde{\Sigma}, \Sigma_0; O(s\log(Mp))) \leqslant O\left(\sigma_x^2 \max\left\{\frac{s\log(Mp)}{\rho_*^3 n}, \sqrt{\frac{s\log(Mp)}{\rho_*^2 n}}\right\}\right) =: \epsilon.$$

By Lemma 15, to prove this corollary it is sufficient to show that $\frac{\epsilon}{\lambda_{\min}(\Sigma_0)} \to 0$. Note also that $M = \|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant \frac{\|X\|_\infty}{\rho_*^2} \leqslant O\left(\frac{\sigma_x\sqrt{\log p}}{\rho_*^2}\right)$ with probability $1 - o(1)$. The condition $\frac{\epsilon}{\lambda_{\min}(\Sigma_0)} \to 0$ can then be satisfied with $\frac{\sigma_x^4 s\log(\sigma_x \log p/\rho_*)}{\rho_*^3\lambda_{\min}^2 n} \to 0$. □

D Proof of Lemma 1

Lemma 16. *Suppose* $\frac{\log p}{\rho_*^4 n} \to 0$ *and* $\widetilde{\nu}_n \asymp \sigma_x^2 b_1\sqrt{\frac{\log p}{\rho_*^2 n}}$. *Then with probability* $1 - o(1)$ *the population precision matrix* Σ_0^{-1} *is a feasible solution to Eq. (9); that is,* $\max\{\|\widetilde{\Sigma}\Sigma_0^{-1} - I_{p\times p}\|_\infty, \|\Sigma_0^{-1}\widetilde{\Sigma} - I_{p\times p}\|_\infty\} \leqslant \nu_n$.

Proof. First by Hölder's inequality we have that

$$\|\widetilde{\Sigma}\Sigma_0^{-1} - I\|_\infty = \|(\widetilde{\Sigma} - \Sigma_0)\Sigma_0^{-1}\|_\infty \leqslant \|\Sigma_0^{-1}\|_{L_1}\|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant b_1\|\widetilde{\Sigma} - \Sigma_0\|_\infty.$$

By Lemma 2, with probability $1 - o(1)$

$$\|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant \varphi_{u,v}\left(\widetilde{\Sigma}, \Sigma_0; 2\log p\right) \leqslant O\left(\sigma_x^2\sqrt{\frac{\log p}{\rho_*^2 n}}\right),$$

provided that $\frac{\log p}{\rho_*^4 n} \to 0$. Subsequently, we have that

$$\|\Sigma_0^{-1}\|_{L_1}\|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant O\left(\sigma_x^2 b_1\sqrt{\frac{\log p}{\rho_*^2 n}}\right) \leqslant \widetilde{\nu}_n \tag{S5}$$

with probability $1 - o(1)$. The $\|\Sigma_0^{-1}\widetilde{\Sigma} - I\|_\infty$ term can be bounded in the same way by noting that $\|\Sigma_0^{-1}\widetilde{\Sigma} - I\|_\infty \leqslant \|\Sigma_0^{-1}\|_{L_\infty}\|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant b_1\|\widetilde{\Sigma} - \Sigma_0\|_\infty$. □

Lemma 17. *Suppose Σ_0^{-1} is a feasible solution to the CLIME optimization problem in Eq. (9). Then* $\max\{\|\widehat{\Theta}\|_{L_1}, \|\widehat{\Theta}\|_{L_\infty}\} \leqslant \|\Sigma_0^{-1}\|_{L_1}$ *and* $\|\widehat{\Theta} - \Sigma_0^{-1}\|_\infty \leqslant 2\widetilde{\nu}_n \|\Sigma_0^{-1}\|_{L_1}$.

Proof. We first establish that $\|\widehat{\Theta}\|_{L_1} \leqslant \|\Sigma_0^{-1}\|_{L_1}$. In [S42] it is proved that the solution set of Eq. (9) is identical to the solution set of

$$\widehat{\Theta} = \{\widehat{\omega}_i\}_{i=1}^p, \quad \widehat{\omega}_i \in \operatorname{argmin}_{\omega_i \in \mathbb{R}^p} \left\{ \|\omega_i\|_1 : \|\widetilde{\Sigma}\omega_i - e_i\|_\infty \leqslant \widetilde{\nu}_n \right\}.$$

Because Σ_0^{-1} belongs to the feasible set of the above constrained optimization problem, we have that $\|\widehat{\omega}_i\|_1 \leqslant \|\Sigma_0^{-1}\|_{L_1}$ for all $i = 1, \cdots, p$ and hence $\|\widehat{\Theta}\|_{L_1} \leqslant \|\Sigma_0^{-1}\|_{L_1}$. The inequality $\|\widehat{\Theta}\|_{L_\infty} \leqslant \|\Sigma_0^{-1}\|_{L_1}$ can be proved by applying the same argument to $\widehat{\Theta}^\top$.

We next prove the infinity norm bound fot the estimation error $\widehat{\Theta} - \Sigma_0^{-1}$. By triangle inequality,

$$\|\Sigma_0(\widehat{\Theta} - \Sigma_0^{-1})\|_\infty \leqslant \|\widetilde{\Sigma}\widehat{\Theta} - I\|_\infty + \|(\widetilde{\Sigma} - \Sigma_0)\widehat{\Theta}\|_\infty \leqslant \widetilde{\nu}_n + \|(\widetilde{\Sigma} - \Sigma_0)\widehat{\Theta}\|_\infty.$$

Using Hölder's inequality, we have that

$$\|(\widetilde{\Sigma} - \Sigma_0)\widehat{\Theta}\|_\infty \leqslant \|\widehat{\Theta}\|_{L_1} \|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant \|\Sigma_0^{-1}\|_{L_1} \|\widetilde{\Sigma} - \Sigma_0\|_\infty \leqslant \widetilde{\nu}_n.$$

Here the last inequality is due to Eq. (S5). Subsequently, $\|\Sigma_0(\widehat{\Theta} - \Sigma_0^{-1})\|_\infty \leqslant 2\widetilde{\nu}_n$. Applying Hölder's inequality again we obtain

$$\|\widehat{\Theta} - \Sigma_0^{-1}\|_\infty \leqslant \|\Sigma_0^{-1}\|_{L_1} \|(\widetilde{\Sigma} - \Sigma_0)\widehat{\Theta}\|_\infty \leqslant 2\widetilde{\nu}_n \|\Sigma_0^{-1}\|_{L_1}.$$

□

To translate the infinity-norm estimation error Σ_0^{-1} into an L_1-norm bound that we desire, we need the following lemma that establishes basic inequality of the estimation error:

Lemma 18. *Suppose Σ_0^{-1} is a feasible solution to Eq. (9). Then under Assumption (A5) we have that* $\max\{\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_1}, \|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_\infty}\} \leqslant 2b_0 \|\widehat{\Theta} - \Sigma_0^{-1}\|_\infty$.

Proof. Let $\widehat{\omega}_i$ and $\widehat{\omega}_{0i}$ be the ith columns of $\widehat{\Theta}$ and Σ_0^{-1}, respectively. Let J_i denote the support size of $\widehat{\omega}_{0i}$. Definte $\widehat{h} = \widehat{\omega}_i - \omega_{0i}$. We then have that

$$\|\widehat{\omega}_i\|_1 = \|\omega_{0i} + \widehat{h}_{J_i^c}\|_1 + \|\widehat{h}_{J_i}\|_1 \geqslant \|\omega_{0i}\|_1 - \|\widehat{h}_{J_i^c}\|_1 + \|\widehat{h}_{J_i}\|_1.$$

On the other hand, $\|\widehat{\omega}_i\|_1 \leqslant \|\omega_{0i}\|_1$ as shown in the proof of Lemma 17. Subsequently, $\|\widehat{h}_{J_i^c}\|_1 \leqslant \|\widehat{h}_{J_i}\|_1$ and hence

$$\|\widehat{\omega}_i - \omega_{0i}\|_1 = 2\|\widehat{h}_{J_i}\|_1 \leqslant 2|J_i|\|\widehat{h}\|_\infty \leqslant 2b_0 \|\widehat{\omega}_i - \omega_{0i}\|_\infty.$$

Because the above inequality holds for all $i = 1, \cdots, p$, we conclude that $\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_1} \leqslant 2b_0 \|\widehat{\Theta} - \Sigma_0^{-1}\|_\infty$. The bound for $\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_\infty}$ can be proved by applying the same argument to $\widehat{\Theta}^\top$. □

Combining all the above lemmas, we have that with probability $1 - o(1)$

$$\max\{\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_1}, \|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_\infty}\} \leqslant 2\widetilde{\nu}_n b_0 \|\Sigma_0^{-1}\|_{L_1} \leqslant O\left\{ \sigma_x^2 b_0 b_1^2 \sqrt{\frac{\log p}{\rho_*^2 n}} \right\}.$$

E Proofs of the other technical lemmas

E.1 Proof of Lemma 4

We first show that under the conditions on n, $\widetilde{\lambda}_n$ and $\check{\lambda}_n$ specified in the lemma, the true regression vector β^* is feasible to both optimization problems with high probability; that is, $\|\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\beta^*\|_\infty \leqslant \widetilde{\lambda}_n$ and $\|\frac{1}{n}\widetilde{X}^\top y - \Sigma_0\beta^*\|_\infty \leqslant \check{\lambda}_n$ with probability $1 - o(1)$.

Consider $\widehat{\beta}_n$ first. Apply $y = X\beta^* + \varepsilon$ and Definition 1, we have that with probability $1 - o(1)$

$$\begin{aligned}\left\|\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\beta^*\right\|_\infty &\leqslant \left\|\left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^*\right\|_\infty + \left\|\left(\widetilde{\Sigma} - \Sigma_0\right)\beta^*\right\|_\infty + \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty \\ &\leqslant \left\{\varphi_{u,v}\left(\frac{1}{n}\widetilde{X}^\top X, \Sigma_0; \log p\right) + \varphi_{u,v}\left(\widetilde{\Sigma}, \Sigma_0; \log p\right)\right\}\|\beta^*\|_2 + \sigma_\varepsilon^2\varphi_{\varepsilon,\infty}\left(\frac{1}{n}\widetilde{X}\right).\end{aligned}$$

Now apply Lemmas 2 and 3: with probability $1 - o(1)$

$$\left\|\frac{1}{n}\widetilde{X}^\top y - \widetilde{\Sigma}\beta^*\right\|_\infty \leqslant O\left\{\sigma_x\sqrt{\frac{\log p}{n}}\left(\frac{\sigma_x\|\beta^*\|_2}{\rho_*} + \frac{\sigma_\varepsilon}{\sqrt{\rho_*}}\right)\right\} \leqslant \widetilde{\lambda}_n,$$

provided that $\frac{\log p}{\rho_*^4 n} \to 0$. The same line of argument applies to the second inequality by the following decomposition: under the condition that $\frac{\log p}{\rho_*^2 n} \to 0$, with probability $1 - o(1)$

$$\begin{aligned}\left\|\frac{1}{n}\widetilde{X}^\top y - \Sigma_0\beta^*\right\|_\infty &\leqslant \left\|\left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^*\right\|_\infty + \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty \\ &\leqslant \varphi_{u,v}\left(\frac{1}{n}\widetilde{X}^\top X, \Sigma_0, \log p\right)\|\beta^*\|_2 + \sigma_\varepsilon^2\varphi_{\varepsilon,\infty}\left(\frac{1}{n}\widetilde{X}\right) \\ &\leqslant O\left\{\sigma_x\sqrt{\frac{\log p}{\rho_* n}}\left(\sigma_x\|\beta^*\|_2 + \sigma_\varepsilon\right)\right\} \leqslant \check{\lambda}_n.\end{aligned}$$

We are now ready to prove Lemma 4. We only prove the assertion involving $\widehat{\beta}_n$, because the same argument applies for $\check{\beta}_n$ as well. Let $\widehat{h} = \widehat{\beta}_n - \beta^*$. Because $J_0 = \mathrm{supp}(\beta^*)$, we have that

$$\|\widehat{\beta}_n\|_1 = \|\beta^* + \widehat{h}_{J_0}\|_1 + \|\widehat{h}_{J_0^c}\|_1 \geqslant \|\beta^*\|_1 - \|\widehat{h}_{J_0}\|_1 + \|\widehat{h}_{J_0^c}\|_1.$$

On the other hand, because both $\widehat{\beta}_n$ and β^* are feasible, by definition of the optimization problem we have that $\|\widehat{\beta}_n\|_1 \leqslant \|\beta^*\|_1$. Combining both chains of inequalities we arrive at $\|\widehat{h}_{J_0^c}\|_1 \leqslant \|\widehat{h}_{J_0}\|_1$, which is to be demonstrated.

E.2 Proof of Lemma 7

Proposition 4. *Suppose $X \sim \mathcal{N}(\mu, \nu^2)$ for $\mu \in \mathbb{R}$ and $\nu > 0$. Then for any $b \in \mathbb{R}$ and $a > 0$, it holds that*

$$\mathbb{E}\frac{1}{\sqrt{2\pi a^2}}\exp\left\{-\frac{(X-b)^2}{2a^2}\right\} = \sqrt{\frac{\nu^2}{a^2+\nu^2}}\exp\left\{-\frac{(\mu-b)^2}{2(a^2+\nu^2)}\right\}.$$

Proof. Because $X \sim \mathcal{N}(\mu, \nu^2)$,

$$
\begin{aligned}
&\sqrt{2\pi\nu^2}\mathbb{E}\exp\left\{-\frac{(X-b)^2}{2a^2}\right\} \\
&= \int \exp\left\{-\frac{(x-\mu)^2}{2\nu^2} - \frac{(x-b)^2}{2a^2}\right\} \mathrm{d}x \\
&= \int \exp\left\{-\frac{(a^2+\nu^2)x^2 - 2(a^2\mu+\nu^2 b)x + a^2\mu^2 + \nu^2 b^2}{2a^2\nu^2}\right\} \mathrm{d}x \\
&= \int \exp\left\{-\frac{1}{2a^2\nu^2}\left[(a^2+\nu^2)\left(x - \frac{a^2\mu+\nu^2 b}{a^2+\nu^2}\right)^2 - \frac{(a^2\mu+\nu^2 b)^2}{a^2+\nu^2} + \nu^2 b^2 + a^2\mu^2\right]\right\} \mathrm{d}x \\
&= \exp\left\{-\frac{(\mu-b)^2}{2(a^2+\nu^2)}\right\} \int \exp\left\{-\frac{a^2+\nu^2}{2a^2\nu^2}\left(x - \frac{a^2\mu+\nu^2 b}{a^2+\nu^2}\right)^2\right\} \mathrm{d}x \\
&= \exp\left\{-\frac{(\mu-b)^2}{2(a^2+\nu^2)}\right\} \sqrt{\frac{2\pi a^2\nu^2}{a^2+\nu^2}}.
\end{aligned}
$$

The proposition is then proved by multiplying both sides by $\sqrt{2\pi a^2/\nu^2}$. □

We now consider the likelihood $p(y, x_{\text{obs}}; \beta, \Sigma)$. Integrating out the missing parts x_{mis} we have

$$
\begin{aligned}
p(y, x_{\text{obs}}; \beta, \Sigma) &= p(x_{\text{obs}}) \int \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{-\frac{(y - x_{\text{obs}}^\top \beta_{\text{obs}} - x_{\text{mis}}^\top \beta_{\text{mis}})^2}{2\sigma\varepsilon^2}\right\} \mathrm{d}P(x_{\text{mis}}|x_{\text{obs}}) \\
&= p(x_{\text{obs}})\mathbb{E}_u\left[\exp\left\{-\frac{(y - x_{\text{obs}}^\top \beta_{\text{obs}} - u)^2}{2\sigma_\varepsilon^2}\right\}\middle| x_{\text{obs}}\right],
\end{aligned}
$$

where $u = x_{\text{mis}}^\top \beta_{\text{mis}}$ follows conditional distribution $u|x_{\text{obs}} \sim \mathcal{N}(\mu, \nu^2)$ with $\mu = x_{\text{obs}}^\top \Sigma_{12}\Sigma_{22}^{-1}\beta_{\text{mis}}$ and $\nu^2 = \beta_{\text{mis}}^\top \Sigma_{22:1}\beta_{\text{mis}}$. Applying Proposition 4 with $a = \sigma$ and $b = y - x_{\text{obs}}^\top \beta_{\text{obs}}$, we have

$$
\begin{aligned}
&\mathbb{E}_u\left[\exp\left\{-\frac{(y - x_{\text{obs}}^\top \beta_{\text{obs}} - u)^2}{2\sigma_\varepsilon^2}\right\}\middle| x_{\text{obs}}\right] \\
&\qquad = \frac{1}{\sqrt{2\pi(\sigma_\varepsilon^2 + \beta_{\text{mis}}^\top \Sigma_{22:1}\beta_{\text{mis}})}} \exp\left\{-\frac{(y - x_{\text{obs}}^\top \beta_{\text{obs}} - \beta_{\text{mis}}^\top \Sigma_{21}\Sigma_{11}^{-1}x_{\text{obs}})^2}{2(\sigma_\varepsilon^2 + \beta_{\text{mis}}^\top \Sigma_{22:1}\beta_{\text{mis}})}\right\}.
\end{aligned}
$$

Finally, $R \perp\!\!\!\perp x$, $x_{\text{obs}} \sim \mathcal{N}_q(0, \Sigma_{11})$ and hence

$$
p(x_{\text{obs}}) = \rho^q(1-\rho)^{p-q} \cdot \frac{1}{\sqrt{(2\pi)^q|\Sigma_{11}|}} \exp\left\{-\frac{1}{2}x_{\text{obs}}^\top \Sigma_{11}^{-1}x_{\text{obs}}\right\}.
$$

E.3 Proof of Lemma 8

We prove this lemma by discussing three cases separately when at least one covariate of x_{s-1} and x_j are missing. Assume in each case Σ_0 and Σ_1 are partitioned as in Lemma 7; that is, $\Sigma_0 = [\Sigma_{011} \;\; \Sigma_{012}; \Sigma_{021} \;\; \Sigma_{022}]$ and $\Sigma_1 = [\Sigma_{111} \;\; \Sigma_{112}; \Sigma_{121} \;\; \Sigma_{122}]$.

1. *Both x_{s-1} and x_j are missing.* In this case $\Sigma_{011} = \Sigma_{111} = I_{q\times q}$ and $\Sigma_{012} = \Sigma_{112} = \Sigma_{021}^\top = \Sigma_{121}^\top = 0_{q\times(p-q)}$. Therefore, $\Sigma_{011} = \Sigma_{111}$ and the first two terms in $p(y, x_{\text{obs}}; \beta^*, \Sigma_0)$ and $p(y, x_{\text{obs}}; \beta_1, \Sigma_1)$ are identical. In addition, $\Sigma_{022:1} = \Sigma_{022} = I - \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top)$ and

$\Sigma_{122:1} = \Sigma_{122} = I + \gamma(e_{s-1}e_j^\top + e_j e_{s-1}^\top)$. Subsequently, $\beta_{0\text{mis}}^\top \Sigma_{022:1}\beta_{0\text{mis}} = \|\beta_{0\text{mis}}\|_2^2 - 2\gamma\beta_{0,s-1}\beta_{0j} = \|\beta_{0\text{mis}}\|_2^2 - 2\widetilde{a}^2\gamma^2$, $\beta_{1\text{mis}}^\top \Sigma_{122:1}\beta_{1\text{mis}} = \|\beta_{1\text{mis}}\|_2^2 + 2\gamma\beta_{1,s-1}\beta_{1j} = \|\beta_{1\text{mis}}\|_2^2 - 2\widetilde{a}^2\gamma^2$. Because $\|\beta_{0\text{mis}}\|_2^2 = \|\beta_{1\text{mis}}\|_2^2$ regardless of which covariates are missing, we have that $\beta_{0\text{mis}}^\top \Sigma_{022:1}\beta_{0\text{mis}} = \beta_{1\text{mis}}^\top \Sigma_{122:1}\beta_{1\text{mis}}$ and hence the last term in $p(y, x_{\text{obs}}; \beta^*, \Sigma_0)$ and $p(y, x_{\text{obs}}; \beta_1, \Sigma_1)$ are identical, because $\beta_{0\text{mis}}^\top \Sigma_{021}\Sigma_{011}^{-1} = \beta_{1\text{mis}}^\top \Sigma_{121}\Sigma_{111}^{-1} = 0$ and $\beta_{0\text{obs}} = \beta_{1\text{obs}}$ when x_j is missing.

2. *x_{s-1} is observed but x_j is missing.* In this case, $\Sigma_{011} = \Sigma_{111} = I_{q\times q}$, $\Sigma_{022} = \Sigma_{122} = I_{(p-q)\times(p-q)}$, $\Sigma_{012} = \Sigma_{021}^\top = -\gamma e_{s-1}e_j^\top$ and $\Sigma_{112} = \Sigma_{121}^\top = \gamma e_{s-1}e_j^\top$. Therefore, $\Sigma_{011} = \Sigma_{111} = I$ and hence the first two terms in the likelihood are identical. In addition, $\Sigma_{022:1} = I - \gamma^2 e_j e_j^\top = \Sigma_{122:1}$ and hence $\beta_{0\text{mis}}^\top \Sigma_{022:1}\beta_{0\text{mis}} = \beta_{1\text{mis}}^\top \Sigma_{122:1}\beta_{1\text{mis}} = \|\beta_{\text{mis}}\|_2^2 - \widetilde{a}^2\gamma^4$. Finally, $\beta_{0\text{obs}} = \beta_{1\text{obs}}$ when x_j is missing and $\beta_{0\text{mis}}^\top \Sigma_{021}\Sigma_{011}^{-1} = \beta_{1\text{mis}}^\top \Sigma_{121}\Sigma_{111}^{-1} = -\widetilde{a}^2\gamma^2$. Therefore the last term in both likelihoods are the same as well.

3. *x_j is observed but x_{s-1} is missing.* In this case, $\Sigma_{011} = \Sigma_{111} = I_{q\times q}$, $\Sigma_{022} = \Sigma_{122} = I_{(p-q)\times(p-q)}$, $\Sigma_{012} = \Sigma_{021}^\top = -\gamma e_j e_{s-1}^\top$ and $\Sigma_{112} = \Sigma_{121}^\top = \gamma e_j e_{s-1}^\top$. Therefore, $\Sigma_{011} = \Sigma_{111} = I$ and hence the first two terms in the likelihood are identical. In addition, $\Sigma_{022:1} = I - \gamma^2 e_{s-1}e_{s-1}^\top = \Sigma_{122:1}$ and hence $\beta_{0\text{mis}}^\top \Sigma_{022:1}\beta_{0\text{mis}} = \beta_{1\text{mis}}^\top \Sigma_{122:1}\beta_{1\text{mis}} = \|\beta_{\text{mis}}\|_2^2 - \widetilde{a}^2\gamma^2$. Finally, $\beta_{0\text{obs}}^\top x_{\text{obs}} + \beta_{0\text{mis}}^\top \Sigma_{021}\Sigma_{011}^{-1}x_{\text{obs}} = \beta_{0\text{obs},<s}^\top x_{\text{obs},<s} + \beta_{0j}x_j - \gamma\beta_{0,s-1}x_j = \beta_{0\text{obs},<s}^\top x_{\text{obs},<s}$ because $\beta_{0j} = \widetilde{a}\gamma$ and $\beta_{0,s-1} = \widetilde{a}$. Similarly, $\beta_{1\text{obs}}^\top x_{\text{obs}} + \beta_{1\text{mis}}^\top \Sigma_{121}\Sigma_{111}^{-1}x_{\text{obs}} = \beta_{1\text{obs},<s}^\top x_{\text{obs},<s} + \beta_{1j}x_j + \gamma\beta_{1,s-1}x_j = \beta_{1\text{obs},<s}^\top x_{\text{obs},<s}$. Because $\beta_{0\text{obs},<s} = \beta_{1\text{obs},<s}$, we conclude that the last term of both likelihoods are the same.

E.4 Proof of Lemma 9

We first prove the upper bound for $\|r_n\|_\infty$. By Hölder's inequality,

$$\|r_n\|_\infty \leqslant \sqrt{n}\|\widehat{\Theta}\widetilde{\Sigma} - I\|_\infty \|\widehat{\beta}_n - \beta^*\|_1 \leqslant \sqrt{n}\widetilde{\nu}_n\|\widehat{\beta}_n - \beta^*\|_1,$$

where the last inequality is due to Eq. (9).

We next focus on $\|\widetilde{r}_n\|_\infty$. Apply Hölder's inequality and triangle inequality:

$$\begin{aligned}\|\widetilde{r}_n\|_\infty &\leqslant \sqrt{n}\|\widehat{\Theta} - \Sigma_0^{-1}\|_{L_\infty}\left(\|\Delta_n\beta^*\|_\infty + \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty\right)\\ &\leqslant 2\sqrt{n}b_0b_1\widetilde{\nu}_n\left(\|\Delta_n\beta^*\|_\infty + \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty\right).\end{aligned}$$

Here in the second line we invoke the conclusion in Lemma 1. It then suffices to upper bound $\|\Delta_n\beta^*\|_\infty$ and $\|\frac{1}{n}\widetilde{X}^\top\varepsilon\|_\infty$. With Definition 1, it holds with probability $1 - o(1)$ that

$$\begin{aligned}\|\Delta_n\beta^*\|_\infty &\leqslant \left\|\left(\frac{1}{n}\widetilde{X}^\top X - \Sigma_0\right)\beta^*\right\|_\infty + \left\|\left(\widetilde{\Sigma} - \Sigma_0\right)\beta^*\right\|_\infty\\ &\leqslant \left[\varphi_{u,v}\left(\frac{1}{n}\widetilde{X}^\top X, \Sigma_0; \log p\right) + \varphi_{u,v}\left(\widetilde{\Sigma}, \Sigma_0; \log p\right)\right]\|\beta^*\|_2\end{aligned}$$

and

$$\left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty \leqslant \sigma_\varepsilon\varphi_{\varepsilon,\infty}\left(\frac{1}{n}\widetilde{X}\right).$$

By Lemmas 2 and 3, if $\frac{\log p}{\rho_*^4 n} \to 0$ then

$$\|\Delta_n\beta^*\|_\infty \leqslant O_{\mathbb{P}}\left\{\sigma_x^2\|\beta^*\|_2\sqrt{\frac{\log p}{\rho_*^2 n}}\right\} \quad \text{and} \quad \left\|\frac{1}{n}\widetilde{X}^\top\varepsilon\right\|_\infty \leqslant O_{\mathbb{P}}\left\{\sigma_x\sigma_\varepsilon\sqrt{\frac{\log p}{\rho_* n}}\right\}.$$

F Tail inequalities

Lemma 19 (Sub-Gaussian concentration inequality)**.** *Suppose X is a univariate sub-Gaussian random variable with parameter $\sigma > 0$; that is, $\mathbb{E}X = 0$ and $\mathbb{E}e^{tX} \leqslant e^{\sigma^2 t^2/2}$ for all $t \in \mathbb{R}$. Then*

$$\Pr\left[|X| \geqslant \epsilon\right] \leqslant 2e^{-\frac{\epsilon^2}{2\sigma^2}}, \qquad \forall t > 0;$$
$$\mathbb{E}|X|^r \leqslant r \cdot 2^{r/2} \cdot \sigma^r \cdot \Gamma\left(\frac{r}{2}\right), \qquad \forall r = 1, 2, \cdots$$

Lemma 20 (Sub-exponential concentration inequality)**.** *Suppose $X_1, \cdots, X_n$ are i.i.d. univariate sub-exponential random variables with parameter $\lambda > 0$; that is, $\mathbb{E}X_i = 0$ and $\mathbb{E}e^{tX_i} \leqslant e^{t^2\lambda^2/2}$ for all $|t| \leqslant 1/\lambda$. Then*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right] \leqslant 2\exp\left\{-\frac{n}{2}\min\left(\frac{\epsilon^2}{\lambda^2}, \frac{\epsilon}{\lambda}\right)\right\}.$$

Lemma 21 (Hoeffding inequality)**.** *Suppose $X_1, \cdots, X_n$ are independent univariate random variables with $X_i \in [a_i, b_i]$ almost surely. Then for all $t > 0$, we have that*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i\right| > t\right] \leqslant 2\exp\left\{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

Lemma 22 (Bernstein inequality, support condition)**.** *Suppose $X_1, \cdots, X_n$ are independent random variables with zero mean and finite variance. If $|X_i| \leqslant M < \infty$ almost surely for all $i = 1, \cdots, n$, then*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > t\right] \leqslant 2\exp\left\{-\frac{\frac{1}{2}n^2t^2}{\sum_{i=1}^n \mathbb{E}X_i^2 + \frac{1}{3}Mnt}\right\}, \qquad \forall t > 0.$$

Lemma 23 (Bernstein inequality, moment condition)**.** *Suppose $X_1, \cdots, X_n$ are independent random variables with zero mean and $\mathbb{E}|X_i|^2 \leqslant \sigma^2 < \infty$. Assume in addition that there exists some positive number $L > 0$ such that*

$$\mathbb{E}|X_i|^k \leqslant \frac{1}{2}\sigma^2 L^{k-2} k!, \qquad \forall k > 1.$$

Then we have that

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > t\right] \leqslant 2\exp\left\{-\frac{nt^2}{2(\sigma^2 + Lt)}\right\}, \qquad \forall t > 0.$$

Lemma 24 ([S18])**.** *Suppose $X = (X_1, \cdots, X_p)$ is a p-dimensional zero-mean sub-Gaussian random vector; that is, there exists $\sigma > 0$ such that*

$$\mathbb{E}\exp\left\{\alpha^\top X\right\} \leqslant \exp\left\{\|\alpha\|_2^2\sigma^2/2\right\}, \qquad \forall \alpha \in \mathbb{R}^p.$$

Let A be a $p \times p$ positive semi-definite matrix. Then for all $t > 0$,

$$\Pr\left[X^\top AX > \sigma^2\left(\mathrm{tr}(A) + 2\sqrt{\mathrm{tr}(A^2)t} + 2\|A\|_{\mathrm{op}}t\right)\right] \leqslant e^{-t}.$$

References

[S42] Cai, T., Liu, W., & Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607.

[S25] Loh, P.-L., & Wainwright, M. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, *40*(3), 1637–1664.

[S44] van de Geer, S. (2010). *Empirical Processes in M-Estimation*. Cambridge University Press.