Semi-parametric efficiency bounds and efficient estimation for high-dimensional models

Sara van de Geer Jana Janková

ETH Zürich

Abstract

Asymptotic lower bounds for estimation play a fundamental role in assessing the quality of statistical procedures. In this paper we consider the possibility of establishing semi-parametric efficiency bounds for high-dimensional models and construction of estimators reaching these bounds. We propose a local uniform asymptotic unbiasedness assumption for high-dimensional models and derive explicit lower bounds on the variance of any asymptotically unbiased estimator. We show that an estimator obtained by de-sparsifying (or de-biasing) an ℓ_1-penalized M-estimator is asymptotically unbiased and achieves the lower bound on the variance: thus it is asymptotically efficient. In particular, we consider the linear regression model, Gaussian graphical models and Gaussian sequence models under mild conditions.

Furthermore, motivated by the results of Le Cam on local asymptotic normality, we show that the de-sparsified estimator converges to the limiting normal distribution with zero mean and the smallest possible variance not only pointwise, but locally uniformly in the underlying parameter. This is achieved by deriving an extension of Le Cam's Lemma to the high-dimensional setting.

1 Introduction

Following the development of numerous efficient methods for high-dimensional estimation, more recently the need for statistical inference has emerged. The major approach to estimation in high-dimensions is based on regularized M-estimators, where the regularization is in terms of the ℓ_1-penalty. This approach produces near-oracle estimators under sparsity conditions on the high-dimensional parameter. However, in contrast to the low-dimensional setting, it does not easily yield estimators which are asymptotically normal. This is essentially due to the regularization which introduces bias by shrinking all coefficients towards zero. One stream of work then concentrates on "de-sparsifying" or "de-biasing" ([11], [9], [3]). This approach uses the ℓ_1-regularized M-estimator as an initial estimator and implements a bias correction step. This has been in particular studied for the linear model, generalized linear models and some special cases of non-linear models such as Gaussian graphical models. The work in essence shows an important result: an asymptotically normal estimator for low-dimensional parameters can be constructed.

Further questions being studied concern optimality properties of these de-biased estimators. In particular, what are lower bounds on the rate of convergence attainable in the supremum norm and whether the constructed estimators achieve these optimal rates (see [2] and [3]). The results reveal several things. Firstly, the parametric rate $1/\sqrt{n}$ can be achieved. This basically follows directly from the asymptotic normality of the de-biased estimators, under sufficient sparsity which is of small order $\sqrt{n}/\log p$. Naturally, the parametric rate is optimal: it cannot be improved in order (as was also shown in [2]). On the other hand, if there is insufficient sparsity, in particular when $s \geq n/\log p$, the minimax lower bounds diverge. This is no surprise as oracle inequalities for certain M-estimators have only been shown under the condition $s = o(n/\log p)$. In the intermediate sparsity regime when $\sqrt{n}/\log p \leq s < n/\log p$, the parametric rate cannot be achieved. However, as we discuss in Section 7.3, the sparsity condition $s = o(\sqrt{n}/\log p)$ is essentially necessary for asymptotically normal estimation. Thus the analysis revealed that under sufficient sparsity of small order $\sqrt{n}/\log p$, the parametric rate of order $1/\sqrt{n}$ is optimal, and the de-biased estimator achieves it (in the above mentioned settings).

However, the analysis on minimax rates does not address an important question. The parametric bounds derived do not reveal any explicit lower bounds on the variance. The question of efficiency in the spirit of the famous Cramér-Rao result thus remains open in the high-dimensional setting. This motivates us to pose the following questions. Can we establish lower bounds on the variance, similar to the Cramér-Rao bounds in the (semi-)parametric setting, also in the high-dimensional setting? And if yes, can we construct an estimator that achieves these bounds?

We give an affirmative answer to these questions. We propose Cramér-Rao type bounds on the variance for sparse high-dimensional models. To this end, we propose a uniform asymptotic unbiasedness assumption. This basically measures the rate at which the bias vanishes in shrinking neighbourhoods of the true distribution P_0 of size $1/\sqrt{n}$. We then present a lower bound on the variance of sequences of estimators that are uniformly asymptotically unbiased. This essentially means that we obtain explicit minimax lower bounds.

We further show that one can construct an asymptotically unbiased estimator, which achieves the lower bound. As one might expect, this is the de-biased estimator or an estimator that is in some sense asymptotically equivalent to the de-biased estimator. Thus, compared to previous results, which only showed asymptotic normality or minimaxity (up to order in n) of the de-biased estimator, we show that the de-biased estimator is the best among all asymptotically unbiased estimators: thus in this sense asymptotically efficient.

Furthermore, we extend the work of Le Cam on asymptotic efficiency ([5]) to the high-dimensional setting. In particular, we show that the de-sparsified estimator converges locally uniformly to the limiting normal distribution with zero mean and the smallest possible variance. This involves a careful adjustment of Le Cam's arguments to the high-dimensional setting.

As a by-product of our analysis, we establish new oracle results for the Lasso which hold in expectation. These are needed to claim strong asymptotic unbiasedness of certain de-sparsified estimators.

2 Notation

For a vector $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ we denote its ℓ_p norm by $\|x\|_p := (\sum_{i=1}^p x_i^p)^{1/p}$ for $p = 1, 2, \ldots$. We further let $\|x\|_\infty := \max_{i=1,\ldots,p} |x_i|$ and $\|x\|_0 = |\{i : i \in \{1, \ldots, p\}, x_i = 0\}|$. We denote $\|x\|_n^2 := \|x\|_2^2/n$.

By e_i we denote a p-dimensional vector of zeros with one at position i. For a matrix $A \in \mathbb{R}^{m \times n}$ we let $\|A\|_\infty := \max_{i=1,\ldots,m, j=1,\ldots,n} |A_{ij}|$, $|||A|||_1 := \max_{i=1,\ldots,m} \sum_{j=1}^n |A_{ij}|$. We denote its j-th column by A_j, which is a column vector. We recall here that for symmetric matrices $A, B \in \mathbb{R}^{p \times p}$ it holds that $\text{vec}(A)^T \text{vec}(B) = \text{tr}(AB)$, where $\text{vec}(A)$ is the vectorized version of a matrix A. For matrices $A, B, C \in \mathbb{R}^{p \times p}$, it holds that $A \otimes B \, \text{vec}(C) = \text{vec}(A^T C B)$, where $\otimes$ denotes the Kronecker product. By $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ we denote the minimum and maximum eigenvalue of A, respectively.

For real sequences f_n, g_n, we write $f_n = \mathcal{O}(g_n)$ if $|f_n| \leq C|g_n|$ for some $C > 0$ independent of n and all n. We write $f_n \asymp g_n$ if both $f_n = \mathcal{O}(g_n)$ and $1/f_n = \mathcal{O}(g_n)$ hold. Finally, $f_n = o(g_n)$ if $\lim_{n\to\infty} f_n/g_n = 0$.

We use $\rightsquigarrow$ to denote the convergence in distribution. By $\Phi(\cdot)$ we denote the cumulative distribution function of a standard normal random variable.

3 Organization of the paper

The paper consists of two main parts. In the first part we develop an asymptotic version of a semi-parametric Cramér-Rao lower bound for high-dimensional models. In particular, we consider the linear model, the Gaussian graphical model and the Gaussian sequence model. For each of these models, we establish lower bounds on the variance of any strongly asymptotically unbiased estimator. Consequently, we give a construction of a strongly asymptotically unbiased estimator which is asymptotically efficient, i.e. it reaches the derived lower bound.

The particular sections are divided as follows. In Section 4 we state preliminary results on strong oracle inequalities for the Lasso. In Section 5 we propose a strong asymptotic unbiasedness assumption. Section 6 gives lower bounds on the variance in the linear model, considering random and fixed design. In Section 7 we propose an estimator that is asymptotically efficient for the considered linear model. In Section 8 we derive lower bounds on the variance in Gaussian graphical models. Section 9 then gives a construction of an asymptotically efficient estimator for Gaussian graphical models. In the second part of the paper we extend the results of Le Cam's to the high-dimensional setting, which shows that the desparsified estimator is locally uniform converging to the limiting distribution with zero mean and the smallest possible variance. This extension in contained in Section 11. Finally, the proofs are contained in Section 13.

4 Strong oracle inequalities for the Lasso

We present new results on oracle inequalities for the Lasso in linear regression which will be needed in subsequent sections, but can also be of independent interest. Typical high-dimensional analysis derives oracle inequalities for the Lasso which hold with high-probability (see [1] for an overview of such results). We derive stronger oracle inequalities that hold in expectation.

Consider the linear model

$$Y = X\beta_0 + \epsilon, \tag{1}$$

where X is the $n\times p$ design matrix with independent rows $X_i, i = 1, \ldots, n$, Y is the $n\times 1$ vector of observations and the (unobservable) error $\epsilon \in \mathbb{R}^n$ satisfies $\mathbb{E}\epsilon = 0$ and ϵ_i are independent for $i = 1, \ldots, n$. Moreover, the error ϵ and the design matrix X are independent. The vector $\beta_0 \in \mathbb{R}^p$ is unknown, but assumed to only have s non-zero entries. The quantity s is called the sparsity of β_0.

Consider the Lasso estimator with a tuning parameter λ defined as follows:

$$\hat\beta := \arg\min_{\beta\in\mathbb{R}^p} \|Y - X\beta\|_2^2/n + 2\lambda\|\beta\|_1. \tag{2}$$

The following result shows that the error $\mathbb{E}\|\hat\beta - \beta_0\|_1$ is up to a logarithmic factor of the same order as the oracle error $\mathbb{E}\|\beta_{ora} - \beta_0\|_1 = \mathcal{O}(s/\sqrt{n})$, where β_{ora} is the oracle estimator (i.e. it has knowledge of the set of non-zero entries of β_0). Theorem 1 presented below is actually more general in that it considers the k-th order errors $\mathbb{E}\|\hat{\beta} - \beta_0\|_1^k$ for any fixed $k \in \{1, 2, \dots\}$.

Theorem 1. *Assume the linear model in* (1) *with* $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$*, where* $\sigma_\epsilon^2 = \mathcal{O}(1)$*. Suppose that* $X_i \sim \mathcal{N}(0, \Sigma_0)$ *are independent for* $i = 1, \dots, n$*, where* $\|\Sigma_0\|_\infty = \mathcal{O}(1)$ *and* $\Lambda_{\min}(\Sigma_0) \geq L > 0$ *for a universal constant* L*. Suppose that* $\|\beta_0\|_2 = \mathcal{O}(1)$*,* $\|\beta_0\|_0 \leq s$ *and* $s \log p/n = o(1)$*. Let* $k \in \{1, 2, \dots\}$ *be fixed and let* $\tau > 0$ *fixed be such that* $p^{-\tau/2} = \mathcal{O}((s\lambda^2)^{k/2})$*. Consider the Lasso* $\hat{\beta}$ *defined in* (2) *with tuning parameter* $\lambda \geq c\tau\sqrt{\log p/n}$*, where* c *is a sufficiently large universal constant. Then there exist universal constants* C_1, C_2 *such that*

$$(\mathbb{E}\|\hat{\beta} - \beta_0\|_1^k)^{1/k} \leq C_1 s\lambda.$$

Moreover, for any $\nu > 0$ *it holds with probability at least* $1 - 1/\nu^k$

$$\|\hat{\beta} - \beta_0\|_1 \leq \nu C_1 s\lambda.$$

Theorem 1 is proved in Section 13.1. Taking $k = 1$, under the conditions of Theorem 1 we obtain

$$\mathbb{E}\|\hat{\beta} - \beta_0\|_1 \leq C_1 s\lambda.$$

Moreover, Theorem 1 can be extended to the situation when the errors ϵ_i are independent and sub-Gaussian (with a universal constant) and the design X has independent sub-Gaussian rows (with a universal constant). It can also be easily extended to fixed design, under a compatibility condition on the sample covariance matrix $X^T X/n$.

We note that to obtain strong oracle inequalities for higher powers of the error $\|\hat{\beta} - \beta_0\|_1$, we need to keep increasing τ (because of the condition $p^{-\tau/2} = \mathcal{O}((s\lambda^2)^{k/2})$, where k is the power). However, the regularization parameter λ depends on τ, in particular $\lambda \geq c\tau\sqrt{\log p/n}$. Hence the higher order of error we want to control, the stronger regularization must be chosen.

5 Local uniform asymptotic unbiasedness

In this section we introduce a local uniform asymptotic unbiasedness assumption. Consider the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$, where P_θ is a probability distribution for every $\theta \in \Theta \subset \mathbb{R}^p$. Assume that $P_\theta \in \mathcal{P}$ is dominated by some σ-finite measure for all θ and denote the corresponding probability densities by p_θ. The log-likelihood will be denoted by $\ell_\theta := \log p_\theta$. Further denote the score function by $s_\theta := \frac{\partial \ell_\theta}{\partial \theta}$ and let $I_\theta = \mathbb{E}_\theta s_\theta s_\theta^T$.

Let $g : \mathbb{R}^p \to \mathbb{R}$ and let the parameter of interest be $g(\theta_0)$. Our goal is to derive an asymptotic lower bound for the variance of an estimator T_n of $g(\theta)$, which is in some sense asymptotically unbiased. To this end, we define strong asymptotic unbiasedness as presented below.

Definition 1. *Let* $a \in \mathbb{R}^p$ *and let* $0 < \delta_n \downarrow 0$*. We call* T_n *a strongly asymptotically unbiased estimator of* $g(\theta_0)$ *at* θ_0 *in the direction* a *with rate* δ_n *if for* $m_n := n/\delta_n$ *and for* $\theta := \theta_0 + a/\sqrt{m_n}$ *and for* $\theta := \theta_0$ *it holds that*

$$\sqrt{m_n}(\mathbb{E}_\theta T_n - g(\theta)) = o(1).$$

The motivation for this definition comes from the asymptotic unbiasedness assumption for semi-parametric models. In particular, we consider shrinking neighbourhoods of θ_0 of size $1/\sqrt{n}$, where we require the bias to vanish at a rate $1/\sqrt{n}$. Note that if $\sqrt{n}(\mathbb{E}_\theta(T_n) - g(\theta)) =$

$o(1)$, then one may take e.g. $\delta_n := \sqrt{n}(\mathbb{E}_\theta(T_n) - g(\theta))$. Definition 1 is particularly useful when recognizing the concept of a worst possible sub-direction, as will be discussed later on. Further we consider the following notion which assumes strong asymptotic unbiasedness in every direction within the considered sparse model.

Definition 2. *We say that T_n is strongly asymptotically unbiased for estimation of $g(\theta)$ if for all $\theta \in \Theta$ and $a \in \Theta$ it holds that*

$$\sqrt{n}\left(\mathbb{E}_{\theta+a/\sqrt{n}}T_n - g\left(\theta + \frac{a}{\sqrt{n}}\right)\right) = o(1).$$

6 Lower bounds for the linear model

In this part, we derive lower bounds for the variance of a strongly asymptotically unbiased estimator in a high-dimensional linear regression model. Consider the linear model (1) with $\epsilon \sim \mathcal{N}_n(0, I)$. In the following sections we look first at the case of random Gaussian design matrix and then a fixed design matrix.

6.1 Linear model with random design

Assume that X is a random $n \times p$ matrix independent of ϵ with independent rows $X_i \sim \mathcal{N}(0, \Sigma_0)$ for $i = 1, \ldots, n$. We assume the inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists.

Theorem 2. *Let $a \in \mathbb{R}^p$ be such that $a^T\Sigma_0 a = 1$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta_0)$ at β_0 in the direction a with rate δ_n. Assume moreover that for some $\dot{g}(\beta_0) \in \mathbb{R}^p$ and for $m_n = n/\delta_n$*

$$\sqrt{m_n}\left(g(\beta_0 + a/\sqrt{m_n}) - g(\beta_0)\right) = a^T\dot{g}(\beta_0) + o(1). \tag{3}$$

Then

$$n\text{var}(T_n) \geq [a^T\dot{g}(\beta_0)]^2 - o(1).$$

The condition (3) is a differentiability condition on g. By maximizing the lower bound $[a^T\dot{g}(\beta_0)]^2$ over all a such that $a^T\Sigma_0 a = 1$, we obtain the following corollary.

Corollary 1. *The lower bound $[a^T\dot{g}(\beta_0)]^2$ is maximized at the value*

$$a_0 := \Theta_0\dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T\Theta_0\dot{g}(\beta_0)}.$$

Hence under the conditions of Theorem 2, we get

$$n\text{var}(T_n) \geq \dot{g}(\beta_0)^T\Theta_0\dot{g}(\beta_0) - o(1).$$

Definition 3. *Let g be differentiable at β_0 with derivative $\dot{g}(\beta_0)$. We call*

$$c_0 := \Theta_0\dot{g}(\beta_0)/\dot{g}(\beta_0)^T\Theta_0\dot{g}(\beta_0)$$

the worst possible sub-direction for estimating $g(\beta_0)$.

The motivation for the terminology in Definition 3 is given by Corollary 1. The normalization by $\dot{g}(\beta_0)^T\Theta_0\dot{g}(\beta_0)$ is arbitrary but natural from a projetion theory point of view.

As a special case, consider $g(\beta_0) = \beta_j^0$ for some fixed value of j. Then $\dot{g}(\beta_0) = e_j$, the j-th unit vector in $\mathbb{R}^p$. Clearly, $\Theta_0 e_j = \Theta_j^0$ and $e_j^T \Theta_0 e_j = \Theta_{jj}^0$, where Θ_j^0 is the j-th column of Θ_0 and Θ_{jj}^0 is its j-th diagonal element. It follows that $C_j^0 = \Theta_j^0 / \Theta_{jj}^0$ is the worst possible sub-direction for estimating β_j^0. Corollary 1 implies the lower bound $n\text{var}(T_n) \geq \Theta_{jj} + o(1)$.

Finally, we reformulate Corollary 1 in view of Definition 2. In particular, we assume that T_n is strongly asymptotically unbiased in all directions $a \in \mathcal{B}$, where $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s, \|\beta\|_2 = \mathcal{O}(1)\}$.

Corollary 2. *Let T_n be a strongly asymptotically unbiased estimator of $g(\beta_0)$, and for all $\beta_0 \in \mathcal{B}, a \in \mathcal{B}$ it holds*

$$\sqrt{n}\left(g(\beta_0 + a/\sqrt{n}) - g(\beta_0)\right) = a^T \dot{g}(\beta_0) + o(1).$$

Suppose that $\Theta_0 \dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)} \in \mathcal{B}$ for all $\beta_0 \in \mathcal{B}$ and suppose that $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$. Then it holds

$$n\text{var}_{\beta_0}(T_n) \geq \dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0) - o(1).$$

6.2 Linear model with fixed design

In this section, we assume that the design matrix X is fixed (non-random). Let $\hat{\Sigma} := X^T X/n$ be the sample covariance matrix.

Theorem 3. *Let $a \in \mathbb{R}^p$ be such that $a^T \hat{\Sigma} a \leq 1 + o(1)$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta_0)$ at β_0 in the direction a with rate δ_n. Assume moreover that for some $\dot{g}(\beta_0) \in \mathbb{R}^p$ and for $m_n := n/\delta_n$ it holds that*

$$\sqrt{m_n}\,(g(\beta_0 + a/\sqrt{m_n}) - g(\beta_0)) = a^T \dot{g}(\beta_0) + o(1). \tag{4}$$

Then

$$n\text{var}(T_n|X) \geq [a^T \dot{g}(\beta_0)]^2 - o(1).$$

The following Lemma gives the lower bound in view of Definition 2. It assumes existence of $\hat{\Theta}$, which is a pseudoinverse of $\hat{\Sigma}$ in the following sense: $\|\hat{\Sigma}\hat{\Theta} - I\|_\infty = \mathcal{O}(\lambda)$. A construction of such an estimator is possible and deferred to Section 7.

Lemma 1. *Suppose that $\hat{\Theta}\dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T \hat{\Theta} \dot{g}(\beta_0)} \in \mathcal{B}$ for all $\beta_0 \in \mathcal{B}$ and suppose that $\|\hat{\Sigma}\hat{\Theta} - I\|_\infty = \mathcal{O}(\lambda)$, where $s\lambda = o(1)$. Let T_n be a strongly asymptotically unbiased estimator of $g(\beta_0)$, and assume that for all $\beta_0, a \in \mathcal{B}$ it holds*

$$\sqrt{n}\left(g(\beta_0 + a/\sqrt{n}) - g(\beta_0)\right) = a^T \dot{g}(\beta_0) + o(1).$$

Further assume that $1/\dot{g}(\beta_0)^T \hat{\Theta} \dot{g}(\beta_0) = \mathcal{O}(1)$ and $\|\dot{g}(\beta_0)\|_1 = \mathcal{O}(s)$. Then it holds

$$n\text{var}(T_n|X) \geq \dot{g}(\beta_0)^T \hat{\Theta} \dot{g}(\beta_0) - o(1).$$

7 An asymptotically efficient estimator in the linear model

In this section we consider the construction of a strongly asymptotically unbiased estimator, which achieves the corresponding lower bound on the variance derived in the previous sections (for fixed and random design). We first consider estimation of single elements $g(\beta) = \beta_j$ for some $j \in \{1, \ldots, p\}$ and later estimation of linear functionals $g(\beta) = \xi^T \beta$, where $\xi \in \mathbb{R}^p$ is known. Further extension to other functionals of interest might be also possible, under some conditions on the transformation g.

The problem of estimation of low-dimensional parameters in high-dimensional linear regression has been studied extensively (see [1] for an overview). The Lasso estimator (see [1]) is a prime example. However, Lasso is biased due to the inclusion of the ℓ_1-penalty. A de-sparsified or de-biased version of the Lasso was then considered (see [9]), which was shown to be asymptotically normal in [9]. Here we consider the de-sparsified Lasso estimator and show that it is strongly asymptotically unbiased. However, the analysis is not limited to this example; other de-sparsified estimators (e.g. one based on the square-root Lasso) or other estimators which are in some sense equivalent are likely to be applicable as well.

We consider the linear model (1) with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, where $\sigma_\epsilon = \mathcal{O}(1)$.

7.1 Linear model with random design

Assume that X in (1) is a random $n \times p$ matrix independent of ϵ with independent sub-Gaussian rows X_i with mean zero and covariance matrix Σ_0. The rows of X will be denoted by $X_i, i = 1, \ldots, n$. We assume the inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists.

Consider the Lasso defined in (2) with $\lambda \asymp \sqrt{\log p/n}$. We further need to construct an estimator of Θ_0. Let $\hat{\Theta}_j$ be an estimate of Θ_j^0 be obtained by solving the following program, that will be referred to as *nodewise regression* (see [9]). Denote by X_{-j} the $n \times (p-1)$ matrix obtained by removing the j-th column from X. For $j = 1, \ldots, p$, let

$$\hat{\gamma}_j := \arg \min_{\gamma \in \mathbb{R}^{p-1}} \|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j \|\gamma\|_1, \tag{5}$$

$$\hat{\tau}_j^2 := \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n,$$

$$\hat{\Theta}_{Lasso,j} := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \ldots, -\hat{\gamma}_{j,p})/\hat{\tau}_j^2, \tag{6}$$

where $\lambda_j \asymp \lambda \asymp \sqrt{\log p/n}$ for $j = 1, \ldots, p$. The necessary Karush-Kuhn-Tucker conditions corresponding to the nodewise regression (obtained by replacing derivatives by sub-differentials) imply the condition $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty = O_P(\lambda)$ (see [9]).

Define the de-biased Lasso introduced in [9] by

$$\hat{b} := \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n, \tag{7}$$

and let $\hat{b}_j$ denote its j-th element.

Estimation of β_j^0

We will next show that $\hat{b}_j$ with $\hat{\Theta}_j$ defined by the nodewise regression is strongly asymptotically unbiased. We need the following assumptions and auxiliary Lemma 2. Recall that $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta_0\|_0 = \mathcal{O}(s), \|\beta_0\|_2 = \mathcal{O}(1)\}$. Condition $\|\beta_0\|_0 = \mathcal{O}(s)$ represents the classical sparsity condition on the high-dimensional parameter (here the sparsity s will be specified later). Condition $\|\beta_0\|_2 = \mathcal{O}(1)$ can be justified in terms of the signal-to-noise ratio being bounded. If the signal-to-noise ratio stays bounded, and the variance of the noise is bounded (as assumed above), then the ℓ_2-norm of β also remains bounded. We further consider the following condition on boundedness of eigenvalues of Σ_0 as follows.

(A1) $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$ and $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$.

Condition (A1) guarantees that the compatibility condition is satisfied and that the nodewise regression yields an oracle estimator (see [9], Theorem 2.4). The following Lemma is a direct consequence of Theorem 1.

Lemma 2. *Suppose that condition (A1) is satisfied and suppose that* $s\log p/n = o(1)$. *Let* $\hat{\beta}$ *be the Lasso estimator defined in* (2) *with a sufficiently large tuning parameter of order* $\sqrt{\log p/n}$. *Then for every* $\beta_0 \in \mathcal{B}$

$$\mathbb{E}_{\beta_0}\|\hat{\beta} - \beta_0\|_1 = \mathcal{O}(s\lambda).$$

We consider estimation of $g(\beta) = \beta_j$ and we show strong asymptotic unbiasedness of the de-biased estimator for estimation of β_j^0. To show strong asymptotic unbiasedness, we need to assume the sparsity assumption $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. This condition is necessary as discussed in Section 7.3.

Lemma 3. *Suppose that condition (A1) is satisfied and suppose that* $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. *Let* $\hat{b}_j$ *be defined as in* (7) *with* $\hat{\Theta}_j$ *satisfying* $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j$. *Then for every* $\beta_0 \in \mathcal{B}$

$$\sqrt{n}\mathbb{E}_{\beta_0}(\hat{b}_j - \beta_j^0) = o(1).$$

Finally we show that the de-biased estimator achieves the lower bound on the variance derived in previous section. Thus the de-sparsified estimator is strongly asymptotically unbiased and has the smallest variance among all strongly asymptotically unbiased estimators.

Theorem 4. *Suppose that condition (A1) is satisfied,* $s = o\left(\frac{\sqrt{n}}{\log p}\right)$ *and that* $\|\Theta_j^0\|_0 = \mathcal{O}(s)$. *Let* $\hat{\Theta}_{Lasso,j}$ *be obtained using the nodewise regression as in* (6). *Then* $\hat{b}_j$ *defined in* (7) *using the nodewise regression is strongly asymptotically unbiased and for any strongly asymptotically unbiased estimator* T *of* β_j^0 *it holds for all* $\beta_0 \in \mathcal{B}$

$$\mathrm{var}(T) \geq \mathrm{var}(\hat{b}_j) = \frac{\Theta_{jj}^0 + o(1)}{n}.$$

Estimation of linear functionals

We consider estimation of linear functionals $g(\beta) = \xi^T\beta$, where $\xi \in \mathbb{R}^p$ is a known vector. In this section, we assume that the design is random, however, similar results might be obtained also for fixed design. We define the de-sparsified estimator for estimation of $\xi^T\beta$ as a linear combination ξ of the de-sparsified estimator $\hat{b}$. This yields

$$\hat{b}_\xi := \xi^T\hat{b} = \xi^T\hat{\Theta}(\hat{\beta} - \beta_0) + \xi^T\hat{\Theta}X^T(Y - X\hat{\beta})/n. \tag{8}$$

Then we have the following Lemma, which shows strong asymptotic unbiasedness of $\hat{b}_\xi$.

Lemma 4. *Suppose that condition (A1) is satisfied and* $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. *Let* $\hat{b}_\xi$ *be the estimator defined in* (8). *Assume that* $\|\xi\|_1 = \mathcal{O}(1)$. *Then for every* $\beta_0 \in \mathcal{B}$ *it holds that*

$$\sqrt{n}\mathbb{E}_{\beta_0}(\hat{b}_\xi - \xi^T\beta_0) = o(1).$$

Theorem 5 shows that the de-biased estimator (8) achieves the lower bound on the variance.

Theorem 5. *Suppose that condition (A1) is satisfied and* $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. *Let* $\hat{b}_\xi$ *be the estimator defined in* (8). *Assume that* $\|\Theta_0\xi\|_0 = \mathcal{O}(s)$ *and* $\|\xi\|_1 = \mathcal{O}(1)$. *Then* $\hat{b}_\xi$ *is strongly asymptotically unbiased and for any strongly asymptotically unbiased estimator* T *of* $\xi^T\beta$ *it holds for all* $\beta_0 \in \mathcal{B}$

$$\mathrm{var}_{\beta_0}(T) \geq \mathrm{var}_{\beta_0}(\hat{b}_\xi) = \frac{\xi^T\Theta_0\xi + o(1)}{n}.$$

7.2 Linear model with fixed design

We consider $X \in \mathbb{R}^{n\times p}$ to be a fixed design matrix. Denote the sample covariance matrix by $\hat{\Sigma} := X^TX/n$. An estimate $\hat{\Theta}$ which is a surrogate inverse for $\hat{\Sigma}$ can be obtained in the same way as for the random design, using the nodewise regression (6). The necessary Karush-Kuhn-Tucker conditions of the nodewise regression (obtained by replacing derivatives by sub-differentials) again imply the condition $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty = O_P(\lambda)$. The de-sparsified estimator can then be defined in the same way as for the random design, as in (7).

We consider estimation of $g(\beta) := \beta_j$, although one could further consider estimation of linear functionals, similarly as for the random design. Strong asymptotic unbiasedness for estimation of β_j^0 then follows similarly as in Lemma 3 for all $\beta_0 \in \mathcal{B}$, under $s = o\left(\frac{\sqrt{n}}{\log p}\right)$ and if the compatibility condition (see [1]) is satisfied for X^TX/n with a universal constant. We omit details of the proof of strong asymptotic unbiasedness of $\hat{b}_j$ under fixed design since the proof is analogous to the proof of Lemma 3. We formulate the lower bound for estimation of $g(\beta) := \beta_j$ in the following theorem.

Theorem 6. *Let* $\hat{\Theta}_j$ *be obtained using the nodewise regression as in* (6). *Suppose that* $s = o\left(\frac{\sqrt{n}}{\log p}\right)$, $\|\hat{\Theta}_j\|_0 = \mathcal{O}(s), \|\hat{\Theta}_j\|_2 = \mathcal{O}(1), 1/\hat{\Theta}_{jj} = \mathcal{O}(1)$ *and that the compatibility condition is satisfied for* X^TX/n *with a universal constant. Then* $\hat{b}_j$ *defined in* (7) *using* $\hat{\Theta}_j$ *is strongly asymptotically unbiased and for any strongly asymptotically unbiased estimator* T *of* β_j^0 *it holds for all* $\beta_0 \in \mathcal{B}$

$$\mathrm{var}_{\beta_0}(T|X) \geq \mathrm{var}_{\beta_0}(\hat{b}_j|X) = \frac{\hat{\Theta}_j^T\hat{\Sigma}\hat{\Theta}_j + o(1)}{n} = \frac{\hat{\Theta}_{jj} + o(1)}{n}.$$

7.3 Discussion on asymptotic efficiency in the linear model

To establish asymptotic efficiency of the de-sparsified estimator, we only considered mild conditions analogous to the conditions assumed in [9]. These include conditions on the boundedness of the spectrum of the precision matrix, boundedness of the signal to noise ratio, boundedness of the error variance, sparsity condition on the parameter β_0 and row-sparsity of the precision matrix.

In particular, our analysis requires the sparsity condition $s = o(\sqrt{n}/\log p)$. However, this condition is essentially necessary in the linear regression setting for construction of confidence intervals, as argued in the following. First observe that if the (slightly weaker) condition $s = \mathcal{O}(\sqrt{n}/\log p)$ is not satisfied, then there cannot exist an estimator T_n of $\beta_j^0 \in \mathbb{R}$ such that for all β_0

$$\sqrt{n}(T_n - \beta_j^0)/\sigma_n \rightsquigarrow \mathcal{N}(0,1), \tag{9}$$

assuming that $\sigma_n = \mathcal{O}(1)$. Suppose that there exists an estimator T_n that satisfies (9). Then necessarily $\sqrt{n}(T_n - \beta_j)/\sigma_n = O_P(1)$. By similar reasoning as in [8], we have under the conditions assumed the minimax rates for $|T - \beta_j^0|$ of order $\frac{1}{\sqrt{n}} + \frac{s \log p}{n}$. But then necessarily $s \log p / n = \mathcal{O}(1/\sqrt{n})$, which gives $s = \mathcal{O}(\sqrt{n}/\log p)$. This is only slightly weaker than the condition we require, $s = o(\sqrt{n}/\log p)$.

We remark that the above results can be easily extended to sub-Gaussian design and sub-Gaussian error. The lower bounds clearly hold also for sub-Gaussian designs. As already pointed out, strong oracle inequalities for sub-Gaussian designs and sub-Gaussian error may be easily derived and used to show strong asymptotic unbiasedness of the de-sparsified estimator.

8 Lower bounds for Gaussian graphical models

In this part, we consider efficient estimation of edge weights in Gaussian graphical models. Gaussian graphical models encode conditional dependencies between variables (nodes in the graph) by including an edge between two variables if and only if they are not independent given all the other variables. This corresponds to the problem of estimation of the precision matrix of a multivariate normal distribution, which we now introduce. Let

$$X_i \sim \mathcal{N}_p(0, \Sigma_0), \quad i = 1, \dots, n,$$

where the X_i's are independent for $i = 1, \dots, n$. Denote the precision matrix by $\Theta_0 := \Sigma_0^{-1}$, where the inverse of Σ_0 is assumed to exist. The matrix $\Theta_0 \in \mathbb{R}^{p \times p}$ is unknown, but assumed to only have row-sparsity (column-sparsity) of order s, i.e. let $\max_{i=1,\dots,p} \|\Theta_i\|_0 \leq s$, where Θ_i is the i-th column of the precision matrix.

There have been numerous methods proposed for estimation of the precision matrix in the high-dimensional setting when $p \gg n$. These methods are based on ℓ_1-regularization and thus lead to biased estimators. De-biasing was then studied similarly as in the linear regression, and it was shown that de-biasing leads to estimators which are asymptotically normal. For our further analysis, we consider the de-sparsified nodewise Lasso estimator proposed in [4].

In the following results, we restrict our attention to estimation of single elements of the precision matrix and linear functionals of the precision matrix.

Let $g : \mathbb{R}^{p \times p} \to \mathbb{R}$ and let the parameter of interest be $g(\Theta_0)$. Let T_n be some estimator of $g(\Theta_0)$.

We give some direct arguments for an asymptotic lower bound for the variance of T_n when T_n is strongly asymptotically unbiased. Contrary to previous parts, the parameter is a matrix, therefore instead of vector direction a we shall write the capital letter A.

8.1 Estimation of elements of the precision matrix

As the first step we consider estimation of $g(\Theta_0) = \Theta_{ij}^0$ for some fixed $(i, j) \in \{1, \dots, p\}^2$. The following Theorem gives a lower bound on the variance of any strongly asymptotically unbiased estimator of Θ_{ij}^0. Define

$$\mathcal{G} := \{\Theta \in \mathbb{R}^{p \times p} : \|\Theta_j\|_0 = \mathcal{O}(s), j = 1, \dots, p, 1/\Lambda_{\min}(\Theta) = \mathcal{O}(1), \Lambda_{\max}(\Theta) = \mathcal{O}(1)\}.$$

Theorem 7. *Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\Theta_0) := \Theta_{ij}^0$. Suppose that $(\Theta_i^0 (\Theta_j^0)^T + \Theta_i^0 (\Theta_j^0)^T)/(2\sigma_{ij}) \in \mathcal{G}$, where where $\sigma_{ij}^2 := (\Theta_{ij}^0)^2 + \Theta_{ii}^0 \Theta_{jj}^0$. Then for all $\Theta_0 \in \mathcal{G}$*

$$n \text{var}(T_n) \geq (\Theta_{ij}^0)^2 + \Theta_{ii}^0 \Theta_{jj}^0 - o(1).$$

Theorem 7 follows from a more general result - Theorem 8 in the next section.

8.2 Estimation of linear functionals

One could be further interested in estimation of linear functions of Θ_0, $h(\Theta_0) = \operatorname{tr}(\Psi\Theta_0)$, where $\Psi \in \mathbb{R}^{p\times p}$ is a known matrix. We shall consider the case when Ψ is of rank one, i.e. estimation of functions $g(\Theta_0) = \xi_1^T\Theta_0\xi_2$, where $\xi_1, \xi_2 \in \mathbb{R}^p$ are known vectors.

Theorem 8. *Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\Theta) = \xi_1^T\Theta\xi_2$ at Θ_0 in the direction $A := \Theta_0(\xi_1\xi_2^T + \xi_2\xi_1^T)\Theta_0/(2\sigma)$, where $\sigma^2 = \xi_1^T\Theta_0\xi_1\xi_2^T\Theta_0\xi_2 + (\xi_1^T\Theta_0\xi_2)^2$, with rate δ_n. Then it holds*

$$n\mathrm{var}_{\Theta_0}(T_n) \geq \xi_1^T\Theta_0\xi_1\xi_2^T\Theta_0\xi_2 + (\xi_1^T\Theta_0\xi_2)^2 - o(1).$$

Corollary 3. *Let T_n be a strongly asymptotically unbiased estimator of $g(\Theta_0) = \xi_1^T\Theta_0\xi_2$. Suppose that $\Theta_0(\xi_1\xi_2^T + \xi_2\xi_1^T)\Theta_0/(2\sigma) \in \mathcal{G}$ for all $\Theta_0 \in \mathcal{G}$. Then for all $\Theta_0 \in \mathcal{G}$ it holds that*

$$n\mathrm{var}_{\Theta_0}(T_n) \geq \xi_1^T\Theta_0\xi_1\xi_2^T\Theta_0\xi_2 + (\xi_1^T\Theta_0\xi_2)^2 - o(1).$$

9 An asymptotically efficient estimator for Gaussian graphical models

We consider the de-sparsified nodewise Lasso estimator introduced in [4] and show that this estimator is strongly asymptotically unbiased and reaches the lower bound on variance. To this end, we consider the following construction, which was proposed in [6]. We recall the construction again, although it is identical to the nodewise regression defined in (6). Denote by X_{-j} the $n \times (p-1)$ matrix obtained by removing the j-th column from X. For $j = 1, \dots, p$, let

$$\hat{\gamma}_j := \arg\min_{\gamma\in\mathbb{R}^{p-1}} \|X_j - X_{-j}\gamma\|_n^2 + 2\lambda\|\gamma\|_1, \tag{10}$$

$$\hat{\tau}_j^2 := \|X_j - X_{-j}\hat{\gamma}_j\|_n^2,$$

$$\hat{\Gamma}_j := (-\hat{\gamma}_{j,1}, \dots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \dots, -\hat{\gamma}_{j,p}),$$

and define the nodewise Lasso estimator

$$\hat{\Theta}_j := \hat{\Gamma}_j/\hat{\tau}_j^2. \tag{11}$$

Define the de-sparsified nodewise Lasso (see [4])

$$\hat{T} := \hat{\Theta} + \hat{\Theta}^T - \hat{\Theta}\hat{\Sigma}\hat{\Theta}. \tag{12}$$

We will show that the $\hat{T}_{ij}$ is strongly asymptotically unbiased for estimation of Θ_{ij}^0 and achieves the lower bound on variance.

We introduce further notation: let $\gamma_0 := \arg\min_{\gamma\in\mathbb{R}^{p-1}} \mathbb{E}\|X_j - X_{-j}\gamma\|_n^2$ and let $\tau_j^2 := \mathbb{E}\|X_j - X_{-j}\gamma_0\|_n^2$. To show strong asymptotic unbiasedness, we need a strong oracle inequality for the estimator $\hat{\Theta}$ of Θ_0. Namely, the paper [9] shows that under certain conditions (see Theorem 12) it holds $\|\hat{\Theta}_j - \Theta_j^0\|_1 = \mathcal{O}_P(s\lambda)$. We aim to show a stronger claim, $\mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1 = \mathcal{O}(s\lambda)$. This is a somewhat more difficult task than for the linear regression, since one has to make sure that the estimate of the noise level, $\hat{\tau}_j^2$, does not blow up in expectation. Before establishing strong asymptotic unbiasedness, we thus need the following auxiliary results.

Lemma 5. *Assume that $s \log p/n = o(1)$. Let $k \in \{1, 2, \dots\}$ be fixed and let $\hat{\gamma}_j$ be defined as in* (10) *with a sufficiently large tuning parameter λ of order $\sqrt{\log p/n}$. Then for all $\Theta_0 \in \mathcal{G}$ it holds that*

$$(\mathbb{E}_{\Theta_0}\|\hat{\gamma}_j - \gamma_j^0\|_1^k)^{1/k} = \mathcal{O}(s\lambda).$$

The following Lemma shows that the noise estimator $1/\hat{\tau}_j^2$ does not blow up in expectation.

Lemma 6. *Assume that $s \log p/n = o(1)$. Then for all $\Theta_0 \in \mathcal{G}$ the following statements hold*

1. $\mathbb{E}\frac{1}{\hat{\tau}_j^8} = \mathcal{O}(1)$,
2. $\mathbb{E}|\hat{\tau}_j^2 - \tau_j^2|^2 = \mathcal{O}(s\lambda^2)$,
3. $\mathbb{E}|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}| = \mathcal{O}(\sqrt{s}\lambda)$.

Combination of results in Lemmas 5 and 6 gives the following result.

Lemma 7. *Assume that $s \log p/n = o(1)$. Then for $\hat{\Theta}_j$ defined in* (11) *with a sufficiently large tuning parameter of order $\sqrt{\log p/n}$ it holds for all $\Theta_0 \in \mathcal{G}$ that*

$$\mathbb{E}_{\Theta_0}(\|\hat{\Theta}_j - \Theta_j^0\|_1^2)^{1/2} = \mathcal{O}(s\lambda).$$

Now we are at the point to prove strong asymptotic unbiasedness of $\hat{T}_{ij}$.

Lemma 8. *Assume that $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. Let $\hat{T}_{ij}$ be defined in* (12)*, where $\hat{\Theta}$ is the node-wise Lasso estimator. Then $\hat{T}_{ij}$ is strongly asymptotically unbiased, i.e. for all $\Theta_0 \in \mathcal{G}$ it holds*

$$\sqrt{n}\mathbb{E}_{\Theta_0}(\hat{T}_{ij} - \Theta_{ij}^0) = o(1).$$

Finally, we show that the de-sparsified estimator $\hat{T}_{ij}$ reaches the lower bound on the variance.

Theorem 9. *Assume that $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. Let $\hat{T}_{ij}$ be defined in* (12)*, where $\hat{\Theta}$ is the node-wise Lasso estimator. Then $\hat{T}_{ij}$ is a strongly asymptotically unbiased estimator of Θ_{ij}^0 and for any strongly asymptotically unbiased estimator T of Θ_{ij}^0 it holds for all $\Theta_0 \in \mathcal{G}$*

$$\text{var}_{\Theta_0}(T) \geq \text{var}_{\Theta_0}(\hat{T}_{ij}) = \frac{\Theta_{ii}^0\Theta_{jj}^0 + (\Theta_{ij}^0)^2 + o(1)}{n}.$$

9.1 Discussion on asymptotic efficiency for Gaussian graphical models

The conditions under which we show asymptotic efficiency only include eigenvalue conditions on the true precision matrix and sparsity conditions on columns/rows of the precision matrix. In particular, the condition on row sparsity required is the same as for the linear model: $s = o(\sqrt{n}/\log p)$. In view of results on minimax rates for estimation of elements of precision matrices (see [8]), the condition $s = o(\sqrt{n}/\log p)$ is necessary for construction of confidence intervals.

We remark that extension of the above results to sub-Gaussian observations is straightforward. We note that similarly as for the linear regression, one can consider estimators of linear functionals of the precision matrix. Then one may construct estimators of linear functionals of Θ_0 by taking linear combinations of the de-sparsified nodewise Lasso $\hat{T}$ to construct an asymptotically efficient estimator.

10 Lower bounds for the Gaussian sequence model

Finally, we consider the Gaussian sequence model, which might be viewed as a special case of a linear model with fixed design. However, the number of parameters here is precisely equal to the number of observations. Consider thus the model

$$X_i = \beta_i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ are independent. The following Theorem shows a lower bound on the variance of any strongly asymptotically unbiased estimator.

Theorem 10. *Let $a \in \mathbb{R}^p$ be such that $a^T a = 1$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta_0)$ at β^0 in the direction a with rate δ_n. Assume moreover that for some $\dot{g}(\beta^0) \in \mathbb{R}^p$ and for $m_n = n/\delta_n$*

$$\sqrt{m_n}\left(g(\beta^0 + a/\sqrt{m_n}) - g(\beta^0)\right) = a^T \dot{g}(\beta^0) + o(1).$$

Then

$$n\mathrm{var}(T) \geq [a^T \dot{g}(\beta_0)]^2 + o(1).$$

Similarly as in the linear model, the worst sub-direction is then

$$a = \dot{g}(\beta_0) / \sqrt{\dot{g}(\beta_0)^T \dot{g}(\beta_0)},$$

and the lower bound

$$n\mathrm{var}(T) \geq \dot{g}(\beta_0)^T \dot{g}(\beta_0) + o(1).$$

11 Le Cam's Lemma

In the analysis in the previous part, we have shown for several settings that the de-sparsified estimator is strongly asymptotically unbiased and reaches the lower bound on the variance. This guaranteed us that the de-sparsified estimator is the best among all strongly asymptotically unbiased estimators in terms of variance. In this part, we further show that the convergence of the de-sparsified estimator to the limiting normal distribution with smallest possible variance is locally uniform in the underlying unknown parameter. This is motivated by work of Le Cam on local asymptotic normality ([5]).

The motivation for locally uniform convergence can be seen in the classical examples of superefficiency (see e.g. [10]). They show that pointwise convergence is insufficient for asymptotic efficiency and that we in fact need uniform convergence on shrinking neighbourhoods. We show that for sparse high-dimensional models, asymptotic linearity of an estimator implies this uniform convergence. This is in line with the results of Le Cam (see Lemma 8.14 in [10]). We consider the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$, where

$$\Theta := \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq s, \|\theta\|_2 = \mathcal{O}(1)\}.$$

Note that for many sparse high-dimensional models, one can often show that the de-sparsified estimator T_n is asymptotically linear:

$$T_n - g(\theta) = \frac{1}{n}\sum_{i=1}^n l_\theta(X_i) + o_P(n^{-1/2}),$$

where $\mathbb{E}_\theta l_\theta = 0$ and $\mathbb{E} l_\theta^2 < \infty$. For asymptotically linear estimators, one has the asymptotic variance $V_\theta := \mathbb{E} l_\theta^2$. Now consider the following condition for every $h \in \Theta$

$$P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\beta) = 0.$$

If the condition is satisfied, then the Cauchy-Schwarz inequality implies

$$(h^T \dot{g}(\theta))^2 = (P_\theta l_\theta h^T s_\theta)^2 \le \mathrm{var}(l_\theta)\mathrm{var}(h^T s_\theta) = V_\theta h^T I_\theta h.$$

Hence

$$V_\theta \ge \max_{h\in\Theta} (h^T \dot{g}(\theta))^2 / h^T I_\theta h. \tag{13}$$

Assuming that $I_\theta^{-1}\dot{g}(\theta) \in \Theta$, the right-hand side of (13) is maximized at $I_\theta^{-1}\dot{g}(\theta)$. Hence we obtain the following lower bound on the asymptotic variance

$$V_\theta \ge \dot{g}(\theta)^T I_\theta^{-1} \dot{g}(\theta).$$

Thus for $V_\theta = \dot{g}(\theta)^T I_\theta^{-1}\dot{g}(\theta)$, the lower bound is reached. However, to avoid superefficiency as discussed above, we require uniform convergence. Under the conditions of the central limit theorem, asymptotic linearity implies that

$$\sqrt{n}(T_n - g(\theta))/V_\theta^{1/2} \overset{\theta}{\rightsquigarrow} \mathcal{N}(0,1)$$

for every θ. We show that asymptotic linearity actually implies uniform convergence: the limiting distribution remains the same under a disappearing change in the parameter. In particular, this means that for every $h \in \Theta$ and every $\theta \in \Theta$ it holds that

$$\frac{\sqrt{n}(T_n - g(\theta + h/\sqrt{n}))}{V_\theta^{1/2}} \overset{\theta + h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1).$$

This result is precisely formulated in the following Theorem.

Theorem 11. *Let $g : \mathbb{R}^p \to \mathbb{R}$ satisfy*

$$\sqrt{n}(g(\theta + h/\sqrt{n}) - g(\theta)) = h^T \dot{g}(\theta) + o(1).$$

Suppose that for all $\theta \in \Theta$

$$T_n - g(\theta) = \frac{1}{n}\sum_{i=1}^n l_\theta(X_i) + o_{P_\theta}(n^{-1/2}),$$

where $P_\theta l_\theta = 0$ and $V_\theta := P_\theta l_\theta^2 < \infty$. Suppose that $V_\theta = \mathcal{O}(1)$ and $1/V_\theta = \mathcal{O}(1)$. Let s_θ be the score function, let $I_\theta := \mathbb{E} s_\theta s_\theta^T$ and assume that

$$\|\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) + I_\theta\|_\infty = \mathcal{O}_P(\lambda),$$

where λ is such that $s\lambda = o(1)$. Assume further that $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$. Then for every $h \in \Theta$

it holds that

$$\frac{\sqrt{n}(T_n - g(\theta + h/\sqrt{n})) - (P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\theta))}{V_\theta^{1/2}} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1)$$

The result of Theorem 11 contains a bias term $P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\theta)$ which depends on h. However, in many cases, we have similarly as in the low-dimensional setting (when the number of parameters p is fixed) that $P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\theta) = 0$.

We look at this condition for the linear regression and Gaussian graphical models. First consider the linear model with random design and the parameter of interest $g(\beta) = \beta_j$. Then we have asymptotic linearity of the de-sparsified Lasso (see [9]) with $l_\beta(x_i, y_i) = (\Theta_j^0)^T x_i \epsilon_i$, where Θ_j^0 is the j-th column of the precision matrix. But then

$$P_\theta(l_\theta h^T s_\theta) = (\Theta_j^0)^T \mathbb{E} x_1 \epsilon_1^2 x_1^T h = (\Theta_j^0)^T \mathbb{E}\mathbb{E}(x_1 \epsilon_1^2 x_1^T | x_1) h = \Theta_j^0 \Sigma_0 h = h_j.$$

Therefore in this case indeed $P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\beta) = 0$. Next consider precision matrix estimation with the parameter of interest $g(\Theta) = \Theta_{ij}$. We have asymptotic linearity of the de-sparsified nodewise Lasso (see [4]) with

$$l_\Theta(x) = \text{tr}(\Theta_i \Theta_j^T (xx^T - \Sigma_0)) = \text{vec}(\Theta_i \Theta_j^T)^T \text{vec}(xx^T - \Sigma_0)$$

and $\text{vec}(H)^T \text{vec}(s_\Theta) = \text{vec}(H)^T \text{vec}(xx^T - \Sigma_0)$. Then by some algebra it follows

$$\begin{aligned} P_\Theta(l_\Theta \text{vec}(H)^T \text{vec}(s_\Theta)) &= \text{vec}(\Theta_i \Theta_j^T)^T \Sigma_0 \otimes \Sigma_0 \text{vec}(H) \\ &= \text{vec}(\Sigma_0 \Theta_i \Theta_j^T \Sigma_0)^T \text{vec}(H) \\ &= e_i^T H e_j = H_{ij}. \end{aligned}$$

Hence the condition is satisfied for Gaussian graphical models.

This shows that in the above cases, the bias term vanishes. Hence the de-sparsified estimator converges uniformly to a normal distribution with zero mean and the smallest possible variance.

12 Conclusions

In this paper we precisely formulated the concept of asymptotic efficiency in high-dimensional models. We further analyzed the lower bounds on asymptotic efficiency and whether it is possible to construct an estimator attaining the lower bounds. We showed that indeed construction of asymptotically efficient estimator is possible: a de-sparsified estimator in linear regression and Gaussian graphical models is asymptotically efficient. Our analysis identified the theoretical conditions on the model and on the parameter sparsity under which asymptotic efficiency is attained. The underlying analysis of the asymptotic Cramér-Rao bound involved a detailed study of the remainders.

Furthermore, we showed that asymptotic linearity of an estimator implies that the estimator converges uniformly to the limiting normal distribution with zero mean and smallest possible variance. Thus we extended the classical results of Le Cam (see [5]). In high-dimensional settings, the de-sparsified estimator is asymptotically linear in various settings such as the linear regression, Gaussian graphical models and some cases of generalized linear models (see [9]).

Our analysis considered particular examples of de-sparsified estimators, however, other estimators which are in some sense equivalent to these de-sparsified estimators (such as those based on the square-root Lasso) are applicable.

13 Proofs

We recall a sub-Gaussianity assumption on random vectors (see Section 14 in [1]).

Definition 4. *We say that a random vector $X \in \mathbb{R}^p$ is sub-Gaussian with a constant L if for all $\alpha \in \mathbb{R}^p$ such that $\|\alpha\|_2 = 1$ it holds that*

$$\mathbb{E}e^{(\alpha^T X)^2/L^2} = \mathcal{O}(1).$$

13.1 Proofs for Section 4

In this section we prove the oracle inequality for the Lasso as stated in Theorem 1. We need the following preliminary Lemmas 9, 10 and 11. Lemma 9 below is a version of Theorem 1 in [7]. It gives sufficient conditions under which the restricted eigenvalue condition is satisfied. For the definition of the restricted eigenvalue condition and the compatibility condition, see Section 2.2.2 in [7] and Section 6.13 in [1]. Lemma 10 is a concentration result which follows from more general results for sub-Gaussian random variables in Section 14 in [1]. Lemma 11 is a version of Lemma 5.1 in [9].

Lemma 9. *Let $X \in \mathbb{R}^{n\times p}$, where the rows $X_i \in \mathbb{R}^p, i = 1, \dots, n$ are $\mathcal{N}(0, \Sigma_0)$-distributed. Suppose that $\Sigma_0 := \mathbb{E}X^T X/n$ satisfies the compatibility condition with $\phi > 0$, where $1/\phi = \mathcal{O}(1)$ and that $\|\Sigma_0\|_\infty = \mathcal{O}(1)$. Further suppose that $s \log p/n = o(1)$. Then there exists a universal constant ϕ' such that $1/\phi' = \mathcal{O}(1)$ and such that for any fixed $\tau > 0$ it holds for all n sufficiently large that*

$$P(X \text{satisfies the compatibility condition with } \phi') \geq 1 - p^{-\tau}.$$

Lemma 10. *Suppose that $\epsilon \sim \mathcal{N}_n(0, \sigma_\epsilon^2 I)$ and $X_i, i = 1, \dots, n$ are independent and $\mathcal{N}(0, \Sigma_0)$-distributed. Then for any $\tau > 0$*

$$P(\|\epsilon^T X\|_\infty/n > \tau c_1 \sqrt{\log p/n}) \leq c_2 p^{-\tau},$$

where c_2, c_2 are some universal constants.

Lemma 11. *Assume the linear model in (1) with Gaussian error with variance $\sigma_\epsilon^2 = \mathcal{O}(1)$ and suppose that $X_i, i = 1, \dots, n$ are independent and $\mathcal{N}(0, \Sigma_0)$-distributed. Consider the Lasso estimator $\hat{\beta}$ defined in (2) with tuning parameter $\lambda \geq 2\lambda_0$. Then on the set*

$$\mathcal{T} := \{\|\epsilon^T X\|_\infty/n \leq \lambda_0, X \text{satisfies the compatibility condition with } \phi\}$$

it holds

$$\|\hat{\beta} - \beta_0\|_1 \leq 8\lambda \frac{s}{\phi^2}.$$

Proof of Theorem 1. First we summarize the oracle inequality for the Lasso which holds with high probability. By Lemma 10, for the complement of the set $\mathcal{T}_1 := \{\|\epsilon^T X\|_\infty/n >$

$\tau c_1\sqrt{\log p/n}\}$ it holds that $P(\mathcal{T}_1^c) \leq c_2 p^{-\tau}$ for any $\tau > 0$. The condition $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$ implies that Σ_0 satisfies the compatibility condition with some constant ϕ such that $1/\phi = \mathcal{O}(1)$. Furthermore, by assumption, we have that $s\log p/n = o(1)$ and $\|\Sigma_0\|_\infty = \mathcal{O}(1)$. Define the event $\mathcal{T}_2 := \{X$ satisfies the compatibility condition with $\tilde{\phi}\}$. Then by Lemma 9 it follows that $P(\mathcal{T}_2^c) \leq p^{-\tau}$ (for all n sufficiently large) for some constant $\tilde{\phi}$ such that $1/\tilde{\phi} = \mathcal{O}(1)$. Denote $\mathcal{T} := \mathcal{T}_1 \cap \mathcal{T}_2$. Then $P(\mathcal{T}^c) \lesssim p^{-\tau}$. By Lemma 11, when $\lambda \geq 2\lambda_0 := 2\tau c_1\sqrt{\log p/n}$, on the set $\mathcal{T}$ it holds that $\|\hat{\beta} - \beta_0\|_1 \leq 8\lambda\frac{s}{\tilde{\phi}^2}$.

We now proceed to show that the oracle inequality for the Lasso holds also in expectation. The definition of $\hat{\beta}$ gives

$$\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \|\epsilon\|_n^2 + \lambda\|\beta_0\|_1.$$

Consequently,

$$\|\hat{\beta}\|_1 \leq \|\epsilon\|_n^2/\lambda + \|\beta_0\|_1.$$

Then, and by the triangle inequality

$$\|\hat{\beta} - \beta_0\|_1 \leq \|\hat{\beta}\|_1 + \|\beta_0\|_1 \leq \|\epsilon\|_n^2/\lambda + 2\|\beta_0\|_1,$$

and thus for any $k \in \{1, 2, \dots\}$

$$\mathbb{E}\|\hat{\beta} - \beta_0\|_1^k \leq \mathbb{E}(\|\epsilon\|_n^2/\lambda + 2\|\beta_0\|_1)^k.$$

Then it follows

$$\begin{aligned}\mathbb{E}(\|\epsilon\|_n^2/\lambda + 2\|\beta_0\|_1)^k &= \mathbb{E}\sum_{j=0}^k \binom{k}{j}(\|\epsilon\|_n^2)^j(2\|\beta_0\|_1)^{k-j} \\ &= \sum_{j=0}^k \binom{k}{j}(2\|\beta_0\|_1)^{k-j}\mathbb{E}(\|\epsilon\|_n^2)^j.\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}(\|\epsilon\|_n^2)^j &= \frac{1}{n^j}\sum_{i_1=1}^j \cdots \sum_{i_j=1}^j \mathbb{E}\epsilon_{i_1}^2 \dots \epsilon_{i_j}^2 \\ &\leq \max_{i_1,\dots,i_j} \mathbb{E}\epsilon_{i_1}^2 \dots \epsilon_{i_j}^2.\end{aligned}$$

By assumption we have $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and hence by the well-known formula: $\mathbb{E}\epsilon_i^m = \sigma_\epsilon^m(m-1)!$ if m is even. Furthermore, we assume that $\sigma_\epsilon = \mathcal{O}(1)$. Hence, and by the Cauchy-Schwarz inequality, we can conclude that

$$\max_{i_1,\dots,i_j} \mathbb{E}\epsilon_{i_1}^2 \dots \epsilon_{i_j}^2 = \mathcal{O}(1),$$

because $j \leq k$, where k is fixed. Next observe that by assumption we have $\|\beta_0\|_2 = \mathcal{O}(1)$ and hence

$$\|\beta_0\|_1^k \leq (\sqrt{s}\|\beta_0\|_2)^k \leq \mathcal{O}(s^{k/2}).$$

We can thus conclude that

$$\begin{aligned}\mathbb{E}(\|\epsilon\|_n^2/\lambda + 2\|\beta_0\|_1)^k &= \sum_{j=0}^{k}\binom{k}{j}(2\|\beta_0\|_1)^{k-j}\mathbb{E}(\|\epsilon\|_n^2)^j \\ &\leq \sum_{j=0}^{k}\binom{k}{j}2^{k-j}s^{(k-j)/2}\mathcal{O}(1) \\ &\leq \mathcal{O}(s^{k/2}).\end{aligned}$$

Hence

$$(\mathbb{E}_{\beta_0}\|\hat{\beta}-\beta_0\|_1^k)^{1/k} = \mathcal{O}(s^{1/2}).$$

On the set $\mathcal{T}$ we have $\|\hat{\beta}-\beta_0\|_1 = \mathcal{O}(s\lambda)$ and thus $\|\hat{\beta}-\beta_0\|_1^k = \mathcal{O}(s^k\lambda^k)$, and otherwise (so also on the set $\mathcal{T}^c$) we have the rough bound $\mathbb{E}_{\beta_0}\|\hat{\beta}-\beta_0\|_1^k = \mathcal{O}(s^{k/2})$. Denote by 1_A the indicator function of a set A. Then it follows using the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}_{\beta_0}\|\hat{\beta}-\beta_0\|_1^k &= \mathbb{E}\|\hat{\beta}-\beta_0\|_1^k 1_{\mathcal{T}} + \mathbb{E}\|\hat{\beta}-\beta_0\|_1^k 1_{\mathcal{T}^c} \\ &\leq \mathcal{O}(s^k\lambda^k) + \sqrt{\mathbb{E}\|\hat{\beta}-\beta_0\|_1^{2k}}\sqrt{\mathbb{E}1_{\mathcal{T}^c}} \\ &= \mathcal{O}(s^k\lambda^k) + \sqrt{s^k}\ \sqrt{\mathbb{P}(\mathcal{T}^c)} \\ &\lesssim \mathcal{O}(s^k\lambda^k) + s^{k/2}p^{-\tau/2} \\ &= \mathcal{O}(s^k\lambda^k),\end{aligned}$$

where we used the assumption $p^{-\tau/2} = \mathcal{O}((s\lambda^2)^{k/2})$ which implies that $s^{k/2}p^{-\tau/2} = \mathcal{O}(s^k\lambda^k)$. Hence we conclude that

$$(\mathbb{E}_{\beta_0}\|\hat{\beta}-\beta_0\|_1^k)^{1/k} = \mathcal{O}(s\lambda). \tag{14}$$

The second statement of the Theorem follows by Markov's inequality. Take any $\nu > 0$. Then

$$P(\|\hat{\beta}-\beta_0\|_1 > \nu C_1 s\lambda) = P(\|\hat{\beta}-\beta_0\|_1^k > \nu^k {C_1}^k s^k\lambda^k) \leq \frac{\mathbb{E}\|\hat{\beta}-\beta_0\|_1^k}{C_1^k\nu^k s^k\lambda^k} \leq \frac{1}{\nu^k},$$

where C_1 is the constant from (14). □

13.2 Proofs for Section 6.1

Before proving the statement of Theorem 2, we need the following two auxiliary Lemmas.

Lemma 12. *Let $Z \sim \mathcal{N}(0,1)$. Then for all $t \in \mathbb{R}$*

$$\mathbb{E}\left[e^{tZ-t^2/2} - 1 - tZ\right]^2 = e^{t^2} - 1 - t^2.$$

Moreover, for $2t^2 < 1$ we have

$$\mathbb{E}e^{t^2Z^2} = \frac{1}{\sqrt{1-2t^2}}.$$

Proof of Lemma 12. By direct calculation

$$\mathbb{E}\left[e^{tZ-t^2/2}\right]^2 = \mathbb{E}[2tZ - t^2] = e^{t^2},$$

$$\mathbb{E}[tZ - t^2/2] = 1$$

and

$$\mathbb{E}Ze^{tZ-t^2/2} = t\mathbb{E}[tZ - t^2/2] = t.$$

The first result follows immediately. The second result is also easily found by standard calculations:

$$\mathbb{E}e^{t^2Z^2} = \int e^{t^2z^2}\phi(z)dz = \int \phi(z\sqrt{1-2t^2}) = \frac{1}{\sqrt{1-2t^2}}.$$

□

Lemma 13. *Suppose that $2h^T\Sigma_0 h < 1$. Let $Z = (X, Y)$, denote the corresponding probability measure by ν and the density by p_β. Then it holds*

$$\mathbb{E}_{\beta_0}\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - s_{\beta_0}(Z)^T a/\sqrt{m}\right)^2 = (1 - 2h^T\Sigma_0 h)^{-n/2} - 1 - nh^T\Sigma_0 h.$$

Proof of Lemma 13. Denote the density of Y given X by $p_{\beta_0}(\cdot|X)$, i.e.

$$p_{\beta_0}(y|X) := \prod_{i=1}^n \phi(y_i - x_i^T\beta_0), \quad y = (y_1, \ldots, y_n),$$

where ϕ is the standard normal density.
Given X, the random variable $\epsilon^T Xh$ is $\mathcal{N}(0, nh^T\hat{\Sigma}h)$-distributed. It follows therefore from the first result of Lemma 12 that

$$\mathbb{E}_{\beta_0}\left(\frac{p_{\beta_0+h}(Y - Xh|X) - p_{\beta_0}(Y|X)}{p_{\beta_0}(Y|X)} - s_{\beta_0}(Z)^T h\right)^2 = \mathbb{E}e^{nh^T\hat{\Sigma}h} - 1 - nh^T\Sigma_0 h.$$

Since $X_i h \sim \mathcal{N}(0, h^T\Sigma_0 h)$ for $i = 1, \ldots, n$, we have by the second result of Lemma 12

$$\mathbb{E}e^{(X_i h)^2} = \frac{1}{\sqrt{1 - 2h^T\Sigma_0 h}}.$$

Whence the result. □

Proof of Theorem 2. By assumption (3) and by strong asymptotic unbiasedness, it follows

$$\begin{aligned} a^T\dot{g}(\beta_0) &= \sqrt{m_n}\,(g(\beta_0 + a/\sqrt{m_n}) - g(\beta_0)) + o(1) \\ &= \sqrt{m_n}\left(\mathbb{E}_{\beta_0+a/\sqrt{m_n}}T_n - \mathbb{E}_{\beta_0}T_n\right) + o(1) \end{aligned}$$

Denoting $Z = (X, Y)$, the corresponding probability measure by ν and the density by p_β, we may further rewrite the expressions to obtain

$$\begin{aligned}
\sqrt{m_n}\left(\mathbb{E}_{\beta_0+a/\sqrt{m_n}}T_n - \mathbb{E}_{\beta_0}T_n\right) + o(1) &= \int T_n(z)(p_{\beta_0+a/\sqrt{m_n}}(z) - p_{\beta_0}(z))d\nu(z) \\
&= \mathbb{E}_{\beta_0}T_n(Z)\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} \\
&= \mathbb{E}_{\beta_0}T_n(Z)\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - \frac{s_{\beta_0}(Z)^T a}{\sqrt{m_n}}\right) \\
&\quad + \ \mathbb{E}_{\beta_0}T_n(Z)\frac{s_{\beta_0}(Z)^T a}{\sqrt{m_n}} \\
&= \mathbb{E}_{\beta_0}(T_n(Z) - g(\beta_0))\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - \frac{s_{\beta_0}(Z)^T a}{\sqrt{m_n}}\right) \\
&\quad + \mathbb{E}_{\beta_0}T_n(Z)\frac{s_{\beta_0}(Z)^T a}{\sqrt{m_n}}
\end{aligned}$$

We assume the variance of T is $\mathcal{O}(1)$, otherwise the statement trivially holds. But then

$$\mathbb{E}_{\beta_0}(T_n(Z) - g(\beta_0))^2 = \text{var}(T_n(Z)) + [\mathbb{E}_{\beta_0}(T_n(Z) - g(\beta_0))]^2 = \mathcal{O}(1) + o(1/n) = \mathcal{O}(1).$$

By Lemma 13 and some basic calculations,

$$\mathbb{E}_{\beta_0}\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - s_{\beta_0}(Z)^T a/\sqrt{m}\right)^2 = \mathcal{O}(\delta_n).$$

Consequently, and by the Cauchy-Schwarz inequality, we have the upper bound

$$\begin{aligned}
&\left|\mathbb{E}_{\beta_0}(T_n(Z) - g(\beta_0))\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - s_{\beta_0}(Z)^T a/\sqrt{m}\right)\right| \\
\leq\ &\sqrt{\mathbb{E}_{\beta_0}(T_n(Z) - g(\beta_0))^2}\sqrt{\mathbb{E}\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - s_{\beta_0}(Z)^T a/\sqrt{m}\right)^2} = \mathcal{O}(\delta_n).
\end{aligned}$$

Hence, and since $\delta_n \downarrow 0$

$$\begin{aligned}
a^T \dot{g}(\beta_0) &= \text{cov}(T_n, \epsilon^T X a/\sqrt{m}) + o(1) \\
&\leq \sqrt{n}\sqrt{\text{var}(T_n)} + o(1).
\end{aligned}$$

□

13.3 Proofs for Section 6.2

Proof of Theorem 3. The proof follows the same lines as the proof of Theorem 2. The only difference is that we need to check the condition

$$\mathbb{E}_{\beta_0}\left(\frac{p_{\beta_0+a/\sqrt{m_n}}(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} - s_{\beta_0}(Z)^T a/\sqrt{m}\right)^2 = \mathcal{O}(\delta_n),$$

for fixed design. To this end, consider the density $p_\beta(z)$, which is given by

$$p_\beta(z_i) := p_\beta(x_i, y_i) = \prod_{i=1}^n \phi(y_i - x_i^T \beta) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(Y-X\beta)^T(Y-X\beta)}.$$

For simplicity, denote $h := a/\sqrt{m_n}$. By direct calculation, we obtain

$$\begin{aligned}\frac{p_\beta(Z) - p_{\beta_0}(Z)}{p_{\beta_0}(Z)} &= \frac{e^{-\frac{1}{2}(Y-X(\beta+h))^T(Y-X(\beta+h))} - e^{-\frac{1}{2}(Y-X\beta)^T(Y-X\beta)}}{e^{-\frac{1}{2}(Y-X\beta)^T(Y-X\beta)}} \\ &= e^{-h^T X^T(Y-X\beta) - \frac{1}{2}h^T X^T X h} - 1\end{aligned}$$

We have $h^T X^T (Y - X\beta) = h^T X^T \epsilon \sim \mathcal{N}(0, h^T X^T X h)$. Hence $\mathbb{E} e^{t\epsilon^T X h} = e^{\frac{1}{2}t^2 h^T X^T X h}$. Thus we obtain

$$\begin{aligned}\mathbb{E}\left(e^{-\epsilon^T X h + \frac{1}{2}h^T X^T X h} - 1 - \epsilon^T X h\right)^2 &= e^{h^T X^T X h} - 1 - h^T X^T X h \\ &= \mathcal{O}(h^T X^T X h) = \mathcal{O}(n h^T \hat{\Sigma} h).\end{aligned}$$

Then by the assumption $a^T \hat{\Sigma} a \leq 1 + o(1)$ we obtain

$$\mathcal{O}(n h^T \hat{\Sigma} h) = \mathcal{O}(n a^T \hat{\Sigma} a / m_n) = \mathcal{O}(a^T \hat{\Sigma} a \delta_n) \leq \mathcal{O}(\delta_n).$$

□

Proof of Lemma 1. By the assumptions $\|\hat{\Theta}\hat{\Sigma} - I\|_\infty = \mathcal{O}(\lambda)$, $\|\hat{\Theta}\dot{g}(\beta_0)\|_1 = \mathcal{O}(\sqrt{s}\|\hat{\Theta}\dot{g}(\beta_0)\|_2) = \mathcal{O}(\sqrt{s})$ and $\|\dot{g}(\beta_0)\|_1 = \mathcal{O}(\sqrt{s})$ we obtain

$$\begin{aligned}\dot{g}(\beta_0)^T \hat{\Theta}\hat{\Sigma}\hat{\Theta}\dot{g}(\beta_0)/\dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0) &\leq \dot{g}(\beta_0)^T(\hat{\Theta}\hat{\Sigma} - I)\hat{\Theta}\dot{g}(\beta_0)/\dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0) \\ &\quad + \dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0)/\dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0) \\ &\leq \|\dot{g}(\beta_0)\|_1 \|\hat{\Theta}\hat{\Sigma} - I\|_\infty \|\hat{\Theta}\dot{g}(\beta_0)\|_1 / \dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0) \\ &\quad + 1 \\ &\leq \mathcal{O}(1)s\lambda/\sqrt{\dot{g}(\beta_0)^T\hat{\Theta}\dot{g}(\beta_0)} + 1 = 1 + o(1).\end{aligned}$$

□

13.4 Proofs for Section 7.1

Proof of Lemma 2. We apply Theorem 1. Conditions $\|\beta_0\|_0 \leq s, \|\beta_0\|_2 = \mathcal{O}(1)$ and (A1) imply that conditions of Theorem 1 are satisfied.

□

Proof of Lemma 3. First note that

$$\mathbb{E}\hat{\Theta}_j^T X^T \epsilon / n = \mathbb{E}\mathbb{E}(\hat{\Theta}_j^T X^T \epsilon / n | X) = \mathbb{E}\hat{\Theta}_j^T X^T \mathbb{E}(\epsilon | X)/n = 0.$$

We then have by the definition of $\hat{b}_j$ and Hölder's inequality

$$\begin{aligned}\mathbb{E}_{\beta_0}(\hat{b}_j - \beta_j^0) &= \underbrace{\mathbb{E}_{\beta_0}\hat{\Theta}_j^T X^T \epsilon/n}_{=0} + \mathbb{E}_{\beta_0}(\hat{\Sigma}\hat{\Theta}_j - e_j)^T(\hat{\beta} - \beta_0) \\ &\leq \mathbb{E}_{\beta_0}\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \|\hat{\beta} - \beta_0\|_1.\end{aligned}$$

Hence, by assumption $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda$ and using Lemma 2, we obtain that $\mathbb{E}_{\beta_0}(\hat{b}_j - \beta_j^0) = \mathcal{O}(s\lambda^2) = o(1/\sqrt{n})$, where we used the sparsity assumption.

□

Proof of Theorem 4. By taking $g(\beta) = \beta_j$ and $a := \Theta_j^0/\sqrt{\Theta_{jj}^0}$ in Theorem 2 we obtain the lower bound.

Since $\hat{\Theta}_j$ is constructed using the nodewise regression, it satisfies

$$\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda.$$

Hence Lemma 3 implies that $\hat{b}_j$ is strongly asymptotically unbiased.

Note now that $\mathbb{E}(\Theta_j^0)^T X^T \epsilon/n = 0$ and

$$\text{var}((\Theta_j^0)^T X^T \epsilon/n) = \mathbb{E}(\mathbb{E}[((\Theta_j^0)^T X^T \epsilon/n)^2 | X]) = \mathbb{E}(\Theta_j^0)^T X^T X/n \Theta_j^0) = \Theta_{jj}^0.$$

We then have the following decomposition

$$\hat{b}_j - \beta_j^0 = (\Theta_j^0)^T X^T \epsilon/n + (\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n + (\hat{\Sigma}\hat{\Theta}_j - e_j)^T(\hat{\beta} - \beta_0).$$

Thus, and by the Cauchy-Schwarz inequality and some basic calculations, we have

$$\text{var}(\hat{b}_j - \beta_j^0) = \underbrace{\text{var}((\Theta_j^0)^T X^T \epsilon/n)}_{=\Theta_{jj}^0/n} + \mathcal{O}(\underbrace{\text{var}((\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n)}_{i}) + \mathcal{O}(\underbrace{\text{var}((\hat{\Sigma}\hat{\Theta}_j - e_j)^T(\hat{\beta} - \beta_0))}_{ii}).$$

We show that the terms i and ii are of small order $1/n$.

First consider the term i. By Lemma 5, which is proved in the Section 8.2 on Gaussian graphical models, we have that $\mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1^4 = \mathcal{O}(s^4\lambda^4)$. Hence

$$\begin{aligned} i = \text{var}((\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n) &= \mathbb{E}((\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n)^2 - (\mathbb{E}((\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n))^2 \\ &\leq \mathbb{E}((\hat{\Theta}_j - \Theta_j^0)^T X^T \epsilon/n)^2 \\ &\leq \mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1^2 \|X^T\epsilon\|_\infty^2/n^2 \\ &\leq (\mathbb{E}\|X^T\epsilon\|_\infty^4/n^4)^{1/2} \\ &\leq s^2\lambda^4 = o(1/n), \end{aligned}$$

where we used that $\mathbb{E}\|X^T\epsilon\|_\infty^4/n^4 = \mathcal{O}(\lambda^2)$, which follows by concentration results for sub-Gaussian random variables (see [1]).

For the second term ii, by Lemma 2 it follows

$$\mathbb{E}[(\hat{\Sigma}\hat{\Theta}_j - e_j)(\hat{\beta} - \beta_0)]^2 = \mathcal{O}((s\lambda^2)^2).$$

Hence

$$\begin{aligned} ii &= \mathrm{var}((\hat{\Sigma}\hat{\Theta}_j - e_j)(\hat{\beta} - \beta_0)) \\ &= \mathbb{E}((\hat{\Sigma}\hat{\Theta}_j - e_j)(\hat{\beta} - \beta_0))^2 - (\mathbb{E}(\hat{\Sigma}\hat{\Theta}_j - e_j)(\hat{\beta} - \beta_0))^2 \\ &= \mathcal{O}((s\lambda^2)^2) = o(1/n). \end{aligned}$$

Thus we obtain

$$\mathrm{var}(\hat{b}_j) = \frac{\Theta^0_{jj} + o(1)}{n}.$$

□

13.5 Proofs for Section 7.2

Proof of Theorem 6. The lower bound follows by Theorem 1 (note that $g(\beta) = \beta_j$ and thus the condition on g is satisfied) since by assumption, $\|\hat{\Theta}_j\|_0 = \mathcal{O}(s), \|\hat{\Theta}_j\|_2 = \mathcal{O}(1)$.
Strong asymptotic unbiasedness of $\hat{b}_j$ follows similarly as in Lemma 3 under the assumptions $\|\beta_0\|_0 \leq s, \|\beta_0\|_2 = \mathcal{O}(1)$, $s = o\left(\frac{\sqrt{n}}{\log p}\right)$ and if X^TX/n satisfies the compatibility condition with a universal constant.
First observe that by the assumption on $\hat{\Theta}_j$ we obtain

$$\hat{\Theta}_j^T\hat{\Sigma}\hat{\Theta}_j/\hat{\Theta}_{jj} \leq \|\hat{\Theta}_j^T\|_1\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty/\hat{\Theta}_{jj} + \hat{\Theta}_j^T e_j/\hat{\Theta}_{jj} \leq \mathcal{O}(s\lambda) + 1 = 1 + o(1).$$

Hence for the variance of

$$\hat{b}_j - \beta_j^0 = (\hat{\Theta}_j)^T X^T\epsilon/n + (\hat{\Sigma}\hat{\Theta}_j - e_j)^T(\hat{\beta} - \beta_0),$$

we get

$$\begin{aligned} \mathrm{var}(\hat{b}_j|X) &= \hat{\Theta}_j^T\hat{\Sigma}\hat{\Theta}_j/(\hat{\Theta}_{jj}n) + \mathcal{O}(\|\hat{\Sigma}\hat{\Theta}_j - e_j\|^2_\infty\|\hat{\beta} - \beta_0\|_1^2) \\ &= \hat{\Theta}_j^T\hat{\Sigma}\hat{\Theta}_j/(\hat{\Theta}_{jj}n) + o(1/n) = \hat{\Theta}_{jj}/n + o(1/n). \end{aligned}$$

□

13.6 Proofs for Section 7.1

Proof of Lemma 4. We have

$$\begin{aligned} \mathbb{E}_{\beta_0}(\hat{b}_\xi - \xi^T\beta_0) &= \underbrace{\mathbb{E}_\beta\xi^T\hat{\Theta}X^T\epsilon/n}_{=0} + \mathbb{E}_{\beta_0}\xi^T(\hat{\Sigma}\hat{\Theta} - I)^T(\hat{\beta} - \beta_0) \\ &\leq \mathbb{E}_{\beta_0}\|\xi\|_1\|\hat{\Sigma}\hat{\Theta} - I\|_\infty\|\hat{\beta} - \beta_0\|_1 \\ &= \mathcal{O}(s\lambda^2) = o(1/\sqrt{n}). \end{aligned}$$

□

Proof of Theorem 5. By Corollary 2 with $g(\beta) = \xi^T\beta$ we obtain the lower bound by assumptions $\|\Theta_0\xi\|_0 \leq s$, $\|\xi\|_1 = \mathcal{O}(1)$.
Lemma 4 implies that $\hat{b}_\xi$ is strongly asymptotically unbiased.

It remains to calculate the variance of $\hat{b}_\xi$. Consider the following decomposition

$$\hat{b}_\xi - \xi^T\beta_0 = \xi^T\Theta_0^T X^T\epsilon/n + \xi^T(\hat{\Theta}-\Theta_0)^T X^T\epsilon/n + \xi^T(\hat{\Sigma}\hat{\Theta}-I)^T(\hat{\beta}-\beta_0).$$

Then one can show using basic calculations and the Cauchy-Schwarz inequality that

$$\text{var}(\hat{b}_\xi) \;\;=\;\; \text{var}(\xi^T\Theta X^T\epsilon/n) + \mathcal{O}(\underbrace{\text{var}(\xi^T(\hat{\Theta}-\Theta_0)^T X^T\epsilon/n)}_{i}) + \mathcal{O}(\underbrace{\text{var}(\xi^T(\hat{\Sigma}\hat{\Theta}-I)^T(\hat{\beta}-\beta_0))}_{ii}).$$

We have $\text{var}(X) \le \mathbb{E}X^2$ and hence

$$\begin{aligned} i &= \text{var}(\xi^T(\hat{\Theta}-\Theta_0)^T X^T\epsilon/n) \\ &\le \mathbb{E}(\xi^T(\hat{\Theta}-\Theta_0)^T X^T\epsilon/n)^2 \\ &\le \|\xi\|_1^2\mathbb{E}\left|\!\left|\!\left|\hat{\Theta}-\Theta\right|\!\right|\!\right|_1^2\|X^T\epsilon/n\|_\infty^2 = \mathcal{O}(s^2\lambda^4). \end{aligned}$$

and

$$ii = \mathbb{E}(\xi^T(\hat{\Sigma}\hat{\Theta}-I)^T(\hat{\beta}-\beta_0))^2 \le \|\xi\|_1^2\mathbb{E}\|\hat{\Sigma}\hat{\Theta}-I\|_\infty^2\|\hat{\beta}-\beta_0\|_1^2 = \mathcal{O}(s^2\lambda^4).$$

Thus we conclude $\text{var}(\hat{b}_\xi) = \frac{\xi^T\Theta_0\xi + o(1)}{n}$.

□

13.7 Proofs for Section 8.2

We first need several auxiliary Lemmas (Lemmas 14, 15, 16, 17). Although we state results only for estimation of linear functions, we carry out the calculations below to allow for handling general functionals.

Lemma 14. *Let $x \sim \mathcal{N}(0_p, \Sigma_0)$ and let $\Theta_0 = \Sigma_0^{-1}$. Then for any symmetric $A \in \mathbb{R}^{p\times p}$ it holds*

$$\mathbb{E}e^{tx^TAx} = \left(\frac{\det(\Theta_0)}{\det(\Theta_0 - 2tA)}\right)^{1/2}.$$

Proof of Lemma 14. By direct calculation, we obtain

$$\begin{aligned} \mathbb{E}e^{tx^TAx} &= \int_{\mathbb{R}^p} \frac{\det(\Theta_0)^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}x^T\Theta_0 x} e^{tx^TAx}dx \\ &= \int_{\mathbb{R}^p} \frac{\det(\Theta_0)^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}x^T\Theta_0 x} e^{tx^TAx}dx \\ &= \int_{\mathbb{R}^p} \frac{\det(\Theta_0)^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}x^T(\Theta_0-2tA)x}dx \\ &= \int_{\mathbb{R}^p} \frac{\det(\Theta_0)^{1/2}\det(\Theta_0-2tA)^{1/2}}{(2\pi)^{p/2}\det(\Theta_0-2tA)^{1/2}} e^{-\frac{1}{2}x^T(\Theta_0-2tA)x}dx \\ &= \frac{\det(\Theta_0)^{1/2}}{\det(\Theta_0-2tA)^{1/2}}. \end{aligned}$$

□

Lemma 15.

$$\begin{aligned}\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1-n\mathrm{tr}((\hat{\Sigma}-\Sigma_0)A)\right)^2 &= \frac{\det(\Theta_0+A)^n}{\det(\Theta_0)^n}\left(\frac{\det(\Theta_0)}{\det(\Theta_0+2A)}\right)^{n/2}-1\\ &+\sum_{i=1}^n \mathrm{var}(x_i^T A x_i)\\ &-2n\mathrm{tr}[((\Theta_0+A)^{-1}-\Sigma_0)A].\end{aligned}$$

Proof of Lemma 15. The density is given by

$$p_{\Theta_0}(x_1,\dots,x_n)=\frac{\det(\Theta_0)^{n/2}}{(2\pi)^{np/2}}e^{-\frac{1}{2}\sum_{i=1}^n x_i^T\Theta_0 x_i}$$

Then we have

$$\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1=\frac{\det(\Theta_0+A)^{n/2}e^{-\frac{1}{2}\sum_{i=1}^n x_i^T A x_i}}{\det(\Theta_0)^{n/2}}-1$$

The score function is given by $s_{\Theta_0}(x)=n(\hat{\Sigma}-\Sigma_0)$. Let

$$Z:=\mathrm{vec}(A)^T\mathrm{vec}(s_{\Theta_0}(x))=\mathrm{tr}(\sum_{i=1}^n x_i x_i^T A-n\Sigma_0 A)=\sum_{i=1}^n x_i^T A x_i - n\mathrm{tr}(\Sigma_0 A).$$

First observe that

$$\mathbb{E}Z^2=\mathrm{var}(\sum_{i=1}^n x_i^T A x_i)=\sum_{i=1}^n \mathrm{var}(x_i^T A x_i).$$

We have

$$\begin{aligned}\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1-Z\right)^2 &= \mathbb{E}\left(\frac{\det(\Theta_0+A)^{n/2}e^{-\frac{1}{2}\sum_{i=1}^n x_i^T A x_i}}{\det(\Theta_0)^{n/2}}-1-Z\right)^2\\ &= \frac{\det(\Theta_0+A)^n}{\det(\Theta_0)^n}\mathbb{E}e^{-\sum_{i=1}^n x_i^T A x_i}+1+\mathbb{E}Z^2+\\ &\quad -2\frac{\det(\Theta_0+A)^{n/2}}{\det(\Theta_0)^{n/2}}\mathbb{E}e^{-\frac{1}{2}\sum_{i=1}^n x_i^T A x_i}+2\mathbb{E}Z\\ &\quad -2\frac{\det(\Theta_0+A)^{n/2}}{\det(\Theta_0)^{n/2}}\mathbb{E}Ze^{-\frac{1}{2}\sum_{i=1}^n x_i^T A x_i}.\end{aligned}$$

Using Lemma 14, we obtain

$$\begin{aligned}\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1-Z\right)^2 &= \frac{\det(\Theta_0+A)^n}{\det(\Theta_0)^n}\left(\frac{\det(\Theta_0)}{\det(\Theta_0+2A)}\right)^{n/2}+1+\sum_{i=1}^n \mathrm{var}(x_i^TAx_i)\\ &\quad -2\frac{\det(\Theta_0+A)^{n/2}}{\det(\Theta_0)^{n/2}}\left(\frac{\det(\Theta_0)}{\det(\Theta_0+A)}\right)^{n/2}\\ &\quad -2\frac{\det(\Theta_0+A)^{n/2}}{\det(\Theta_0)^{n/2}}\mathbb{E}Ze^{-\frac{1}{2}\sum_{i=1}^n x_i^TAx_i}\\ &= \frac{\det(\Theta_0+A)^n}{\det(\Theta_0)^n}\left(\frac{\det(\Theta_0)}{\det(\Theta_0+2A)}\right)^{n/2}-1+\sum_{i=1}^n \mathrm{var}(x_i^TAx_i)\\ &\quad -2\,\underbrace{\frac{\det(\Theta_0+A)^{n/2}}{\det(\Theta_0)^{n/2}}\mathbb{E}Ze^{-\frac{1}{2}\sum_{i=1}^n x_i^TAx_i}}_{i}.\end{aligned}$$

Next we calculate i. We have

$$\mathbb{E}e^{tZ}=e^{-n\mathrm{tr}(\Sigma_0A)t}\mathbb{E}e^{t\sum_{i=1}^n x_i^TAx_i}=e^{-n\mathrm{tr}(\Sigma_0A)t}\left(\frac{\det(\Theta_0)}{\det(\Theta_0-2tA)}\right)^{n/2}.$$

We also have

$$\begin{aligned}\mathbb{E}Ze^{tZ} &= (\mathbb{E}e^{tZ})'\\ &= e^{-n\mathrm{tr}(\Sigma_0A)t}\left(\frac{\det(\Theta_0)}{\det(\Theta_0-2tA)}\right)^{n/2} n\left[\mathrm{tr}((\Theta_0-2tA)^{-1}A)-\mathrm{tr}(\Sigma_0A)\right]\end{aligned}$$

Finally, we obtain

$$\begin{aligned}\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1-Z\right)^2 &= \frac{\det(\Theta_0+A)^n}{\det(\Theta_0)^n}\left(\frac{\det(\Theta_0)}{\det(\Theta_0+2A)}\right)^{n/2}-1+\sum_{i=1}^n \mathrm{var}(x_i^TAx_i)\\ &\quad -2n\mathrm{tr}[((\Theta_0+A)^{-1}-\Sigma_0)A]\end{aligned}$$

This finishes the proof.

□

We apply the Lemmas above to the special case when $g(\Theta)=\xi_1^T\Theta\xi_2$.

Lemma 16. *Let* $A:=\Theta_0(\xi_1\xi_2^T+\xi_2\xi_1^T)\Theta_0/(2\sigma\sqrt{m_n})$. *Then*

$$\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)}-1-Z\right)^2=\mathcal{O}(\delta_n).$$

Proof of Lemma 16. We apply Lemma 15 with $A:=\Theta_0(\xi_1\xi_2^T+\xi_2\xi_1^T)\Theta_0/(2\sigma\sqrt{m_n})$, where

$\sigma^2 := \xi_1^T \Theta^0 \xi_1 \xi_2^T \Theta^0 \xi_2 + (\xi_1^T \Theta^0 \xi_2)^2$. Then $\mathrm{tr}(\Sigma_0 A) = \xi_1^T \Theta^0 \xi_2$ and

$$\begin{aligned}
\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)} - 1 - Z\right)^2 &= \underbrace{\left(1 + \frac{(\xi_1^T\Theta^0\xi_2)^2/(4\sigma^2 m_n)}{1+2\xi_1^T\Theta^0\xi_2/(\sigma m_n)}\right)^{n/2} - 1}_{i} \\
&\quad + \frac{n(\xi_1^T\Theta^0\xi_1\xi_1^T\Theta^0\xi_1 + (\xi_1^T\Theta^0\xi_2)^2)}{\sigma^2 m_n} - \frac{n}{\sigma^2 m_n}\frac{(\xi_1^T\Theta^0\xi_2)^2}{1+\xi_1^T\Theta^0\xi_2/\sqrt{m_n}} \\
&= \mathcal{O}(\delta_n) + \mathcal{O}\left(\frac{n}{m_n}\right) = \mathcal{O}(\delta_n),
\end{aligned}$$

where in the last step we used Lemma 17 to conclude that $i = \mathcal{O}(\delta_n)$. □

Lemma 17. *Let $0 < \delta = \delta_n \to 0$ and let $a, b = \mathcal{O}(1)$. Then*

$$\left(1 + \frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})}\right)^{n/2} - 1 = O(\delta).$$

Proof of Lemma 17.

$$\begin{aligned}
\left(1 + \frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})}\right)^{n/2} - 1 &= e^{\frac{n}{2}\log\left(1+\frac{\delta a}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})}\right)} - 1 \\
&= e^{\frac{n}{2}\left[\frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})} + o\left(\frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})}\right)\right]} - 1
\end{aligned}$$

Next

$$\frac{an\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})} = \mathcal{O}(\delta).$$

Hence, and using that $e^x - 1 = o(x)$ for $x \to 0$, we obtain

$$e^{\frac{n}{2}\left[\frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})} + o\left(\frac{a\delta}{\sqrt{n}(b\sqrt{\delta}+\sqrt{n})}\right)\right]} - 1 = \mathcal{O}(\delta).$$

□

Proof of Theorem 8. The proof follows in the same way as the proof of Theorem 2, but we need to use Lemma 16 to conclude that

$$\mathbb{E}\left(\frac{p_{\Theta_0+A}(x)}{p_{\Theta_0}(x)} - 1 - \mathrm{vec}(A)^T\mathrm{vec}(s_{\Theta_0}(x))\right)^2 = \mathcal{O}(\delta_n).$$

□

13.8 Proofs for Section 9

We first recall a version of Theorem 2.4 from [9].

Theorem 12 (a version of Theorem 2.4 in [9])**.** *Suppose that $X \sim \mathcal{N}(0, \Theta_0^{-1})$, assume that $s\log p/n = o(1)$. Consider the nodewise regression estimator $\hat{\Theta}_j$ and the corresponding $\hat{\tau}_j^2$ with $\lambda_j = \lambda \asymp \sqrt{\log p/n}$ for $j = 1, \ldots, p$. Then on the set $\mathcal{T} := \{\|X_{-j}^T(X_j - X_{-j}\gamma_j^0)\|_\infty/n \leq c\lambda\}$*

(where c is some sufficiently large constant), we have the following claims for $j = 1, \ldots, p$ for all $\Theta_0 \in \mathcal{G}$

$$\|\hat{\Theta}_j - \Theta_j^0\|_1 = \mathcal{O}(s\lambda), \quad |\hat{\tau}_j^2 - \tau_j^2| = \mathcal{O}(\sqrt{s \log p/n}).$$

Proof of Lemma 5. Letting $\epsilon_j := X_j - X_{-j}\gamma_j^0$ for $j = 1, \ldots, p$, we have $\epsilon_j \sim N_n(0, \Gamma_j^T \Gamma_j I)$. We have by assumption $1/\Lambda_{\min}(\Theta) = \mathcal{O}(1), \Lambda_{\max}(\Theta) = \mathcal{O}(1)$ that $\Gamma_j^T \Gamma_j = \mathcal{O}(1)$. Further we have $X_{-j,i} \sim N_{p-1}(0, \Sigma_{-j,-j}^0)$. Then under $1/\Lambda_{\min}(\Theta) = \mathcal{O}(1)$, one can check that $\Lambda_{\min}(\Sigma_{-j,-j}^0) \geq L > 0$ and $\|\Sigma_{-j,-j}^0\|_\infty = \mathcal{O}(1)$. Hence the conditions of Theorem 1 are satisfied and it follows that

$$(\mathbb{E}\|\hat{\gamma}_j - \gamma_j\|_1^k)^{1/k} = \mathcal{O}(s\lambda).$$

□

Proof of Lemma 6. **Proof of part 1**
Without loss of generality, let $j = 1$. We first show that $\mathbb{E}\frac{1}{\hat{\tau}_1^8} = \mathcal{O}(n^2)$. First observe that

$$\begin{aligned} \mathbb{P}(\hat{\tau}_1^2 \leq t) &= \mathbb{P}(\hat{\Gamma}_1^T \hat{\Sigma} \hat{\Gamma}_1 + \lambda\|\hat{\gamma}_1\|_1 \leq t) \\ &\leq \mathbb{P}(\hat{\Gamma}_1^T \hat{\Sigma} \hat{\Gamma}_1 \leq t \wedge \lambda\|\hat{\gamma}_1\|_1 \leq t) \end{aligned}$$

Using the following lower bound

$$\begin{aligned} \hat{\Gamma}_1^T \hat{\Sigma} \hat{\Gamma}_1 &= \hat{\Sigma}_{11} - 2\hat{\Sigma}_{1,-1}^T \hat{\gamma}_1 + \hat{\gamma}_1^T \hat{\Sigma}_{-1,-1} \hat{\gamma}_1 \\ &\geq \hat{\Sigma}_{11} - 2|\hat{\Sigma}_{1,-1}^T \hat{\gamma}_1| + \hat{\gamma}_1^T \hat{\Sigma}_{-1,-1} \hat{\gamma}_1 \\ &\geq \hat{\Sigma}_{11} - 2\|\hat{\Sigma}_{1,-1}\|_\infty \|\hat{\gamma}_1\|_1 + \hat{\gamma}_1^T \hat{\Sigma}_{-1,-1} \hat{\gamma}_1 \\ &\geq \hat{\Sigma}_{11} - 2\hat{\Sigma}_{11} t/\lambda \\ &= \hat{\Sigma}_{11}(1 - 2t/\lambda), \end{aligned}$$

we obtain that

$$\begin{aligned} \mathbb{P}(\hat{\Gamma}_1^T \hat{\Sigma} \hat{\Gamma}_1 \leq t \wedge \lambda\|\hat{\gamma}_1\|_1 \leq t) &\leq \mathbb{P}(\hat{\Sigma}_{11}(1 - 2t/\lambda) \leq t \wedge \lambda\|\hat{\gamma}_1\|_1 \leq t) \\ &\leq \mathbb{P}(\hat{\Sigma}_{11}(1 - 2t/\lambda) \leq t). \end{aligned}$$

Next $\hat{\Sigma}_{11} = e_1^T X^T X e_1/n \sim \Sigma_{11}\chi_n^2$ (by assumed Gaussianity of X). Using Chernoff bounds, we have for $Z \sim \chi_n^2$ the following upper bound

$$P(Z \leq x) \leq \left(\frac{x}{n} e^{1 - x/n}\right)^{n/2} = \left(\frac{e}{n}\right)^{n/2} x^{n/2} e^{-x/2}.$$

Hence for $t/\lambda \leq 1/2$ it holds

$$\begin{aligned} \mathbb{P}(\hat{\Sigma}_{11}(1 - 2t/\lambda) \leq t) &= \mathbb{P}\left(\hat{\Sigma}_{11}/\Sigma_{11} \leq \frac{t}{(1 - 2t/\lambda)\Sigma_{11}}\right) \\ &\leq \left(\frac{e}{n}\right)^{n/2} \left(\frac{t}{(1 - 2t/\lambda)\Sigma_{11}}\right)^{n/2} e^{-\left(\frac{t}{(1-2t/\lambda)\Sigma_{11}}\right)/2}. \end{aligned}$$

Hence collecting the above inequalities, we have so far shown that for any $t/\lambda \leq 1/2$ it holds

$$\mathbb{P}(\hat{\tau}_1^2 \leq t) \leq \left(\frac{e}{n}\right)^{n/2} \left(\frac{t}{(1 - 2t/\lambda)\Sigma_{11}}\right)^{n/2} e^{-\left(\frac{t}{(1-2t/\lambda)\Sigma_{11}}\right)/2}. \tag{15}$$

Then by rewriting the expectation as an integral

$$\begin{aligned}\mathbb{E}\frac{1}{\hat{\tau}_1^8} &= \int_0^\infty \mathbb{P}(1/\hat{\tau}_1^8 > x)dx \\ &= \int_0^1 \mathbb{P}(1/\hat{\tau}_1^8 > x)dx + \int_1^{\left(\frac{1}{2}\lambda\right)^{-4}} \mathbb{P}(1/\hat{\tau}_1^8 > x)dx + \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(1/\hat{\tau}_1^8 > x)dx \\ &\leq 1 + \left(\frac{1}{2}\lambda\right)^{-4} + \underbrace{\int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(1/\hat{\tau}_1^8 > x)dx}_{ii}\end{aligned}$$

Next we calculate an upper bound on ii.

$$\begin{aligned}\int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(1/\hat{\tau}_1^8 > x)dx &= \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(1/\hat{\tau}_1^2 > x^{1/4})dx \\ &= \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(\hat{\tau}_1^2 < x^{-1/4})dx\end{aligned}$$

Now we can use the bound (15) since $x^{-1/4} \leq 1/2\lambda$. We obtain

$$\begin{aligned}\int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \mathbb{P}(\hat{\tau}_1^2 < x^{-1/4})dx &= \left(\frac{e}{n}\right)^{n/2} \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \left(\frac{1}{x(1-2/(x\lambda))\Sigma_{11}}\right)^{n/2} e^{-\left(\frac{1}{x(1-2/(x\lambda))\Sigma_{11}}\right)/2}dx \\ &\leq \left(\frac{e}{n}\right)^{n/2} \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \left(\frac{1}{x(1-2/(x\lambda))\Sigma_{11}}\right)^{n/2} \underbrace{e^{-\left(\frac{1}{x(1-2/(x\lambda))\Sigma_{11}}\right)/2}}_{\leq 1 \text{ since } x^{-1/4} \leq \frac{1}{2}\lambda}dx \\ &\leq \left(\frac{e}{n}\right)^{n/2} \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \left(\frac{2}{x\Sigma_{11}}\right)^{n/2} dx \\ &\leq \left(\frac{2e}{\Sigma_{11}n}\right)^{n/2} \int_{\left(\frac{1}{2}\lambda\right)^{-4}}^\infty \left(\frac{1}{x}\right)^{n/2} dx \\ &\leq \left(\frac{2e}{\Sigma_{11}n}\right)^{n/2} \frac{(\log p/n)^{n/4-1/2}}{2n-4} = o(1),\end{aligned}$$

where we used that under $1/\Lambda_{\min}(\Theta) = \mathcal{O}(1)$, it holds that $1/\Theta_{11} = \mathcal{O}(1)$. Hence

$$\mathbb{E}\frac{1}{\hat{\tau}_1^8} = \mathcal{O}\left(\left(\frac{1}{2}\lambda\right)^{-4}\right) = \mathcal{O}(n^2/(\log p)^2).$$

Denote

$$\begin{aligned}\mathcal{T} \;:=\; &\{\|X_{-j}^T(X_j - X_{-j}\gamma_j^0)\|_\infty/n \leq c\lambda, \\ &X_{-j}^T X_{-j}/n \text{ satisfies the compatibility condition with a universal constant}\}.\end{aligned}$$

Then by Theorem 12, on $\mathcal{T}$ we have $1/\hat{\tau}_1^8 = \mathcal{O}(1)$. But then

$$\begin{aligned} \mathbb{E}\frac{1}{\hat{\tau}_1^8} &= \mathbb{E}\frac{1}{\hat{\tau}_1^8}1_{\mathcal{T}} + \mathbb{E}\frac{1}{\hat{\tau}_1^8}1_{\mathcal{T}^c} \\ &\leq \mathcal{O}(1) + \mathcal{O}(n^2)e^{-\tau \log p} = \mathcal{O}(1). \end{aligned}$$

Proof of part 2 First we show that $\hat{\tau}_j^4 = \mathcal{O}(1)$.

$$\begin{aligned} \hat{\tau}_j^2 &= \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n + \lambda\|\hat{\gamma}_j\|_1 \\ &= \hat{\Gamma}_j^T\hat{\Sigma}\hat{\Gamma}_j/n + \lambda\|\hat{\gamma}_j\|_1 \\ &\leq \|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty + \lambda\|\hat{\gamma}_j\|_1 \\ &= \|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty + \lambda\|\gamma_j\|_1 \\ &\quad + \ \|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda(\|\hat{\gamma}_j\|_1 - \|\hat{\gamma}_j\|_1). \end{aligned}$$

Hence by basic calculations

$$\begin{aligned} \mathbb{E}\hat{\tau}_j^4 &= \big\|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda\|\gamma_j\|_1\big]^2 \\ &\quad + \ \left[\|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda\|\gamma_j\|_1\right] \mathcal{O}(\underbrace{\mathbb{E}\left[\|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda(\|\hat{\gamma}_j\|_1 - \|\gamma_j\|_1)\right])}_{i} \\ &\quad + \ \mathcal{O}(\underbrace{\mathbb{E}\left[\|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda(\|\hat{\gamma}_j\|_1 - \|\gamma_j\|_1)\right]^2}_{ii}). \end{aligned}$$

First observe that $\left[\|\Gamma_j\|_1^2\|\Sigma\|_\infty + \lambda\|\gamma_j\|_1\right]^2 = \mathcal{O}(s)$.
We now consider i. By the triangle inequality and Lemma 5

$$\mathbb{E}\lambda(\|\hat{\gamma}_j\|_1 - \|\gamma_j\|_1) \leq \lambda\mathbb{E}\|\hat{\gamma}_j - \gamma_j\|_1 = \mathcal{O}(s\lambda).$$

Further, we have $\mathbb{E}\|\hat{\Gamma}_j - \Gamma_j\|_1^4 = \mathcal{O}(s^4\lambda^4)$ (as in Lemma 5) and $\|\Gamma_j\|_1 = \mathcal{O}(\sqrt{s})$. Hence

$$\mathbb{E}\|\hat{\Gamma}_j\|_1^4 = \mathbb{E}\|\hat{\Gamma}_j - \Gamma_j + \Gamma_j\|_1^4 \leq \mathcal{O}(s^2).$$

Thus

$$\begin{aligned} \mathbb{E}|\|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\Sigma\|_\infty| &\leq \mathbb{E}|\|\hat{\Gamma}_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\hat{\Sigma}\|_\infty| + \mathbb{E}|\|\Gamma_j\|_1^2\|\hat{\Sigma}\|_\infty - \|\Gamma_j\|_1^2\|\Sigma\|_\infty| \\ &= \mathcal{O}(s^2). \end{aligned}$$

Therefore, we conclude that $i = \mathcal{O}(s^2)$. Next we consider ii. This can be bounded similarly, but we also need $(\mathbb{E}\|\hat{\Gamma}_j - \Gamma_j\|_1^8)^{1/8} = \mathcal{O}(s\lambda)$. Then we can show $ii = \mathcal{O}(s^4)$.
Hence, by Theorem 12, on $\mathcal{T}$ we have that $\hat{\tau}_j^2 = \mathcal{O}(1)$, hence it follows

$$\mathbb{E}\hat{\tau}_j^4 = \mathbb{E}\hat{\tau}_j^4 1_{\mathcal{T}} + \mathbb{E}\hat{\tau}_j^4 1_{\mathcal{T}^c} = \mathcal{O}(1).$$

We have under $1/\Lambda_{\min}(\Theta) = \mathcal{O}(1)$, that $\tau_j = \mathcal{O}(1)$ and hence

$$|\hat{\tau}_j^2 - \tau_j^2|^2 \leq |\hat{\tau}_j^2|^2 + 2|\hat{\tau}_j^2||\tau_j^2| + |\tau_j^2|^2 = \mathcal{O}(\hat{\tau}_j^2 + \hat{\tau}_j^4).$$

We can then apply the same procedure:

$$\mathbb{E}|\hat{\tau}_j^2 - \tau_j^2|^2 = \mathbb{E}|\hat{\tau}_j^2 - \tau_j^2|^2 1_{\mathcal{T}} + \mathbb{E}|\hat{\tau}_j^2 - \tau_j^2|^2 1_{\mathcal{T}^c},$$

to show the claim.

Proof of part 3

$$\begin{array}{rcl} \mathbb{E}|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}| & \leq & \mathbb{E}|\frac{\hat{\tau}_j^2 - \tau_j^2}{\hat{\tau}_j^2 \tau_j^2}| \\ & \leq & 1/\tau_j^2 \sqrt{\mathbb{E}|\hat{\tau}_j^2 - \tau_j^2|^2} \sqrt{\mathbb{E} 1/\hat{\tau}_j^4}. \end{array}$$

Using Parts 1 and 2, we obtain the claim.

□

Proof of Lemma 7. By Theorem 12, when $\lambda \geq c\tau\sqrt{\log p/n}$, on the set

$$\mathcal{T} := \{\|\epsilon_j^T X\|_\infty / n \leq 2\sigma\tau\sqrt{\log p/n}\}$$

it holds that

$$\|\hat{\Theta}_j - \Theta_j^0\|_1 = \mathcal{O}(s\lambda), \quad |\hat{\tau}_j^2 - \tau_j^2| = \mathcal{O}(\sqrt{s}\lambda).$$

Next we rewrite

$$\mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1^2 = \mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1^2 1_{\mathcal{T}} + \mathbb{E}\|\hat{\Theta}_j - \Theta_j^0\|_1^2 1_{\mathcal{T}^c}.$$

Then

$$\begin{array}{rcl} \mathbb{E}_{\Theta_0}\|\hat{\Theta}_j - \Theta_j^0\|_1^2 & = & \mathbb{E}_{\Theta_0}\left[\|\hat{\gamma}_j - \gamma_j^0\|_1/\hat{\tau}_j^2 + \|\gamma_j^0\|_1 |1/\hat{\tau}_j^2 - 1/\tau_j^2|\right]^2 \\ & = & \mathbb{E}_{\Theta_0}\|\hat{\gamma}_j - \gamma_j^0\|_1^2/\hat{\tau}_j^4 + 2\mathbb{E}_{\Theta_0}\|\hat{\gamma}_j - \gamma_j^0\|_1/\hat{\tau}_j^2 \|\gamma_j^0\|_1 |1/\hat{\tau}_j^2 - 1/\tau_j^2| \\ & & + \ \mathbb{E}\|\gamma_j^0\|_1^2 |1/\hat{\tau}_j^2 - 1/\tau_j^2|^2 \\ & \leq & \sqrt{\mathbb{E}_{\Theta_0}\|\hat{\gamma}_j - \gamma_j^0\|_1^4} \sqrt{\mathbb{E} 1/\hat{\tau}_j^8} + 2\|\gamma_j^0\|_1 \sqrt{\mathbb{E}_{\Theta_0}\|\hat{\gamma}_j - \gamma_j^0\|_1^2 \mathbb{E}_{\Theta_0}|\hat{\tau}_j^2 - \tau_j^2|^2} \\ & & + \ \|\gamma_j^0\|_1^2 \mathbb{E}|1/\hat{\tau}_j^2 - 1/\tau_j^2|^2 \\ & \leq & s^2\lambda^2 \mathcal{O}(1) + \sqrt{s} s\lambda \sqrt{s}\lambda + s\mathcal{O}(1) s\lambda^2 \\ & = & \mathcal{O}(s^2\lambda^2), \end{array}$$

where in the last display we used Lemmas 5 and 6.

□

Proof of Lemma 8. By the Karush-Kuhn-Tucker conditions corresponding to the nodewise

Lasso estimator, we have $\|\hat{\Sigma}\hat{\Theta} - I\|_\infty = O_P(\lambda)$. Hence, and applying Lemma 7, we obtain

$$\begin{aligned}
\mathbb{E}_{\Theta_0}(\hat{T}_{ij} - \Theta^0_{ij}) &= \underbrace{\mathbb{E}_{\Theta_0}(\Theta^0_i)^T(\hat{\Sigma} - \Sigma_0)\Theta^0_j}_{=0} + \mathbb{E}_{\Theta_0}(\hat{\Theta}_i - \Theta^0_i)^T(\hat{\Sigma}\Theta^0_j - e_j) \\
&+ \mathbb{E}_{\Theta_0}(\hat{\Sigma}\hat{\Theta}_i - e_i)^T(\hat{\Theta}_j - \Theta^0_j) \\
&\leq \mathbb{E}_{\Theta_0}\|\hat{\Theta}_i - \Theta^0_i\|_1\|\hat{\Sigma}\Theta^0_j - e_j\|_\infty + \mathbb{E}_{\Theta_0}\lambda\|\hat{\Theta}_i - \Theta^0_i\|_1 \\
&\leq \lambda\mathbb{E}_{\Theta}\|\hat{\Theta}_i - \Theta^0_i\|_1 + \mathcal{O}(s\lambda^2) \\
&\leq \mathcal{O}(s\lambda^2) = o(1/\sqrt{n}).
\end{aligned}$$

□

Proof of Theorem 9. The lower bound follows by Theorem 8.
The strong asymptotic unbiasedness of $\hat{T}_{ij}$ follows by Lemma 8. It remains to calculate the variance of $\hat{T}_{ij}$. First we have that

$$\mathrm{var}((\Theta^0_i)^T(\hat{\Sigma} - \Sigma_0)\Theta^0_j) = \frac{1}{n}\mathrm{var}((\Theta^0_i)^T X_1 X_1^T \Theta^0_j) = (\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2)/n.$$

By basic calculations, it follows that

$$\begin{aligned}
\mathrm{var}(\hat{T}_{ij}) &= \mathrm{var}((\Theta^0_i)^T\hat{\Sigma}\Theta^0_j - (\hat{\Sigma}\Theta^0_i - e_i)^T(\hat{\Theta}_j - \Theta^0_j) + (\hat{\Theta}_i - \Theta^0_i)^T(\hat{\Sigma}\hat{\Theta}_j - e_j)) \\
&\leq (\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2)/n + \mathcal{O}(\mathbb{E}((\hat{\Sigma}\Theta^0_i - e_i)^T(\hat{\Theta}_j - \Theta^0_j))^2) \\
&+ \mathcal{O}(\mathbb{E}((\hat{\Theta}_i - \Theta^0_i)^T(\hat{\Sigma}\hat{\Theta}_j - e_j))^2) \\
&\leq (\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2)/n + \mathcal{O}(\mathbb{E}\|\hat{\Sigma}\Theta^0_i - e_i\|^2_\infty\|\hat{\Theta}_j - \Theta^0_j\|^2_1) \\
&+ \mathcal{O}(\mathbb{E}\|\hat{\Theta}_i - \Theta^0_i\|^2_1\|\hat{\Sigma}\hat{\Theta}_j - e_j\|^2_1).
\end{aligned}$$

Now we have by Lemma 7

$$\mathbb{E}\|\hat{\Sigma}\Theta^0_i - e_i\|^2_\infty\|\hat{\Theta}_j - \Theta^0_j\|^2_1 \leq \lambda^2\mathbb{E}\|\hat{\Theta}_j - \Theta^0_j\|^2_1 = \mathcal{O}(s^2\lambda^4) = o(1/n).$$

Hence we conclude

$$\mathrm{var}(\hat{T}_{ij}) = (\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2)/n + o(1/n).$$

□

13.9 Proofs for section 10

Proof of Theorem 10. Denote $h := a/\sqrt{m_n}$. The density is given by

$$p_\beta(x) = p_\beta(x_1, \ldots, x_n) = \prod_{i=1}^n \phi(x_i - \beta_i) = \frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}(x-\beta)^T(x-\beta)}.$$

Then we have

$$\begin{aligned}
\frac{p_\beta(x) - p_{\beta_0}(x)}{p_{\beta_0}(x)} &= \frac{e^{-\frac{1}{2}(x-\beta-h)^T(x-\beta-h)} - e^{-\frac{1}{2}(x-\beta)^T(x-\beta)}}{e^{-\frac{1}{2}(x-\beta)^T(x-\beta)}} \\
&= \frac{e^{-\frac{1}{2}(x-\beta)^T(x-\beta)+(x-\beta)h-\frac{1}{2}h^Th}}{e^{-\frac{1}{2}(x-\beta)^T(x-\beta)}} - 1 \\
&= e^{-(x-\beta)h-\frac{1}{2}h^Th} - 1
\end{aligned}$$

We have $\epsilon^T h \sim \mathcal{N}(0, h^T h)$. Then $\mathbb{E}e^{t\epsilon^T h} = e^{\frac{1}{2}t^2 h^T h}$. Hence

$$\begin{aligned}
\mathbb{E}\left(e^{-\epsilon^T h + \frac{1}{2}h^T h} - 1 - \epsilon^T h\right)^2 &= e^{h^T h} - 1 - h^T h \\
&= \mathcal{O}(h^T h).
\end{aligned}$$

Therefore, we can conlude the result as in the proof of Theorem 2. □

13.10 Proofs for Section 11

Proof of Theorem 11. Let

$$\Lambda_n := \sum_{i=1}^n \log p_{\theta+h/\sqrt{n}}(X_i) - \log p_\theta(X_i).$$

Under P_θ, by a two-term Taylor expansion we obtain

$$\Lambda_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T s_\theta(X_i) + \frac{1}{2}h^T\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i)h + o(h^T\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i)h).$$

By assumption, we have that $\|\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) + I_\theta\|_\infty = \mathcal{O}_P(\lambda)$. Note that every $h \in \Theta$ is s-sparse and furthermore we assume that $\|h\|_2 = \mathcal{O}(1)$. Then $\|h\|_1 = \mathcal{O}(\sqrt{s})$. Hence

$$\|h^T(\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) + I(\theta))h\|_\infty \le \|h\|_1^2\|\sum_{i=1}^n \dot{s}_\theta(X_i) + I_\theta\|_\infty = \mathcal{O}_P(s\lambda) = o_P(1).$$

Therefore,

$$\Lambda_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T s_\theta(X_i) - \frac{1}{2}h^T I(\theta)h + o_P(1).$$

We introduce the following notation. Let

$$V := \begin{pmatrix} V_\theta & P_\theta(l_\theta h^T s_\theta) \\ P_\theta(l_\theta h^T s_\theta) & h^T I(\theta) h \end{pmatrix}$$

Furthermore, we denote the entries of the matrix V by v_{ij}. Then by the central limit theorem we have for any $a \in \mathbb{R}^2$ that

$$a^T V^{-1/2}(\sqrt{n}(T_n - g(\theta)), \Lambda_n + \frac{1}{2}h^T I(\theta)h) \rightsquigarrow a^T Z, \text{ where } Z \sim \mathcal{N}(0, I_2).$$

Then by the Wold device we have

$$Z_n := V^{-1/2} \begin{pmatrix} \sqrt{n}(T_n - g(\theta)) \\ \Lambda_n + \frac{1}{2} h^T I(\theta) h \end{pmatrix} \overset{\theta}{\rightsquigarrow} \mathcal{N}_2(0, I) \sim Z.$$

Now let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. We may write

$$\mathbb{E}_{\theta+h/\sqrt{n}} f \left(\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}} \right) = \mathbb{E}_\theta f \left(\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}} \right) e^{\Lambda_n}.$$

Consider the function

$$\psi(x_1, x_2) = \begin{pmatrix} 1/\sqrt{v_{11}} & 0 \\ 0 & 1 \end{pmatrix} \left[V^{1/2}(x_1, x_2) + \begin{pmatrix} -v_{12} \\ -v_{22}/2 \end{pmatrix} \right]$$

Let $X_n := (X_{n,1}, X_{n,2}) = \psi(Z_n) = \left(\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}}, \Lambda_n \right)$. Then we have

$$\mathbb{E}_\theta f \left(\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}} \right) e^{\Lambda_n} = \mathbb{E}_\theta f(X_{n,1}) e^{X_{n,2}}.$$

Similarly,

$$\begin{aligned} U_n &:= (U_{n,1}, U_{n,2}) = \psi(Z) = \begin{pmatrix} 1/\sqrt{v_{11}} & 0 \\ 0 & 1 \end{pmatrix} \left[V^{1/2} \mathcal{N}(0, I_2) + \begin{pmatrix} -v_{12} \\ -v_{22}/2 \end{pmatrix} \right] \\ &= \mathcal{N} \left(\begin{pmatrix} -\frac{v_{12}}{\sqrt{v_{11}}} \\ -\frac{v_{22}}{2} \end{pmatrix}, \begin{pmatrix} 1 & \frac{v_{12}}{\sqrt{v_{11}}} \\ \frac{v_{12}}{\sqrt{v_{11}}} & v_{22} \end{pmatrix} \right). \end{aligned}$$

Since we know that $Z_n \rightsquigarrow Z$, we hope that in some sense $X_n = \psi(Z_n)$ is close to $U_n = \psi(Z)$. Note that the function ψ depends on n.
We aim to apply Lemma 22 with $X_n = \psi(Z_n)$ and $U_n = \psi(Z)$ defined above and with the function $g(x_1, x_2) = f(x_1)e^{x_2}$. By Lemma 21, we have that $\lim_{m\to\infty} \lim_{n\to\infty} \mathbb{E} e^{U_{n,2}} \mathbf{1}_{B_m^c}(U_n) = 0$, where $B_m^c := \{x \in \mathbb{R}^2 : \|x\|_2 \geq m\}$.
Hence we get by the second part of Lemma 22

$$\lim_{n\to\infty} |\mathbb{E}g(X_n) - \mathbb{E}g(U_n)| = 0.$$

Next we calculate $\mathbb{E}g(U_n)$. We have

$$\mathbb{E}_\theta g(U_n) = \mathbb{E} f(U_{n,1}) e^{U_{n,2}} = \int_{\mathbb{R}^2} f(u_1)\, e^{u_2} f_{U_n}(u_1, u_2) du,$$

where f_Y denotes the density of a random variable Y. We use Lemma 18 to obtain that $f_{U_n}(u)e^{u_2} = f_Y(u)$, where

$$Y \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ v_{22}/2 \end{pmatrix}, \begin{pmatrix} 1 & \frac{v_{12}}{\sqrt{v_{11}}} \\ \frac{v_{12}}{\sqrt{v_{11}}} & v_{22} \end{pmatrix} \right).$$

Hence

$$\mathbb{E}_\theta g(U_n) \quad = \quad \int_{\mathbb{R}^2} f(u_1)\, f_Y(u_1, u_2) du = \mathbb{E} f(Y_1),$$

where $Y \sim \mathcal{N}(0,1)$. Hence for any bounded continuous function f we have shown

$$\lim_{n\to\infty} \left|\mathbb{E}_{\theta+h/\sqrt{n}} f\left(\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}}\right) - \mathbb{E}f(Y)\right| = 0.$$

By the Portmanteau Lemma (note that Y in the above display does not depend on n), we thus have

$$\frac{\sqrt{n}(T_n - g(\theta)) - v_{12}}{\sqrt{v_{11}}} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1).$$

Therefore, by the assumption on g, we get

$$\frac{\sqrt{n}(T_n - g(\theta)) + h^T\dot{g}(\theta) - P_\theta l_\theta h^T s_\theta}{{V_\theta}^{1/2}} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1).$$

□

Lemma 18. *Let $Z \in \mathbb{R}^2$ be $\mathcal{N}(\mu, \Sigma)$-distributed, where*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Suppose that $\mu_2 = -\sigma_{22}/2$. Let $Y \in \mathbb{R}^2$ be $\mathcal{N}(\mu + a, \Sigma)$-distributed, with

$$a = \begin{pmatrix} \sigma_{12} \\ \sigma_{22} \end{pmatrix}.$$

Let ϕ_Z be the density of Z and ϕ_Y be the density of Y. Then we have the following equality for all $z = (z_1, z_2) \in \mathbb{R}^2$:

$$\phi_Z(z)e^{z_2} = \phi_Y(z).$$

Proof. The density of Z is

$$\phi_Z(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)}.$$

It holds that

$$\Sigma^{-1}a = (0,1)^T.$$

Then

$$\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) = \frac{1}{2}(z-\mu-a)^T\Sigma^{-1}(z-\mu-a) + a^T\Sigma^{-1}(z-\mu) - \frac{1}{2}a^T\Sigma^{-1}a.$$

We also have

$$a^T\Sigma^{-1}(z-\mu) - \frac{1}{2}a^T\Sigma^{-1}a = (0,1)^T(z-\mu) - \frac{1}{2}(0,1)^T a = z_2 - \mu_2 - \frac{1}{2}\sigma_{22} = z_2.$$

□

Lemma 19. *Let μ and Σ be defined as follows*

$$\mu = \begin{pmatrix} -\frac{v_{12}}{\sqrt{v_{11}}} \\ -\frac{v_{22}}{2} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \frac{v_{12}}{\sqrt{v_{11}}} \\ \frac{v_{12}}{\sqrt{v_{11}}} & v_{22} \end{pmatrix}.$$

Suppose that $V_\theta = \mathcal{O}(1), 1/V_\theta = \mathcal{O}(1)$ and $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$. (The relationship between these quantities and the v_{ij}'s is given in the proof of Le Cam's lemma). Then

$$\|\mu\|_2^2 = \mathcal{O}(1) \qquad \textit{and} \qquad \Lambda_{\max}(\Sigma) = \mathcal{O}(1).$$

Proof. First observe that $v_{12}^2 = (\mathbb{E} l_\theta h^T s_\theta)^2 \le \mathbb{E} l_\theta^2 \mathbb{E}(h^T s_\theta)^2 = V_\theta h \mathbb{E} s_\theta s_\theta^T h \le V_\theta \Lambda_{\max}(\mathbb{E} s_\theta s_\theta^T) h^T h$. Then by assumption $\Lambda_{\max}(\mathbb{E} s_\theta s_\theta^T) = \mathcal{O}(1)$, $V_\theta = \mathcal{O}(1)$ and since $h^T h = \mathcal{O}(1)$, we have that $(\mathbb{E} l_\theta h^T s_\theta)^2 = \mathcal{O}(1)$. Also observe that $v_{22} = h^T I_\theta h \le \Lambda_{\max}(I_\theta) h^T h = \mathcal{O}(1)$ by assumption $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$.
Then, and by $1/V_\theta = \mathcal{O}(1)$, it follows that

$$\|\mu\|_2^2 = v_{12}^2/v_{11} + v_{22}^2/4 = (P_\theta l_\theta h^T s_\theta)^2/V_\theta + (h^T I_\theta h)^2/4 = \mathcal{O}(1).$$

We proceed to check that the eigenvalues of Σ are bounded. We have

$$\lambda_{1,2} = \frac{1 + v_{22} \pm \sqrt{D}}{2},$$

where $D = (1 + v_{22})^2 - 4(v_{22} - v_{12}^2/v_{11}) = (1 - v_{22})^2 + 4v_{12}^2/v_{11}$. Clearly, $D \ge 0$, and as above, one sees that $D = \mathcal{O}(1)$. Hence also $\Lambda_{\max}(\Sigma) = \mathcal{O}(1)$. □

Lemma 20. *Suppose that $X \sim \mathcal{N}_d(\mu, \Sigma)$, where $\|\mu\|^2 \le K$, $\Lambda_{\max}(\Sigma) \le L$ and assume that Σ is invertible. Then it holds that for all $r > \max\{K, Ld\}$*

$$P(\|X\|_2^2 > r) \le \left(\frac{r-K}{2Ld} e^{1 - \frac{r-K}{2Ld}}\right)^{d/2} + \left(\frac{r^2}{4LdK} e^{1 - \frac{r^2}{4LdK}}\right)^{d/2}.$$

Proof. First consider $Y \sim N_d(0, \Sigma)$ and denote $Z := \Sigma^{-1/2} Y \sim N_d(0, I)$. Note that $Z^T Z \sim \chi_d^2$. Since for any $y \in \mathbb{R}^d$ we have $y^T y \le \Lambda_{\max}(\Sigma) y^T \Sigma^{-1} y \le L y^T \Sigma^{-1} y$, then it follows

$$\begin{aligned} P(\|Y\|_2^2 > r) &\le P(L Y^T \Sigma^{-1} Y > r) \\ &= P\left(Z^T Z > \frac{r}{L}\right) \\ &\le \left(\frac{r}{Ld} e^{1 - \frac{r}{Ld}}\right)^{d/2}, \end{aligned} \tag{16}$$

where we used a Chernoff bound for χ_d^2 in the last inequality, which holds provided that $r > Ld$.
Consider now for any $r > K$ (assume that $K > 0$, otherwise if $K = 0$ we are done)

$$\begin{aligned} P(\|X\|_2^2 > r) &= P(\|X - \mu + \mu\|_2^2 > r) \\ &\le P(\|X - \mu\|_2^2 + 2\|X - \mu\|_2 \|\mu\|_2 + \|\mu\|_2^2 > r) \\ &\le P(\|X - \mu\|_2^2 + 2\|X - \mu\|_2 \sqrt{K} + K > r) \\ &\le P\left(\|X - \mu\|_2^2 > \frac{r-K}{2}\right) + P\left(\|X - \mu\|_2 > \frac{r}{2\sqrt{K}}\right) \\ &= P\left(\|X - \mu\|_2^2 > \frac{r-K}{2}\right) + P\left(\|X - \mu\|_2^2 > \frac{r^2}{4K}\right). \end{aligned}$$

Now since $X-\mu \sim N_d(0,\Sigma)$, we can apply (16) to conclude that

$$\begin{aligned} P(\|X\|_2^2 > r) &\leq P\left(\|X-\mu\|_2^2 > \frac{r-K}{2}\right) + P\left(\|X-\mu\|_2^2 > \frac{r^2}{4K}\right) \\ &\leq \left(\frac{r-K}{2Ld}e^{1-\frac{r-K}{2Ld}}\right)^{d/2} + \left(\frac{r^2}{4LdK}e^{1-\frac{r^2}{4LdK}}\right)^{d/2}, \end{aligned}$$

which holds if $r > \max\{K, Ld\}$. □

Lemma 21. *Suppose that*

$$U_n \sim \mathcal{N}\left(\left(\begin{array}{c} -\frac{v_{12}}{\sqrt{v_{11}}} \\ -\frac{v_{22}}{2} \end{array}\right), \left(\begin{array}{cc} 1 & \frac{v_{12}}{\sqrt{v_{11}}} \\ \frac{v_{12}}{\sqrt{v_{11}}} & v_{22} \end{array}\right)\right).$$

Suppose that $V_\theta = \mathcal{O}(1), 1/V_\theta = \mathcal{O}(1)$, $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$. (The relationship between these quantities and the v_{ij}'s is given in the proof of Le Cam's lemma).
Then it holds that

$$\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c} = 0,$$

where $B_m^c = \{x \in \mathbb{R}^2 : \|x\|_2 > m\}$.

Proof. By Lemma 18 we have that

$$\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c}(X) = \mathbb{E}\mathbf{1}_{B_m^c}(Y),$$

where

$$Y \sim \mathcal{N}\left(\left(\begin{array}{c} 0 \\ v_{22}/2 \end{array}\right), \left(\begin{array}{cc} 1 & \frac{v_{12}}{\sqrt{v_{11}}} \\ \frac{v_{12}}{\sqrt{v_{11}}} & v_{22} \end{array}\right)\right).$$

Further we have

$$\mathbb{E}\mathbf{1}_{B_m^c}(Y) = P(Y \in B_m^c) = P(\|Y\|_2 > m).$$

Denote $\mu_Y = \mathbb{E}Y$ and $\Sigma_Y = \text{var}(Y)$. By Lemma 20, for $m > \max\{Ld, K\}$ we have

$$P(\|Y\|_2^2 > m) \leq \left(\frac{m-K}{2Ld}e^{1-\frac{m-K}{2Ld}}\right)^{d/2} + \left(\frac{m^2}{4LdK}e^{1-\frac{m^2}{4LdK}}\right)^{d/2}, \tag{17}$$

where $\|\mu_Y\|_2^2 \leq K$ and $\Lambda_{\max}(\Sigma_Y) \leq L$ and $d=2$. By Lemma 19, we have $L = \mathcal{O}(1)$ and by assumption $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$ we have $K = \mathcal{O}(1)$. Therefore, and using (17), we can obtain an upper bound on $P(\|Y\|_2^2 > m)$ that depends on m but does not depend on n. This upper bound tends to zero for $m \to \infty$, therefore we have shown that

$$\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c}(U_n) = 0.$$

□

Lemma 22. *Assume the conditions of Theorem 11. Suppose that $Z_n \rightsquigarrow Z$, where Z is a random vector with values in $\mathbb{R}^2$. Let $X_n = \psi(Z_n)$ and $U_n = \psi(Z)$ with ψ as in Theorem 11. Then the following statements hold.*

1. *For any function $f : \mathbb{R}^2 \to \mathbb{R}$ which is bounded and continuous it holds that*

$$\lim_{n\to\infty} \mathbb{E}f(X_n) - \mathbb{E}f(U_n) = 0.$$

2. *Let f be any bounded and continuous function $f : \mathbb{R} \to \mathbb{R}$. Suppose that*

$$\lim_{m\to\infty} \lim_{n\to\infty} \mathbb{E}e^{U_{n,2}} \mathbf{1}_{B_m^c} = 0,$$

where $B_m^c := \{x \in \mathbb{R}^d : \|x\|_2 > m\}$. Then it holds that

$$\lim_{n\to\infty} \mathbb{E}f(X_{n,1})e^{X_{n,2}} - \mathbb{E}f(U_{n,1})e^{U_{n,2}} = 0.$$

Proof. We first prove the first statement. Let $\epsilon > 0$ and let $f : \mathbb{R}^2 \to \mathbb{R}$ be continuous and bounded.
Consider the map

$$x \mapsto \psi(x_1, x_2) = \underbrace{\begin{pmatrix} 1/\sqrt{v_{11}} & 0 \\ 0 & 1 \end{pmatrix}}_{D} \left[V^{1/2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -v_{12} \\ -v_{22}/2 \end{pmatrix} \right].$$

The map ψ is linear, i.e. $\psi(x) = Ax + b$ for some $A \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$ (A, b depending on n). Observe that for any $x \in \mathbb{R}^2$

$$\|Ax\|_2^2 = x^T A^T A x = x^T DVDx \le \Lambda_{\max}(DVD) x^T x.$$

By Lemma 19 we have that $\Lambda_{\max}(DVD) = \mathcal{O}(1)$ and $\|b\|_2 = \mathcal{O}(1)$. Therefore, for all $x \in \mathbb{R}^2$

$$\|Ax + b\|_2 = \mathcal{O}(\|x\|_2). \tag{18}$$

Take a compact rectangle $R \subset \mathbb{R}^2$ not depending on n and such that $P(Z \notin R) < \epsilon$.
Divide the rectangle R into a finite number of non-overlapping rectangles of diameter at most $\delta/L^{1/2}$, where L is a universal constant such that $L \ge \Lambda_{\max}(DVD)$. By construction, the number of these rectangles, denote it N, does not depend on n. So we have $R = \cup_{j=1}^N R_j$, where each R_j is a rectangle of diameter at most $\delta/L^{1/2}$.
For all $x, y \in R_j$ it holds that $\|x - y\|_2 \le \delta/L^{1/2}$ and thus

$$\|\psi(x) - \psi(y)\|_2 = \|A(x - y)\|_2 \le L^{1/2}\|x - y\|_2 \le \delta. \tag{19}$$

Note that by (18), there exists a compact set S not depending on n such that $\psi(R) \subset S$ for all n. The continuous function f is uniformly continuous on the compact set S. Hence for our ϵ there exists a $\delta > 0$ such that for all $z, v \in S$ it holds that if $\|z - v\|_2 < \delta$ then $|f(z) - f(v)| < \epsilon$. But then since for all $x, y \in R_j$ we have that $\psi(x), \psi(y) \in S$, we obtain by (19) and the absolute continuity of f that

$$|f(\psi(x)) - f(\psi(y))| < \epsilon$$

for all n. Take a point x_j from each set R_j and define $f_\epsilon = \sum_{j=1}^N f(\psi(x_j))\mathbf{1}_{R_j}$. Then $|f(\psi(x)) - f_\epsilon(x)| < \epsilon$ for all $x \in R$ (and all n) and hence if f takes values in $[-K, K]$, we have the following upper bounds

$$|\mathbb{E}f(\psi(Z)) - \mathbb{E}f_\epsilon(Z)| \leq \epsilon + 2KP(Z \notin R), \tag{20}$$

$$|\mathbb{E}f(\psi(Z_n)) - \mathbb{E}f_\epsilon(Z_n)| \leq \epsilon + 2KP(Z_n \notin R), \tag{21}$$

$$|\mathbb{E}f_\epsilon(Z_n) - \mathbb{E}f_\epsilon(Z)| \leq \sum_{j=1}^{N} |P(Z_n \in R_j) - P(Z \in R_j)||f(\psi(x_j))|. \tag{22}$$

Since $Z_n \rightsquigarrow Z$, for all $j = 1, \dots, N$ it holds

$$|P(Z_n \in R_j) - P(Z \in R_j)| \to 0.$$

Similarly,

$$|P(Z \notin R) - P(Z_n \notin R)| = |P(Z \in R) - P(Z_n \in R)| \to 0.$$

Finally, by construction we have $P(Z \notin R) < \epsilon$. We thus conclude that the upper bounds (20), (21) and (22) can be made smaller than $C\epsilon$ for n sufficiently large. The claim follows by combining the three upper bounds.

Next we prove the second statement. Denote $g(x_1, x_2) = f(x_1)e^{x_2}$. We write $g = g^+ - g^-$, where $g^+ = \max\{g, 0\}$ is the positive part and $g^- := \max\{-g, 0\}$ is the negative part. We first prove for the positive part g^+ that

$$\lim_{n\to\infty} \mathbb{E}g^+(X_n) - \mathbb{E}g^+(U_n) = 0. \tag{23}$$

For every m, since g^+ is non-negative, it holds that $g^+(x) \geq g^+(x)\mathbf{1}_{B_m}(x)$, where $B_m := \{x \in \mathbb{R}^2 : \|x\|_2 \leq m\}$. Hence

$$\begin{aligned}\mathbb{E}g^+(X_n) - \mathbb{E}g^+(U_n) &\geq \mathbb{E}g^+(X_n)\mathbf{1}_{B_m} - \mathbb{E}g^+(U_n) \\ &= [\mathbb{E}g^+(X_n)\mathbf{1}_{B_m} - \mathbb{E}g^+(U_n)\mathbf{1}_{B_m}] \\ &\quad + [\mathbb{E}g^+(U_n)\mathbf{1}_{B_m} - \mathbb{E}g^+(U_n)]\end{aligned}$$

We have $\mathbf{1}_{B_m} - 1 = -\mathbf{1}_{B_m^c}$. Taking limes inferior of both sides, it follows that

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}g^+(X_n) - \mathbb{E}g^+(U_n) &\geq \liminf_{n\to\infty}[\mathbb{E}g^+(X_n)\mathbf{1}_{B_m}(X_n) - \mathbb{E}g^+(U_n)\mathbf{1}_{B_m}] \\ &\quad + \liminf_{n\to\infty} -\mathbb{E}g^+(U_n)\mathbf{1}_{B_m^c}.\end{aligned}$$

For every fixed m, the function $x \mapsto g^+(x)\mathbf{1}_{B_m}(x)$ is bounded since g is continuous on the compact set B_m. We may thus apply the first result of the lemma to conclude

$$\liminf_{n\to\infty} \mathbb{E}g^+(X_n)\mathbf{1}_{B_m} - \mathbb{E}g^+(U_n)\mathbf{1}_{B_m} = 0.$$

Therefore, we have

$$\liminf_{n\to\infty} \mathbb{E}g^+(X_n) - \mathbb{E}g^+(U_n) \geq \liminf_{n\to\infty} -\mathbb{E}g^+(U_n)\mathbf{1}_{B_m^c}. \tag{24}$$

Next since $|f^+| \leq K$ we have $|-\mathbb{E}g^+(U_n)\mathbf{1}_{B_m^c}| \leq K\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c}$. Then the assumption

$$\lim_{m\to\infty}\lim_{n\to\infty} \mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c} = 0$$

implies that also

$$\lim_{m\to\infty}\liminf_n -\mathbb{E}g^+(U_n)\mathbf{1}_{B_m^c} = -\lim_{m\to\infty}\limsup_n \mathbb{E}g^+(U_n)\mathbf{1}_{B_m^c} = 0,$$

so we conclude that

$$\liminf_{n\to\infty} \mathbb{E}g^+(X_n) - \mathbb{E}g^+(U_n) \geq 0. \tag{25}$$

Now similarly, since $K - f^+ \geq 0$ (K is an upper bound on f), we have that

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{E}(K - f^+(X_{n,1}))e^{X_{n,2}} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}} \\
&\geq \liminf_{n\to\infty} \mathbb{E}(K - f^+(X_{n,1}))e^{X_{n,2}}\mathbf{1}_{B_m} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}\mathbf{1}_{B_m} \\
&\quad + \liminf_{n\to\infty} \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}\mathbf{1}_{B_m} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}.
\end{aligned}$$

By the the first part of the lemma, we have that for every m it holds

$$\liminf_n \mathbb{E}(K - f^+(X_{n,1}))e^{X_{n,2}}\mathbf{1}_{B_m} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}\mathbf{1}_{B_m} = 0,$$

since the function $(x_1, x_2) \mapsto (K - f^+(x_1))e^{x_2}\mathbf{1}_{B_m}(x_1, x_2)$ is bounded and continuous. For the second term, we have since $|K - f^+| \leq 2K$

$$|-\mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}\mathbf{1}_{B_m^c}| \leq 2K\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c}.$$

Hence by the assumption $\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{E}e^{U_{n,2}}\mathbf{1}_{B_m^c} = 0$, we have that

$$\liminf_{n\to\infty} -\mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}}\mathbf{1}_{B_m^c} = 0.$$

Thus we conclude that

$$\liminf_{n\to\infty} \mathbb{E}(K - f^+(X_{n,1}))e^{X_{n,2}} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}} \geq 0.$$

Now note that

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{E}(K - f^+(X_{n,1}))e^{X_{n,2}} - \mathbb{E}(K - f^+(U_{n,1}))e^{U_{n,2}} \\
&= \liminf_n -\mathbb{E}f^+(X_{n,1})e^{X_{n,2}} + \mathbb{E}f^+(U_{n,1})e^{U_{n,2}} \\
&= -\limsup_n \mathbb{E}f^+(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^+(U_{n,1})e^{U_{n,2}}.
\end{aligned}$$

So in conclusion we have shown that

$$\limsup_n \mathbb{E}f^+(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^+(U_{n,1})e^{U_{n,2}} \leq 0 \leq \liminf_n \mathbb{E}f^+(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^+(U_{n,1})e^{U_{n,2}}.$$

This proves (23).
The same procedure can be used for the negative part f^- (since f^- is also bounded and positive) to show that

$$\lim_{n\to\infty} \mathbb{E}f^-(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^-(U_{n,1})e^{U_{n,2}} = 0.$$

We then conclude that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}f(X_{n,1})e^{X_{n,2}} - \mathbb{E}f(U_{n,1})e^{U_{n,2}} &\leq \lim_{n\to\infty} |\mathbb{E}f^{+}(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^{+}(U_{n,1})e^{U_{n,2}}| \\ &\quad + \lim_{n\to\infty} |\mathbb{E}f^{-}(X_{n,1})e^{X_{n,2}} - \mathbb{E}f^{-}(U_{n,1})e^{U_{n,2}}| \\ &= 0.\end{aligned}$$

□

References

[1] P. Bühlmann and S. van de Geer. Statistics for high-dimensional data. *Springer*, 2011.

[2] T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *ArXiv: 1506.05539*, 2015.

[3] J. Jankova and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205 –1229, 2014.

[4] J. Jankova and S. van de Geer. Honest confidence regions and optimality for high-dimensional precision matrix estimation. *ArXiv:0710.2044*, 2015.

[5] L. Le Cam. Locally asymptotically normal families of distributions. *University of California Publications in Statistics*, 3:3798, 1960.

[6] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 06 2006.

[7] G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[8] Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *ArXiv: 1309.6024*, sep 2013.

[9] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2013.

[10] A. van der Vaart. Asymptotic statistics. *Cambridge University Press*, 2000.

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