UCLouvain
Institute of Statistics, Biostatistics and Actuarial Sciences

Distributed estimation and inferential tools for graphical models

Ensiyeh Nezakati Rezazadeh

A thesis submitted to the
UCLouvain
in partial fulfillment of the
requirements for the degree of
Doctor of Sciences

Thesis Committee:

Prof. **Eugen Pircalabelu**	*UCLouvain*	Advisor
Prof. **Johan Segers**	*UCLouvain*	Secretary
Prof. **Gerda Claeskens**	*KU Leuven*	
Prof. **Jacob Bien**	*University of Southern California (USC)*	
Prof. **Rainer von Sachs**	*UCLouvain*	Chairman

November, 2023

Acknowledgements

First of all, I would like to thank my supervisor Prof. Eugen Pircalabelu, for his support and continuous guidance over the past four years. Eugen, you helped me so much with my thesis, and you were always available for discussions, even during the busy periods of work. Thank you for those long, helpful meetings with no clear ending time which tended to infinity. Thank you for the freedom you gave me to come to your office whenever I wanted without fixing any appointments, discussions next to your door, and giving me advices on my problems. This thesis would certainly not be possible without your precious support. I would also like to express my gratitude to the jury members, including Professors Johan Segers, Gerda Claeskens, and Jacob Bien, for their constructive remarks and suggestions that significantly enhanced the quality of my thesis. In particular, I would like to thank Prof. Johan Segers for his comments on the theoretical part of the thesis in our annual meetings, which improved the quality of this thesis a lot. A special thanks goes to Prof. Rainer von Sachs for his incredible support and kindness during my PhD, and also for accepting to be the president of the jury.

Working in the Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA) made my life more pleasant and happier. Unfortunately, the first two years passed by during COVID. I did not see my colleagues that much, it was not possible to socialize with them, and it was quite hard. But in the last two years, when COVID finished, new colleagues arrived, we saw each other every day, and it showed me how wonderful and exciting it is to work in ISBA. Gradually, my colleagues became my nice friends! I want to express my big thanks first to Anas for his huge support on everything, listening patiently to me talking about my life and work, accompanying and helping me with my problems, making the days much more fun on the third floor, coffee/tea breaks, lunch discussions, evening shops, movie/series suggestions (discussions), etc. Thank you Anas for all those happy and pleasant moments you made in these two years.

My next thanks goes to Charlotte, who became a part of the institute in my final year of PhD, but made the happiest and the most pleasant moments for me! Charlotte, You and I grew close very swiftly. You always gave me a nice view from

the office in front, spread positive energy around yourself, and understood me so well when I was talking about myself! Thank you for accompanying me whenever I needed you, making our funny words/stickers, and guessing my songs! Thanks to Aigerim, who has been always a complete guideline book for me, who knows everything, and always provides me with information to deal with the problems that I have with living in Belgium! Thanks to Benjamin D. and Hortense for making the office atmosphere so friendly and pleasant. Thanks to Keivan for his positive energy, making fun, dancing in front of our door, and his afternoon "quotes". Thanks to Vanessa and Chikèola for their support and positive guidance whenever I was sad, to Patricia for making the third floor funnier, and to Mengxue for helping me patiently with my shopping with her beautiful smile. I would also like to thank all PhD and post-doc students at ISBA, including Antoine, Aurèle, Charles, Edouard, Fanny, Gabriel, Gilles, Hugues, Jean-Loup, John-John, Kassu, Lara, Lise, Madeline, Mathilde, Morine, Nathalie Lu., Oussama, Quentin, Rebecca, Sophie, Stefka, and Stèphane.

I am grateful to the administrative staff including Nadja, Nancy, Sophie, and Tatiana, for their availability and help with practical concerns. I also thank the SMCS team including Alain, Aurélie, Benjamin C., Catherine, Céline, Nathalie Le., Lieven, Séverine, and Vincent, for being such nice and kind colleagues on the third floor.

Finally, this work could not have been achieved without the support of my family. My special thanks go to my parents, who spread positive energy from a distance with their nice prayers for my future. Last but not the least thanks goes to my husband. Ali, we both know how difficult it was (and still is) for you to leave your job, your father, and your stable life to come with me on this complicated journey. I appreciate your encouragement and emotional support throughout this experience, especially in the last year of my PhD when it was so difficult to deal with me, and I was complaining about everything. Achieving this step would not be possible without your support. I hope happy/easy days come back again to our lives, and we experience more joyful moments together.

Contents

CHAPTER 1

Introduction

This thesis combines two statistical concepts: probabilistic graphical models and distributed algorithms. This type of 'divide and conquer' algorithms are becoming increasingly popular these days due to new challenges in collecting datasets, including security and privacy concerns, limited capacity of systems, and the presence of heterogeneous datasets from various locations. The broad scope of the thesis is to contribute to the estimation and inferential tools for Gaussian graphical models in a distributed setting. Throughout this chapter, we commence with an introduction to graphical models in Section 1.1, covering Markov properties and Gaussian graphical models. We also review estimation methods such as graphical Lasso and debiased graphical Lasso. In order to provide statistical inference for graphical models, an introduction to hypothesis testing is presented in Section 1.2. An overview of the motivation behind the distributed setting and the outline of the thesis are provided in Sections 1.3 and 1.4, respectively. Finally, some notation that is employed throughout the thesis is provided in Section 1.5.

1.1 Graphical models

Graphical models can be regarded as visual maps that help understanding complex systems. They are used in statistics, machine learning and artificial intelligence, and they show how different units are connected and dependent on each other. By representing the connections as edges in a graph, one can comprehend the overall structure of a system. Graphical models are used in many areas, like language processing, computer vision and decision-making. They are a valuable tool for understanding and solving real-world problems. In this section, we provide an introduction on graphical models, which is mostly retrieved from Lauritzen (1996), Bühlmann and van de Geer (2011) and Hastie et al. (2015).

A graph as a mathematical object, may be defined as a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

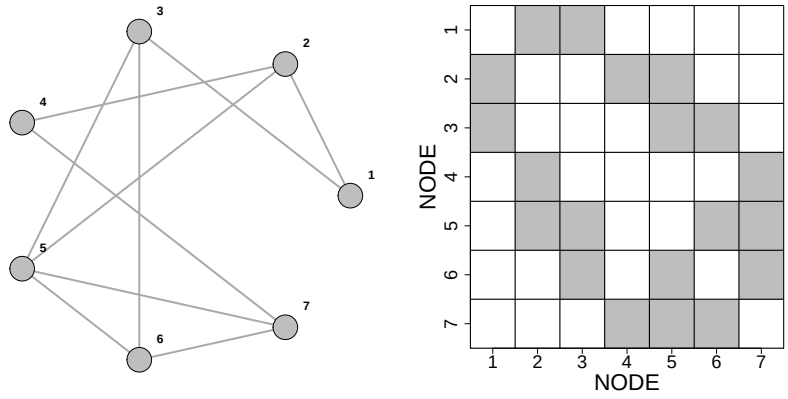


Figure 1.1: An undirected graph with nodes $\{1, 2, 3, 4, 5, 6, 7\}$.

$\mathcal{V}$ is a set of nodes or vertices and $\mathcal{E}$ is a set of edges. The set of edges $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$ consisting of ordered pairs of distinct nodes. Each edge is associated with a pair of nodes. Edges may in general be directed or undirected, but in this thesis, we deal with graphs with undirected edges only. An undirected edge between nodes $a, b \in \mathcal{V}$ is drawn if $(a, b) \in \mathcal{E}$ and $(b, a) \in \mathcal{E}$. Figure 1.1, on the left hand, presents an undirected graph with nodes $\mathcal{V} = \{1, 2, 3, 4, 5, 6, 7\}$ and edge set $\mathcal{E} = \{(1, 2), (1, 3), (2, 4), (2, 5), (3, 5), (3, 6), (4, 7), (5, 6), (5, 7), (6, 7)\}$. Note that due to the definition of the undirected edge, both pairs (a, b) and (b, a) should be included in the edge set $\mathcal{E}$. However, for the sake of visual simplicity in this thesis, the pair (b, a) which we consider equivalent to the pair (a, b) is not presented in the edge set $\mathcal{E}$. Undirected graphs are typically visualized by representing circles or points as nodes, and lines as edges.

The neighbors or adjacency set of a node a in an undirected graph $\mathcal{G}$ are denoted by $\text{adj}(\mathcal{G}, a) = \{b \in \mathcal{V}, b \neq a; \quad (a, b) \in \mathcal{E} \text{ and } (b, a) \in \mathcal{E}\}$. For example, the adjacency set of node 1 in Figure 1.1 is $\text{adj}(\mathcal{G}, 1) = \{2, 3\}$. The right hand side of Figure 1.1 represents the correponding matrix to the adjacency of all the nodes. The grey blocks in each row present the neighbors of the corresponding node on that line, and the white blocks correspond to the nodes that are not neighbors to the specified node on that line. In a graphical model with node set $\mathcal{V} = \{1, \ldots, p\}$, the nodes correspond to the index set of a collection of random variables

$$\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top \sim \mathcal{P},$$

where $\mathcal{P}$ is the probability distribution of $\dot{\mathbf{Y}}$. In this thesis, we denote a graph $\mathcal{G}$ with its nodes and edges as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. However, in some references (see Bühlmann and van de Geer, 2011) the pair $(\mathcal{G}, \mathcal{P})$ is referred to as a graphical model, indicating the probability distribution of the random variables associated to the graph.

1.1.1 Markov properties for undirected graphs

Graphs can be used to represent conditional independence structures in the form of Markov properties. Conditional independence between two random variables is defined as follows.

Definition 1.1.1. Random variables X and Y are conditionally independent given the random variable Z if the conditional distribution of X given Y and Z is the same as the conditional distribution of X given Z. One then writes $X \perp Y \mid Z$ or $X \perp_{\mathcal{P}} Y \mid Z$ if one wants to emphasize that independence depends on a specific probability distribution $\mathcal{P}$.

Markov properties associated with an undirected graph $\mathcal{G}$ are either pairwise or global.

Definition 1.1.2. An undirected graph $\mathcal{G}$ satisfies the **pairwise Markov property** if for any pair of unconnected nodes $(a, b) \notin \mathcal{E}$, $a \neq b$, we have

$$Y^a \perp Y^b \mid \dot{\mathbf{Y}}^{\mathcal{V}\setminus\{a,b\}},$$

where $\dot{\mathbf{Y}}^{\mathcal{V}\setminus\{a,b\}}$ is the vector $\dot{\mathbf{Y}}$ without elements Y^a and Y^b.

For example in Figure 1.1, the pairwise Markov property states that $1 \perp 5 \mid \{2,3,4,6,7\}$ and $4 \perp 6 \mid \{1,2,3,5,7\}$. Based on this definition, an undirected graph $\mathcal{G}$ is a **conditional independence graph**, if the pairwise Markov property holds. The global Markov property, which is defined below, is a stronger property. To define this property, we need a preliminary concept. A **path** is a sequence of nodes $\{j_1, \ldots, j_\ell\}$ such that $(j_i, j_{i+1}) \in \mathcal{E}$ for $i = 1, \ldots, \ell-1$. For a triple of disjoint sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathcal{V}$, we say that $\mathcal{C}$ **separates** $\mathcal{A}$ and $\mathcal{B}$ if every path from $j \in \mathcal{A}$ to $k \in \mathcal{B}$ contains a node in $\mathcal{C}$.

Definition 1.1.3. An undirected graph $\mathcal{G}$ satisfies the **global Markov property** if for any triple of disjoint sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ such that $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$, we have

$$\dot{\mathbf{Y}}^{\mathcal{A}} \perp \dot{\mathbf{Y}}^{\mathcal{B}} \mid \dot{\mathbf{Y}}^{\mathcal{C}},$$

where $\dot{\mathbf{Y}}^{\mathcal{A}} = \{Y^j; j \in \mathcal{A}\}$, and corresponding for $\dot{\mathbf{Y}}^{\mathcal{B}}$ and $\dot{\mathbf{Y}}^{\mathcal{C}}$.

In general, the global Markov property implies the pairwise Markov property. However, Lauritzen (1996) showed that the inverse holds for a large class of models with positive and continuous density with respect to the Lebesgue measure. A typical example of conditional independence graph is the Gaussian graphical model which is explained in the sequel.

1.1.2 Gaussian graphical models

A Gaussian graphical model (GGM) is a conditional independence graph with a multivariate Gaussian distribution associated. Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with node set $\mathcal{V}$ and edge set $\mathcal{E}$. A random vector $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top$ is said to satisfy the (undirected) Gaussian graphical model with graph $\mathcal{G}$, if $\dot{\mathbf{Y}}$ has a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with constant mean vector $\boldsymbol{\mu}$ and a positive definite $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, with density

$$f_{\dot{\mathbf{Y}}}(\dot{\mathbf{y}}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\{\frac{-1}{2}(\dot{\mathbf{y}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\dot{\mathbf{y}} - \boldsymbol{\mu})\}.$$

Since our primary interest is to estimate the graph structure, without loss of generality, we assume that the distribution has mean zero. However, there are some cases for which the mean of the model is not constant. Sometimes the variables of the graph are affected by some covariates. These models are known as **covariate adjusted Gaussian graphical models**, and the effect of covariates is present in the mean structure of the model. Formally, consider $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top$ as a q-dimensional vector of covariates. Conditionally on covariates $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, we suppose that $\dot{\mathbf{Y}} \mid \dot{\mathbf{x}} \sim \mathcal{N}_p(\boldsymbol{\Gamma}^\top \dot{\mathbf{x}}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Gamma}$ is the $q \times p$ matrix of regression coefficients. This model is broadly investigated in Chapter 4.

As the multivariate Gaussian distribution has a positive and continuous density, the pairwise and global Markov properties are equivalent for this graphical model. It is shown in Lauritzen (1996) that the entries of the inverse covariance matrix of the multivariate Gaussian distribution correspond to the edges of the graph. Formally,

$$(a, b) \notin \mathcal{E} \Leftrightarrow Y^a \perp Y^b \mid \dot{\mathbf{Y}}^{\mathcal{V}\setminus\{a,b\}} \Leftrightarrow \boldsymbol{\Sigma}^{-1}_{ab} = 0,$$

where $\boldsymbol{\Sigma}^{-1}_{ab}$ denotes the (a, b)-th entry of the inverse covariance matrix $\boldsymbol{\Sigma}$. This implies that the zero pattern of the inverse covariance matrix represents the structure of the graph. The inverse covariance matrix is referred to as the **precision** or **concentration** matrix and it is denoted by $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. The goal is to estimate the precision matrix $\boldsymbol{\Theta}$ in the setting where $p \gg n$, and then one can estimate the structure of the graph.

The most natural candidate for an estimator of the precision matrix is the inverse of the sample covariance matrix. However, when $p \gg n$, the sample covariance matrix is singular with probability one. To allow for consistent estimation in the setting $p \gg n$, a **sparsity** assumption will be imposed on the model. In general, if we have thousands of variables, the sparsity assumption expresses that one random variable is dependent on a few other random variables, out of the thousands of variables. For example, consider the health status of a person. The sparsity assumption considers that the health status can be predicted based on a few factors among several thousands biomarkers, which is more realistic than considering a model where all the biomarkers would be informative to predict the

state of health. For GGMs several works (see for example, Ravikumar et al., 2011; Jankova and van de Geer, 2015) impose sparsity assumptions on the maximal node degree, i.e., the number of non-zero entries per row of the precision matrix $\mathbf{\Theta}$. Imposing sparsity conditions on $\mathbf{\Theta}$ has proven to be a useful strategy. For example, suppose that d is the maximal node degree of the precision matrix $\mathbf{\Theta}$, which is defined as

$$d := \max_{a \in \mathcal{V}} \#\{b \in \mathcal{V}, b \neq a : \quad \mathbf{\Theta}_{ab} \neq 0\},$$

where $\#A$ is the cardinality of an arbitrary set A, and $\mathbf{\Theta}_{ab}$ is the (a,b)-th entry of $\mathbf{\Theta}$. Imposing sparsity assumptions on $\mathbf{\Theta}$ implies that the number of non-zero entries per row of the precision matrix $\mathbf{\Theta}$ cannot grow too fast with increasing p, and as a result, there should be lots of zero entries in the precision matrix. In general, there are two approaches for estimating $\mathbf{\Theta}$: local (nodewise) methods (Meinshausen and Bühlmann, 2006), and global methods (Yuan and Lin, 2007; Friedman et al., 2008). For local methods, the procedure focuses on estimating the neighborhood (or adjacency) set of each node in the underlying graphical model separately. For global methods, the procedure involves maximizing a global likelihood function which depends on $\mathbf{\Theta}$. In this thesis, we mostly focus on the second approach which is known as **graphical Lasso**, and is described below. For a short description on the method of nodewise Lasso, the reader can refer to Section 4.9.1.

1.1.3 Graphical Lasso

In this section, we review the problem of graph selection and the application of ℓ_1-regularized likelihood methods to address it. Suppose that we have n i.i.d. samples $\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n$ from $\dot{\mathbf{Y}}$, where $\dot{\mathbf{Y}} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, but the underlying graph structure is unknown. To estimate the graph structure using data, we use likelihood-based methods in conjunction with ℓ_1-regularization. In the Gaussian case, this approach leads to model selection based on a log-determinant convex program with ℓ_1-regularization. This problem is also known as (inverse) covariance selection. Consider the rescaled negative log-likelihood function (up to a constant) of the multivariate Gaussian distribution, which is of the form

$$\ell(\mathbf{\Theta}) = \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}), \tag{1.1.1}$$

where $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \dot{\mathbf{Y}}_i \dot{\mathbf{Y}}_i^\top$ is the sample (empirical) covariance matrix. The log-determinant function is defined on the space of positive definite matrices as

$$\log\det(\mathbf{\Theta}) = \begin{cases} \sum_{j=1}^{p} \log(\Lambda_j(\mathbf{\Theta})); & \text{if } \mathbf{\Theta} \succ \mathbf{0} \\ -\infty; & \text{otherwise,} \end{cases}$$

where $\Lambda_j(\mathbf{\Theta})$ is the j-th eigenvalue of $\mathbf{\Theta}$ and $\mathbf{\Theta} \succ \mathbf{0}$ denotes that $\mathbf{\Theta}$ is a positive definite matrix of dimension $p \times p$. The objective function (1.1.1) is an example

of a log-determinant program, which is strictly convex, so its minimum must be unique. Denote by $\hat{\mathbf{\Theta}}$ the optimizer of (1.1.1).

In classical statistics, it is known that $\hat{\mathbf{\Theta}}$ converges to the true value as the sample size n tends to infinity and p stays fixed. However, in high-dimensional settings, when $p \gg n$, the sample covariance matrix $\hat{\mathbf{\Sigma}}$ is not invertible, which implies that the estimator for $\mathbf{\Theta}$ obtained based on (1.1.1) does not exist. Hence, one must consider certain constraints or regularized forms of MLE estimation. Moreover, as explained in the previous sub-section, one may be interested in having a sparse estimator of the precision matrix, and hence easier to interpret. To consider a constraint which imposes sparsity, we add the ℓ_1 off-diagonal norm to the objective function (1.1.1), and the optimization problem turns to

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta} \succ \mathbf{0}} \{\text{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda \|\mathbf{\Theta}\|_{1,\text{off}}\}, \tag{1.1.2}$$

where $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal norm of the matrix, defined as $\|\mathbf{\Theta}\|_{1,\text{off}} = \sum_{a \neq b} |\mathbf{\Theta}_{ab}|$, and $\lambda > 0$ is a regularization (penalty) parameter. Problem (1.1.2) is an example of a penalized log-determinant program, it is convex and its solution is referred to as the graphical Lasso.

As the absolute value function is not differentiable in 0, the partial derivatives route for solving (1.1.2) is not applicable. An alternative method to solve (1.1.2), is to use the sub-differential equation, also known as the Karush-Kuhn-Tucker (KKT) condition, which is of the form

$$\hat{\mathbf{\Sigma}} - \mathbf{\Theta}^{-1} + \lambda \mathbf{\Psi} = 0, \tag{1.1.3}$$

with the symmetric matrix $\mathbf{\Psi}$ having diagonal elements equal to zero, and for $a \neq b$, $\mathbf{\Psi}_{ab} = \text{sign}(\mathbf{\Theta}_{ab})$ if $\mathbf{\Theta}_{ab} \neq 0$, and $\mathbf{\Psi}_{ab} \in [-1, 1]$ if $\mathbf{\Theta}_{ab} = 0$. This equation is nothing else than the first derivative of (1.1.2), where the first derivative of the absolute value function is replaced by its sub-differential. The most natural method to solve (1.1.3), is the first-order block coordinate-descent approach, introduced by d'Aspremont et al. (2008), and refined by Friedman et al. (2008). For more details on the algorithm, the reader can refer to Friedman et al. (2008). Due to the ℓ_1 penalty which is added to problem (1.1.2), the graphical Lasso estimator $\hat{\mathbf{\Theta}}$ is biased. To visualize this bias with a toy example, consider a random sample of size $n = 100000$ from a p-dimensional Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with $p = 4$. The adjacency matrix and the corresponding graph structure are shown in Figure 1.2. It is observed that elements $\mathbf{\Theta}_{12}$ and $\mathbf{\Theta}_{13}$ have non-zero values, and the other off-diagonal entries are zero. To estimate $\mathbf{\Theta}$ by the graphical Lasso method, we set the regularization parameter to $\lambda = \sqrt{\log(p)/n}$, which is of the same order as the one introduced in Ravikumar et al. (2011), when providing statistical guarantees. The histogram of $\hat{\mathbf{\Theta}}_{ab} - \mathbf{\Theta}_{ab}$ over $R = 1000$ different repetitions is presented in Figure 1.3 for $1 \leq a \leq b \leq 4$. The blue vertical lines are drawn at the empirical mean values over 1000 repetitions, and the red lines are drawn at zero. If there is no persisting bias, the blue line should coincide, or be close, to the red line. It is

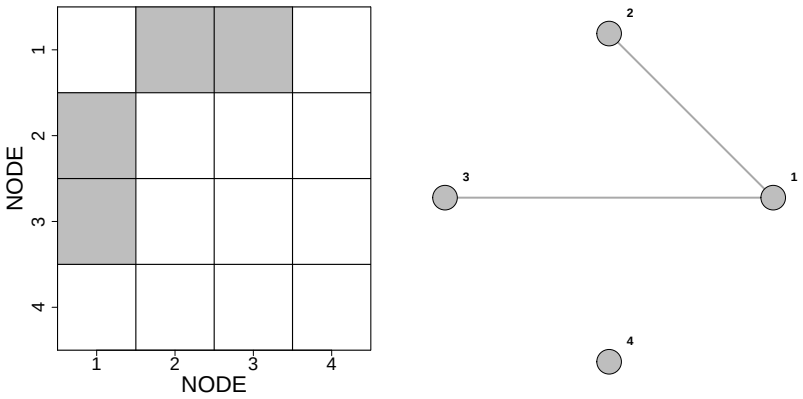


Figure 1.2: Visual representation of the adjacency matrix (on the left) and its graph structure (on the right) with $p = 4$ nodes.

observed that for some entries there are large differences between blue and red lines, which means that the empirical mean of $\hat{\mathbf{\Theta}}_{ab}$ is not close to $\mathbf{\Theta}_{ab}$, and so there is some bias associated with this estimation procedure, due to the introduction of the penalty. In the next sub-section, a debiased graphical Lasso estimator is introduced to remove the bias associated with the penalty term.

1.1.4 Debiased graphical Lasso

To modify the graphical Lasso estimator and remove the bias term associated with the penalty, Jankova and van de Geer (2015) proposed an estimator by inverting the KKT condition (1.1.3). This estimator is of the form

$$\hat{\mathbf{\Theta}}^d = 2\hat{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}, \tag{1.1.4}$$

with $\hat{\mathbf{\Theta}}$ the estimator obtained based on (1.1.2). The debiased estimator is also known as the **desparsified** graphical Lasso estimator, as the estimated matrix is not sparse anymore. Jankova and van de Geer (2015) showed the pointwise asymptotic normality of this estimator under certain conditions. Formally, for every $(a, b) \in \mathcal{V} \times \mathcal{V}$,

$$\sqrt{n}(\hat{\mathbf{\Theta}}^d_{ab} - \mathbf{\Theta}_{ab})/\sigma_{ab} = \mathbf{Z}_{ab} + o_p(1), \tag{1.1.5}$$

where $\hat{\mathbf{\Theta}}^d_{ab}$ is the (a, b)-th entry of $\hat{\mathbf{\Theta}}^d$, and $\mathbf{Z}_{ab}$ converges in distribution to $\mathcal{N}(0, 1)$. To show a visualization of the asymptotic unbiasedness of this estimator, we considered the same toy example as in the previous sub-section, and we computed the debiased estimates $\hat{\mathbf{\Theta}}^d_{ab}$ for $1 \leq a \leq b \leq 4$. Figure 1.4 presents the histogram of

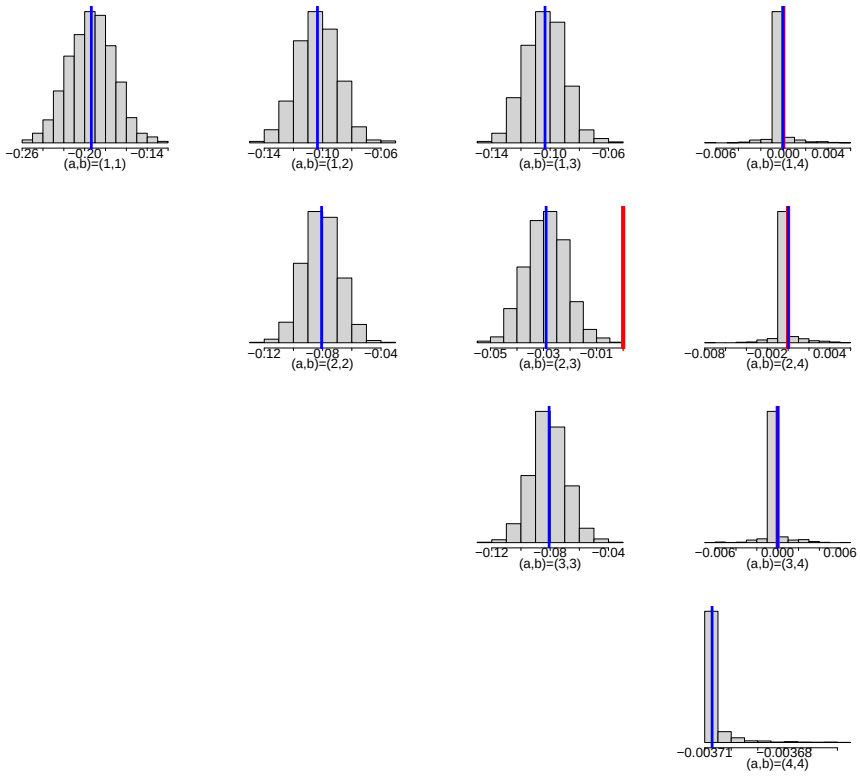


Figure 1.3: Histogram of $\hat{\boldsymbol{\Theta}}_{ab} - \boldsymbol{\Theta}_{ab}$ corresponding to the pairs (a, b), $1 \leq a \leq b \leq 4$, when $n = 100000$ and $p = 4$ over $R = 1000$ repetitions.

$\hat{\boldsymbol{\Theta}}^d_{ab} - \boldsymbol{\Theta}_{ab}$ over $R = 1000$ repetitions for $1 \leq a \leq b \leq 4$. Comparing it to Figure 1.3, it is observed that the bias drastically decreased, especially for the non-active components (a pair (a, b) is called non-active if $\boldsymbol{\Theta}_{ab} = 0$, and it is called active if $\boldsymbol{\Theta}_{ab} \neq 0$) like $\boldsymbol{\Theta}_{14}$ and $\boldsymbol{\Theta}_{24}$, for which the empirical bias is quite close to zero.

The asymptotic normality can be used for testing and constructing confidence intervals within the classical framework. In this thesis, this asymptotic normality is constructed in view of the distributed setting. Before describing the distributed procedure, we provide a short description of the preliminaries of hypothesis testing procedures, which are needed in this thesis, especially in Chapter 3.

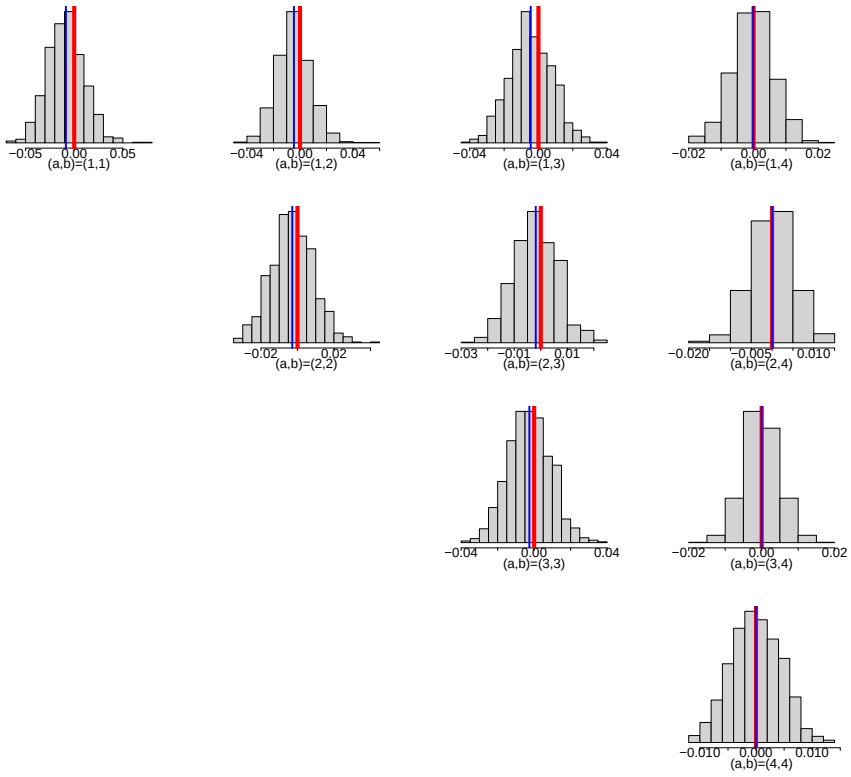


Figure 1.4: Histograms of $\hat{\mathbf{\Theta}}^d_{ab} - \mathbf{\Theta}_{ab}$ corresponding to the pairs $1 \leq a \leq b \leq 4$, when $n = 100000$ and $p = 4$ over $R = 10000$ repetitions.

1.2 Hypothesis testing

Hypothesis testing is a statistical technique used to make inferences and draw conclusions about a population based on a sample of data. It involves formulating two competing hypotheses, denoted as the null hypothesis (H_0) and the alternative hypothesis (H_1). These hypotheses are constructed based on the research question. For example, consider the problem of testing whether $\mathbf{\Theta}_{ab}$ is equal to zero or not. The simple null hypothesis represents the default assumption (in this example, $\mathbf{\Theta}_{ab} = 0$). It states that the partial correlation between variables corresponding to nodes a and b is zero. The alternative hypothesis, represents the proposition that contradicts the null hypothesis (in this example, $\mathbf{\Theta}_{ab} \neq 0$). It states that the partial correlation between variables corresponding to the nodes a and b is not

zero. Given the sample, one must choose between H_0 or H_1. This decision is built based on a critical region.

A subset of the sample space for which H_0 will be rejected is called **rejection region** or **critical region**. The complement of the rejection region is called the **acceptance region**. Typically, the critical region is specified in terms of a **test statistic** $T(\dot{\mathbf{Y}}_1, \dot{\mathbf{Y}}_2, \ldots, \dot{\mathbf{Y}}_n)$, a function of the sample. For example, in the precision matrix estimation, the test statistic can be constructed using the debiased estimator $\hat{\mathbf{\Theta}}^d$, which is a function of the sample data. The procedure for constructing the test statistic for the hypothesis test $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ versus $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$ using the debiased estimator is further explained in Chapter 3. Making a statistical decision always involves uncertainties, and thus making errors. There are two types of errors in statistical testing that are of interest, which are briefly explained below.

Definition 1.2.1. Type I error, also known as **false positive**, is the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis when the results can be attributed to the null.

Type I error occurs when one is observing a difference, while there is no difference in reality. The probability of making a type I error for a test with rejection region $R(\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n)$ is $\mathbb{P}(R(\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n) \mid H_0 \text{ is true})$, called **significance level**, and denoted as α.

Definition 1.2.2. Type II error, also known as **false negative**, is the error of not rejecting a null hypothesis when the alternative hypothesis is true. In other words, this is the error of failing to conclude there is a difference when there actually is one.

The probability of making a type II error for a test with rejection region $R(\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n)$ is $1 - \mathbb{P}(R(\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n) \mid \text{for a given alternative in } H_1)$, and is denoted as β. The **power** of the test is also defined as
$\mathbb{P}(R(\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n) \mid \text{for a given alternative in } H_1)$.

1.2.1 Multiple hypothesis testing

The multiple hypothesis testing problem occurs when a number of individual hypothesis tests are considered simultaneously to draw one single conclusion. For example, in graphical models, suppose that we would like to test simultaneous hypotheses $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ versus $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$ for all $1 \leq a < b \leq p$ in order to estimate the entire structure of the graph. In this case, if the significance level for a given test is α, the Family-Wise Error Rate (FWER) will increase as the number of tests increases. More precisely, assuming all the tests are independent, if M tests are performed, the FWER will be given by $1 - (1 - \alpha)^M \approx M\alpha$. Thus, in order to retain the same overall rate of false positives in a series of multiple tests, the error

rate for each individual test should be smaller. Intuitively, reducing the significance level for each test by the number of tests will result in an overall α which does not exceed the desired limit. For example, to obtain the overall α of 0.05 with ten tests, one requires the significance level $0.05/10 = 0.005$ for each test. This technique, known as **Bonferroni correction**, was first introduced by Holm (1979). He presented a modified version of the procedure called the Holm-Bonferroni method, which provides a more powerful correction for controlling the error rate in multiple hypothesis testing. Other multiple testing correction procedures could be also used.

However, in large-scale multiple testing, the traditional Bonferroni correction is too conservative. For example, for testing $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$, $1 \leq a < b \leq p$ with $p = 100$, one has $p(p-1)/2 = 4950$ tests, and by considering $\alpha = 0.05$, the significance level for each test will be $0.05/4950 = 1.010101 \times 10^{-5}$, which is too conservative. The test would make many false negative errors, resulting in a low statistical power. In order to be able to identify as many significant comparisons as possible while still maintaining a low false positive rate, the False Discovery Rate (FDR) control is proposed. In the next sub-section, an introduction to the false discovery rate is provided, as it is needed in Chapter 3.

1.2.2 False discovery rate

For large-scale multiple testing one can control the false discovery rate instead of imposing a Bonferroni correction. The false discovery rate, introduced by Benjamini and Hochberg (1995), is the **expected proportion** of type I errors. In testing $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$, $1 \leq a < b \leq p$, consider $\hat{S}$ as the set of selected pairs (a, b) by a procedure. Formally,

$$\text{FDR} = \mathbb{E}\left[\frac{\#\{(a,b) \in \hat{S} : \quad \mathbf{\Theta}_{ab} = 0\}}{\max\{\hat{s}, 1\}}\right], \tag{1.2.1}$$

where $\hat{s}$ is the cardinality of the set $\hat{S}$. Equation (1.2.1) implies that the FDR is the expected proportion of false discoveries among all discoveries, where a false discovery occurs when $\mathbf{\Theta}_{ab}$ is truly zero, but we estimate it as non-zero. The FDR control is a less conservative approach than the traditional Bonferroni correction. Suppose that there are 100 hypotheses to be tested. Then the FDR value of 0.05 means that at most 5% of "declared" positive results are expected to be false positives, while the Bonferroni-corrected testwise significance level is $0.05/100 = 0.0005$, which means that at most 5% of results among "all results" would be falsely positive. As such, the Bonferroni correction is more stringent as it significantly reduces the chance of making any false positives. The FDR control methods can typically be applied via p-value thresholds, such as Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). The BH procedure was shown to be able to control FDR in independent or positive dependent cases for regression models (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Afterwards, Barber and Candès

(2015) originally proposed the knockoff method to control FDR for linear models. This method constructs the knockoff variables as the covariates based on the original variables to identify the truly crucial variables. Candès et al. (2018) and Fan et al. (2019) further generalized it to high-dimensional nonlinear models. Another approach relevant for this thesis for controlling the FDR, is constructing the test statistics based on debiased estimators. These methods start from the regularized estimator, use different techniques to construct the debiased estimator, and then draw statistical inferences on the asymptotic normality. In this line of research, we can refer to Javanmard and Javadi (2019) and Cui et al. (2021), where the authors used debiased estimators in high-dimensional linear and generalized linear regression models to control FDR in hypothesis testing. A similar approach to control FDR in high-dimensional Gaussian graphical models is proposed in Chapter 3. In the next section, a short motivation is provided for the distributed procedure, which is an essential concept in this thesis.

1.3 Distributed settings

Consider Facebook's collection dataset which is over 200 terabytes of daily information of users (Vagata and Wilfong, 2017). This huge dataset includes text messages exchanged between users, user-uploaded photos, and the webpages visited by users. Machine learning algorithms use this dataset to predict different quantities such as determining offline friendships between users (Curtiss et al., 2013), identifying users in uploaded images (Taigman et al., 2014), and determining which advertisements are effective for different users (He et al., 2014). Developing algorithms that can work on such a large scale is a crucial problem in today's machine learning research.

Parallel algorithms are one of the solutions to these large scale problems, but designing them is difficult. Most of the parallel algorithms work well only for certain architectures. The simplest architecture is the shared memory model, which corresponds to a typical desktop computer with multiple cores. Cores can communicate with each other without much delay. However, sometimes it is expensive for processors to communicate with each other. In these cases, the **distributed algorithms** are the alternatives.

In distributed algorithms, the dataset is too large to fit in the memory of a single machine. So it is partitioned onto a cluster of machines connected by a network. Frameworks such as Message Passing Interface, known as MPI, (Clarke et al., 1994), MapReduce (Dean and Ghemawat, 2008) and Spark (Meng et al., 2016) have been developed to help managing the communication. In a typical distributed environment, all machines are located in the same physical location enjoying high speed network connections. As such communication takes on the order of milliseconds.

But some examples are more extreme. In some cases the dataset is not available at a central location. For example, a major web search engine has to save its data

on a range of platforms, storage units, and even geographical locations (Corbett et al., 2013); cellphones as processors have limited and unstable network connectivity, sometimes communication takes a long time, even hours or days; a supply chain company, a government, and hospitals all have tremendous amount of data, which are stored over different agencies or areas. In some cases, it is costly or prohibitive to transport all the data. For example, for hospital datasets, consider the information on a rare disease, which is available only for some particular hospitals. Only a small group of researchers can study that rare disease. Due to legal restrictions related to ethics, privacy and data protection, this rare disease information cannot be shared with others. To tackle the above problems, **federated learning** (McMahan et al., 2017) comes into play. In this method, sensitive datasets held by clients, like hospitals or cellphones, do not leave their own boundaries. Instead, each client trains using its own data, and only the updates are shared. Formally, in a federated learning strategy, instead of transferring data from local processors, summary of datasets (updates) are transferred. For example, consider K different locations with independent datasets. One common approach is to first obtain an estimator at each local station $\hat{\theta}_k$, $k = 1, \ldots, K$, and then transfer the local estimates to the central station for aggregation. An interesting challenge is now how to aggregate these statistics into one final summarizing statistic.

In distributed problems, local estimators can be aggregated using two primary methods: the "simple average" and the "one-step estimator". The latter method combines the simple averaging estimator with Newton's method, generating a one-step estimator. Various scenarios have been investigated, including sparse high-dimensional estimation in linear regression, M-estimators, Bayesian frameworks, situations where the number of machines increases as the sample size increases, and many more. A comprehensive study of distributed statistical inference in these settings was conducted by Jordan et al. (2018). Additionally, Chen and Xie (2014), Lee et al. (2017) and Battey et al. (2018) focused on high-dimensional, sparse parameter vector estimation using the penalized M-estimators. The one-step method has been known for some time and has the advantage of making a consistent estimator as efficient as the maximum likelihood estimators or M-estimators with just a single Newton-Raphson iteration. Efficiency here refers to the relative efficiency converging to one. For further details on the general one-step approach, we refer to van der Vaart (2000). Examples of its application can be found in Bickel (1975), Fan and Chen (1999), and Zou and Li (2008) among many others.

1.4 Outline of the thesis

The objective of this thesis is to explore estimation and inference for Gaussian graphical models when the dataset is distributed across several locations. As described in the first part of this chapter, distributed algorithms are used when aggregating all the data in one location is not feasible due to privacy concerns,

and limited capacity of machines. This setup presents a natural problem where datasets from different locations may have different sizes, resulting in an unbalanced distributed setting with different sub-sample sizes. In this context, a key question is:

Can one devise an algorithm for estimation and inference in Gaussian graphical models under an unbalanced distributed setting that ensures the proposed aggregated solution remains consistent and asymptotically efficient, similar to the centralized procedure that would be obtained if all the data were available at a central location?

To address this question, we explore different scenarios. First, we consider a zero-mean Gaussian graphical model and develop the estimation and inferential tools within this family under the unbalanced distributed setting. Afterwards, we delve into the performance of the proposed procedure by investigating the false discovery rate under this setting. Consequently, we develop a procedure to control the false discovery rate for the hypothesis tests related to the entries of the precision matrix. Finally, to broaden the scope of our results, we extend our study to a non-zero Gaussian graphical model, where the mean of the model mean is affected by some covariates. We thoroughly investigate the estimation and inference procedures within this extended setting as well. As such, the outline of this thesis is as follows.

The second chapter of this thesis delves into the estimation of Gaussian graphical models in an unbalanced distributed framework. This chapter focuses on providing an effective approach when dealing with heterogeneous machines or datasets from different sources with different sizes, preventing aggregation onto a single machine. The primary goal is to propose a novel aggregated estimator for the precision matrix, substantiated by both theoretical and practical arguments. This estimator is built based on the asymptotic distribution of the debiased estimators using the sub-samples. Theoretical justifications are presented, outlining the limit distribution and convergence rate of this estimator, especially under sparsity conditions concerning the true precision matrix while controlling for the number of machines involved. Additionally, a procedure for conducting statistical inference is proposed. This chapter also includes a comprehensive simulation study and a real-world data example to demonstrate the distributed estimator's performance, which is found to be comparable to that of the non-distributed estimator utilizing the full dataset.

The third chapter of this thesis introduces a novel multiple testing procedure designed to address the issue of conditional dependence in Gaussian graphical models. The aim of this study is to determine whether the conditional dependence between variables is positive or negative. This chapter proposes the construction of different test statistics using debiased and distributed estimators within a multiple testing framework. We demonstrate that the proposed procedure can effectively control the directional false discovery rate asymptotically at a pre-defined level.

This particular false discovery rate focuses on the sign of the estimation. Additionally, this chapter highlights that, with mild conditions imposed on the non-zero entries of the precision matrix, an asymptotic power of one can be achieved. The theoretical results are further supported and validated through various simulation scenarios and real data examples, which enable a comprehensive investigation of the performance of the proposed procedure.

The fourth chapter of this thesis presents a novel distributed estimation and inferential framework tailored for addressing sparse multivariate regression and conditional Gaussian graphical models within the unbalanced splitting setting. In this chapter, the focus is on scenarios where the number of covariates, responses, and machines grow with the sample size, while still imposing sparsity constraints. To handle this, we first propose debiased estimators for both the coefficient matrix and the precision matrix. Then, to construct the final aggregated estimator, we use the same technique as in Chapter 2, which is based on the asymptotic distribution of the debiased estimators using the sub-samples. Theoretical guarantees for the final estimators are provided. As the number of machines increases with the sample size, these aggregated estimators demonstrate efficiency and asymptotic normality. This indicates their potential to deliver robust and accurate results in such distributed settings.

The fifth chapter of this thesis presents the implementation of the procedures developed in Chapters 2-4 in `R` with our own package **DistributedGGL**. This chapter provides documentation and illustration of the proposed methods in `R` along with some examples.

The sixth chapter of this thesis summarizes the achievements and limitations of this work. It also proposes several extensions and research perspectives, which close the thesis.

1.5 Preliminary notation

For the reader's convenience, this section presents some notation that is used throughout this thesis.

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of dimensions $m \times n$ and $p \times q$, we denote $\mathbf{A} \otimes \mathbf{B}$ as the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ which is defined as a $pm \times qn$ block matrix with $\mathbf{A}_{ab}\mathbf{B}$ for the block (a, b), where $\mathbf{A}_{ab}$ is the (a, b)-th element of matrix $\mathbf{A}$, $a = 1, \ldots, m$ and $b = 1, \ldots, n$. For two sequences $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$, $b_n = O(a_n)$ if there exist positive numbers M_0 and N_0 such that $|\frac{b_n}{a_n}| \leq M_0$ for all $n \geq N_0$. We write $b_n \asymp a_n$ if both $b_n = O(a_n)$ and $a_n = O(b_n)$ hold. Similarly, for a random sequence $\{Y_n; n \geq 1\}$, we write $Y_n = O_p(a_n)$ if for every $\epsilon > 0$, there exist finite numbers $M_0 > 0$ and $N_0 > 0$ such that $\mathbb{P}(|\frac{Y_n}{a_n}| > M_0) < \epsilon$ for all $n \geq N_0$. Furthermore, $b_n = o(a_n)$ if $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$. In the case of a random sequence $\{Y_n; n \geq 1\}$, we write $Y_n = o_p(a_n)$ if $\frac{Y_n}{a_n} \xrightarrow{p} 0$, as $n \to \infty$, where the notation $\xrightarrow{p}$ denotes convergence in probability. For a matrix $\mathbf{A}$, we use the notation

$\|\mathbf{A}\|_\infty = \max_a \sum_b |\mathbf{A}_{ab}|$ and $\|\mathbf{A}\|_\infty = \max_{a,b} |\mathbf{A}_{ab}|$ for the matrix and elementwise ℓ_∞ norms, respectively. The same symbol $\|\mathbf{x}\|_\infty = \max_b |\mathbf{x}_b|$ is used for the ℓ_∞ norm of a vector $\mathbf{x}$. Moreover, $\|\mathbf{A}\|_1 = \sum_{a,b} |\mathbf{A}_{ab}|$ and $\|\mathbf{x}\|_1 = \sum_b |\mathbf{x}_b|$ are used for the ℓ_1 norm of a matrix $\mathbf{A}$ and of a vector $\mathbf{x}$, respectively. Finally, we use $\|\mathbf{A}\|_F = \sqrt{\sum_{a,b} \mathbf{A}_{ab}^2} = \sqrt{\text{trace}(\mathbf{A}\mathbf{A}^\top)}$ for the Frobenius norm of a matrix and $\|\mathbf{x}\|_2 = \sqrt{\sum_b \mathbf{x}_b^2}$ for the ℓ_2 norm of a vector $\mathbf{x}$.

CHAPTER 2

Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models

This chapter studies the estimation of Gaussian graphical models in the unbalanced distributed framework. It provides an effective approach when the available machines are of different powers or when the existing dataset comes from different sources with different sizes and cannot be aggregated on one single machine. In this chapter, we propose a new aggregated estimator of the precision matrix and justify such an approach by both theoretical and practical arguments. The limit distribution and convergence rate for this estimator are provided under sparsity conditions on the true precision matrix and controlling for the number of machines. Furthermore, a procedure for performing statistical inference is proposed. On the practical side, using a simulation study and a real data example, we show that the performance of the distributed estimator is similar to that of the non-distributed estimator that uses the full data. The content of this chapter is based on the paper

> Nezakati, E. and E. Pircalabelu (2023). Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models. *Statistics and Computing* 33, 1–14.

2.1 Introduction

Precision matrix estimation plays an important role in statistical and machine learning, especially in the framework of probabilistic graphical modeling. A large body of literature studied the estimation problem of the precision matrix for Gaussian graphical models. We refer to Meinshausen and Bühlmann (2006), Friedman et al. (2008), Cai et al. (2011), Wang (2014) and Wang et al. (2016) among many others for a treatment on the subject.

Many datasets nowadays may be too large to fit and be read efficiently into a single ordinary machine. Moreover, sometimes datasets from different sources are private and due to security and privacy concerns, one is not allowed to aggregate all the data at one location. For instance, in Federated machine learning (McMahan et al., 2017), different datasets with different sizes are used for training across multiple local machines without any exchanging and sharing. As such, estimation of the precision matrix via decentralized, unbalanced machines is of contemporary interest. Much attention in distributed estimation has focused on the setting where the sample size n is large and most approaches propose splitting the observations into K independent sub-samples that are analyzed in parallel. Once the analysis is performed, estimates are combined together into a final estimate that is treated as the final output of the method. In this chapter we focus on the case where the number of variables p can grow with n, the total sample size, such that $\log p/n_k \to 0$, where n_k is the sample size at the level of the k-th machine with $k = 1, \ldots, K$. As a consequence, a sparsity assumption is imposed on the true precision matrix as a function of sub-sample sizes n_k and p.

A wide range of literature studied combination methods for parallel estimators in the context of linear, generalized linear models, kernel and Bayesian estimators, see for instance Zhang et al. (2015), Lee et al. (2017), Battey et al. (2018), Xu et al. (2019) and Xue and Liang (2019) among many others. Regarding the estimation of the precision matrix, there is limited literature using parallel analysis. Arroyo and Hou (2016) studied the problem of estimating the precision matrix for Gaussian graphical models from a set of K balanced distributed sub-samples via a simple average method. They performed an additional local thresholding step on each machine, in order to obtain a sparse estimator. Wang and Cui (2021) proposed a distributed estimator of the sparse precision matrix in Transelliptical graphical models by debiasing a D-trace Lasso-type estimator and then by applying a hard threshold on the aggregated estimator which is obtained by simple averaging. Nevertheless, in some recent approaches like Federated learning, when some of the available machines are more powerful than others, it is not efficient to distribute a dataset on different machines with equal sizes. Instead, one could place larger datasets on the powerful machines and smaller datasets on the others. In this case, one deals with unbalanced sub-samples and just taking a simple average is not an optimal approach for aggregating estimators.

Meta-analysis is also a known method for combining summary statistics from independent studies. Xie et al. (2011) and Liu et al. (2015) developed general meta-analysis frameworks using confidence distributions to combine multiple heterogeneous estimators derived from individual studies. Since splitting a dataset incurs some efficiency loss, deriving an upper bound on the number of machines is of interest to illustrate the efficiency of the model. Recently, Tang et al. (2020) developed a strategy using an aggregating method based on confidence distributions to combine debiased Lasso-type estimators in generalized linear models. In their framework, the number of machines can also diverge with n. They showed that as

K increases at the rate of $O(n^{1/2-\epsilon})$, $\epsilon \in (0, 1/2]$, the combined estimator achieves the same estimation efficiency as the centralized maximum likelihood estimator. In this chapter, a pseudo likelihood function is constructed based on the asymptotic distribution of the debiased estimators to aggregate the unbalanced estimators of the precision matrix for Gaussian graphical models. An upper bound on K is derived to guarantee the consistency and asymptotic normality of the estimator. It is shown that this upper bound is of order $O(n^{1/3}/(d\log(p)))$, which is a function of the total sample size, the sparsity d and the number of variables, p. See Section 2.2 in this chapter for a definition of d.

The rest of the chapter is organized as follows. In Section 2.2, notation and preliminaries are introduced. A methodology for performing distributed estimation based on sub-samples is presented in Section 2.3. We continue in Section 2.4 by proposing a final estimator that pools together separate estimators. Theoretical properties of this estimator are also investigated in this section. In Section 2.5, the performance of the estimator is evaluated at the hand of a controlled simulation study and in Section 2.6, the performance on a real data set is illustrated. We close with a discussion on the method and possible extensions in Section 2.7. Proofs of the supporting lemmas and theorems, and more simulation results can be found in the Appendix, at the end of the chapter.

2.2 Preliminaries

Consider a zero-mean random variable Y. This random variable is called sub-Gaussian if there exists a constant γ such that $\mathbb{E}(\exp(tY)) \leq \exp(\gamma^2 t^2/2)$ for all $t \in \mathbb{R}$. The bound on the moment generating function implies the tail bound $\mathbb{P}(|Y| > y) \leq 2\exp(-y^2/2\gamma^2)$ for all $y > 0$. Similarly, a zero-mean random vector $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top$ with covariance matrix $\boldsymbol{\Sigma}$ is sub-Gaussian if all normalized components $Y^a/\sqrt{\boldsymbol{\Sigma}_{aa}}, a = 1, \ldots, p$, are sub-Gaussian random variables with a common parameter $\gamma > 0$, where by $\boldsymbol{\Sigma}_{aa}$ we denote the a-th diagonal element of $\boldsymbol{\Sigma}$. The model in this thesis is defined under the Gaussianity assumption, which is a special case of sub-Gaussianity. In fact, a multivariate zero-mean Gaussian vector is a prime example of a sub-Gaussian random vector. Consider the Gaussian graphical model, where a p-dimensional random vector $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top$ follows a multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with mean vector zero and covariance matrix $\boldsymbol{\Sigma}$. The components of $\dot{\mathbf{Y}}$ correspond to the vertex set $\mathcal{V} = \{1, \ldots, p\}$ of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the edge set $\mathcal{E}$ describes the conditional dependence between every pair of components $Y^1, \ldots, Y^p$. A pair (a, b) is included in the edge set $\mathcal{E}$ if and only if the variables Y^a and Y^b are dependent given all remaining variables. Under the multivariate Gaussian distribution, a pair of variables is conditionally independent given all remaining variables if and only if the corresponding entry in the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is zero. Elements of the precision matrix may thus be interpreted as edge weights. Denote the maximum

number of non-zero (active) off-diagonal entries of $\boldsymbol{\Theta}$ per row, i.e., the maximal node degree, with d and the index set of non-zero off-diagonal entries of $\boldsymbol{\Theta}$ with S, formally, $S = \{(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b : \quad \boldsymbol{\Theta}_{ab} \neq 0\}$, having cardinality s. The set of non-active components is denoted by S^c.

2.3 Distributed penalized estimation

In this section, we propose distributed estimators of the precision matrix $\boldsymbol{\Theta}$ using only sub-sample information. First, denote the Hessian of the negative log-likelihood function $\ell(\boldsymbol{\Theta}) \propto \text{trace}(\hat{\boldsymbol{\Sigma}}_k \boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta})$, where $\hat{\boldsymbol{\Sigma}}_k = \mathbf{Y}_k^\top \mathbf{Y}_k / n_k$ is the sample covariance based on the k-th sub-sample (properly defined later in the section), evaluated at the true precision matrix $\boldsymbol{\Theta}$ by $\mathbf{H}$, which can be shown to be $\mathbf{H} = \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}$ (see Ravikumar et al., 2011). By definition, $\mathbf{H}$ is a $p^2 \times p^2$ matrix indexed by the pair of elements from the vertex set, such that $\mathbf{H} = [\mathbf{H}_{(a,b),(c,d)}]$, where $(a,b),(c,d) \in \mathcal{V} \times \mathcal{V}$. Ravikumar et al. (2011) showed that in the case of the multivariate Gaussian distribution, $\mathbf{H}_{(a,b),(c,d)}$ is simplified to $\mathbf{H}_{(a,b),(c,d)} = \mathbb{C}\text{ov}(Y^a Y^b, Y^c Y^d)$. Let $\mathbb{S} = \{S \cup \{(1,1),(2,2),\ldots,(p,p)\}\}$ with cardinality ϖ, which is equal to $\varpi = s + p$, and its complement set with $\mathbb{S}^c$. The following regularity assumptions (see for example, Ravikumar et al., 2011; Jankova and van de Geer, 2015) are considered for the theoretical guarantees when estimating $\boldsymbol{\Theta}$:

> (A1) The irrepresentability condition holds for the true precision matrix $\boldsymbol{\Theta}$, i.e., there exists $\alpha_1 \in (0,1]$ such that $\max_{e \in \mathbb{S}^c} \|\mathbf{H}_{e\mathbb{S}} (\mathbf{H}_{\mathbb{S}\mathbb{S}})^{-1}\|_1 \leq 1 - \alpha_1$, where $\mathbf{H}_{\mathbb{S}\mathbb{S}} \in \mathbb{R}^{\varpi \times \varpi}$ is a sub-matrix of $\mathbf{H}$ whose rows and columns are indexed by the elements of $\mathbb{S}$. Moreover, e is a pair $(a,b) \in \mathbb{S}^c$ such that $\mathbf{H}_{eS}$ is an ϖ-dimensional column vector with elements $\mathbf{H}_{e,(c,d)}$, where $(c,d) \in \mathbb{S}$.

Assumption (A1) is necessary and sufficient for the graphical Lasso estimator to exhibit model selection consistency (Tropp, 2006; Zhao and Yu, 2006). In fact, it limits the influence that the non-edge based variables (i.e., the pairs in $\hat{S}^c$) can have on the edge based variables (i.e., the pairs in $\hat{S}$). For example, for the multivariate Gaussian distribution, consider the zero-mean edge random variables

$$X_{(a,b)} := Y^a Y^b - \mathbb{E}(Y^a Y^b),$$

for all $a,b \in \{1,2,\ldots,p\}$. Then, $\mathbf{H}_{(a,b),(c,d)} = \mathbb{E}(X_{(a,b)} X_{(c,d)})$. Defining $X_S := \{X_{(a,b)}; \quad (a,b) \in S\}$, the irrepresentability condition is of the form $\max_{e \in \mathbb{S}^c} \|\mathbb{E}(X_e X_S^\top)\mathbb{E}(X_S X_S^\top)^{-1}\|_1 \leq 1 - \alpha_1$, which forces the requirement that the edge based variables (i.e., $X_{(a,b)}$ that $(a,b) \in S$) should not be highly correlated with the non-edge based variables (i.e., $X_{(a,b)}$ that $(a,b) \in S^c$). The reader can refer to Ravikumar et al. (2011) for more details.

(A2) The eigenvalues of the precision matrix $\boldsymbol{\Theta}$ are bounded from above and below, i.e., there exists a constant $\Lambda \geq 1$ such that $\frac{1}{\Lambda} \leq \Lambda_{\min}(\boldsymbol{\Theta}) \leq \Lambda_{\max}(\boldsymbol{\Theta}) \leq \Lambda$, where $\Lambda_{\min}(\boldsymbol{\Theta})$ and $\Lambda_{\max}(\boldsymbol{\Theta})$ are the minimum and maximum eigenvalues of $\boldsymbol{\Theta}$, respectively.

Assumption (A2) is needed to control the convergence rate of the debiased estimator via controlling the magnitude of the elements of $\boldsymbol{\Theta}$, as well as to insure $\boldsymbol{\Theta}$ is invertible.

Consider now an i.i.d. sample of size n from $\dot{\mathbf{Y}}$ and suppose that it is divided randomly into K non-overlapping sub-samples each with size n_k for the k-th sub-sample, $k = 1, \ldots, K$, such that $n_k/n \to c_k \in (0,1)$ as $n_k \to \infty$, and $\lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1$. Denote $n_\dagger = \min_{1\leq k\leq K} n_k$. Consider $\dot{\mathbf{Y}}_{l,k} = (Y_{l,k}^1, \ldots, Y_{l,k}^p)^\top$, $l = 1, \ldots, n_k$, and denote the k-th sub-sample in matrix form as $\mathbf{Y}_k = [\dot{\mathbf{Y}}_{1,k}, \dot{\mathbf{Y}}_{2,k}, \ldots, \dot{\mathbf{Y}}_{n_k,k}]^\top$, which is of dimension $n_k \times p$ and obtained by appending column vectors $\dot{\mathbf{Y}}_{1,k}, \ldots, \dot{\mathbf{Y}}_{n_k,k}$ one after another and then transposing. By splitting the sample data into K disjoint sub-samples, one can analyze each sub-sample per machine in parallel. The goal is to estimate the structure of the graph $\mathcal{G}$, or equivalently, find the zero pattern of $\boldsymbol{\Theta}$, by combining K estimators, each obtained using a particular sub-sample. The usual assumption in this context is that the underlying graph is sparse, which means that a certain bound is imposed on the maximum node degree of the precision matrix. Due to the relation between the graph edges and the entries of the precision matrix in Gaussian graphical models, this sparsity reflects that the related graph has a rather low number of edges. To impose this sparsity condition on the estimation, one common approach is to add an ℓ_1 penalty to the negative log-likelihood function. This kind of regularization effectively forces some of the elements of $\boldsymbol{\Theta}$ to zero, thus resulting in sparse solutions. There exists a wide variety of methods making use of ℓ_1 regularization, see for example Friedman et al. (2008), Hsieh et al. (2014), Cai et al. (2016), etc. For each sub-matrix $\mathbf{Y}_k$ ($k = 1, \ldots, K$), the graphical Lasso estimator $\hat{\boldsymbol{\Theta}}_k$, defined by Friedman et al. (2008), is the solution to the optimization problem

$$\hat{\boldsymbol{\Theta}}_k = \arg\min_{\boldsymbol{\Theta}\in S_{++}^p} \left\{ \operatorname{trace}(\hat{\boldsymbol{\Sigma}}_k \boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \lambda_k \|\boldsymbol{\Theta}\|_{1,\text{off}} \right\}, \tag{2.3.1}$$

where $\lambda_k > 0$ is the penalty term, $\hat{\boldsymbol{\Sigma}}_k = \mathbf{Y}_k^\top \mathbf{Y}_k / n_k$ is the sample covariance based on the k-th sub-sample, $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal penalty of the matrix defined as $\|\boldsymbol{\Theta}\|_{1,\text{off}} = \sum_{a\neq b} |\boldsymbol{\Theta}_{ab}|$, the quantity $\boldsymbol{\Theta}_{ab}$ is the (a,b)-th element of $\boldsymbol{\Theta}$ and S_{++}^p is the space of positive definite matrices of dimension $p \times p$. The graphical Lasso estimator (2.3.1) is biased due to the ℓ_1 penalty which is added to the negative log-likelihood function. This estimator satisfies the Karush-Kuhn-Tucker (KKT) condition (see for example, Ravikumar et al., 2011) as $\hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Theta}}_k^{-1} + \lambda_k \hat{\mathbf{D}}_k = \mathbf{0}$, where the matrix $\hat{\mathbf{D}}_k$ belongs to the sub-differential of the off-diagonal norm $\|\cdot\|_{1,\text{off}}$ evaluated at $\hat{\boldsymbol{\Theta}}_k$. Jankova and van de Geer (2015) proposed the debiased graphical

Lasso estimator to correct the bias, which for our setting can be constructed using the k-th sub-sample as

$$\hat{\mathbf{\Theta}}_k^d = 2\hat{\mathbf{\Theta}}_k - \hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_k\hat{\mathbf{\Theta}}_k. \tag{2.3.2}$$

This debiased estimator has the appealing property that each entry of the matrix is asymptotically normal and one can easily construct confidence intervals for the entries of $\mathbf{\Theta}$. It has been shown in Jankova and van de Geer (2015) that for every $(a,b) \in \mathcal{V} \times \mathcal{V}$,

$$\sqrt{n_k}(\hat{\mathbf{\Theta}}_{ab,k}^d - \mathbf{\Theta}_{ab})/\sigma_{ab} = \sqrt{n_k}\{\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}\}_{ab}/\sigma_{ab} + \sqrt{n_k}\mathbf{R}_{ab,k}/\sigma_{ab}, \tag{2.3.3}$$

where $\hat{\mathbf{\Theta}}_{ab,k}^d$ and $\mathbf{R}_{ab,k}$ are the (a,b)-th element of $\hat{\mathbf{\Theta}}_k^d$ and of a remainder term, called $\mathbf{R}_k$, respectively, and $\mathbf{W}_k = \hat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}$. Moreover, $\sigma_{ab}^2 = \mathbf{\Theta}_{aa}\mathbf{\Theta}_{bb} + \mathbf{\Theta}_{ab}^2$, under the multivariate Gaussian distribution. Theorem 1 of Jankova and van de Geer (2015) guarantees that under assumptions (A1) and (A2), with tuning parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$ and under the sparsity condition $d^{3/2} = o(\sqrt{n_k}/(C\log(p)))$ with

$$C = \max\left\{\frac{\kappa_{\mathbf{H}}}{\alpha_1^2}, \frac{\kappa_{\mathbf{H}}^2}{\alpha_1^{9/8}} n_k^{-1/4}(\log(p))^{1/8}, \frac{\max\{\kappa_{\mathbf{\Sigma}}\kappa_{\mathbf{H}}, \kappa_{\mathbf{\Sigma}}^3\kappa_{\mathbf{H}}^2\}^{3/2}}{\alpha_1^{3/2}}(n_k\log(p))^{-1/4}\right\},$$

the random sequence $\sqrt{n_k}\{\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}\}_{ab}/\sigma_{ab}$ converges weakly to $\mathcal{N}(0,1)$, where $\kappa_{\mathbf{\Sigma}} := |||\mathbf{\Sigma}|||_\infty$, $\kappa_{\mathbf{H}} := |||(\mathbf{H}_{SS})^{-1}|||_\infty$ and α_1 is defined in (A1). Moreover, the elementwise ℓ_∞ norm of $\mathbf{R}_k$ is of order

$$\|\mathbf{R}_k\|_\infty = O_p\left(\frac{1}{\alpha_1^2}\kappa_{\mathbf{H}}\max\{d^{3/2}\log(p)/n_k, \frac{1}{\alpha_1}\kappa_{\mathbf{H}}d^2(\log(p)/n_k)^{3/2}\}\right), \tag{2.3.4}$$

and by letting $1/\alpha_1 = O(1)$, $\kappa_{\mathbf{\Sigma}} = O(1)$, $\kappa_{\mathbf{H}} = O(1)$, under the mentioned sparsity condition, $\mathbf{R}_{ab,k}$ is $o_p(1/\sqrt{n_k})$. Furthermore, it follows that the convergence rate of the debiased estimator $\hat{\mathbf{\Theta}}_k^d$ is of order

$$\begin{aligned}\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty = O_p\Big(&\max\{d\sqrt{\log(p)/n_k}, 1/\alpha_1^2\kappa_{\mathbf{H}}d^{3/2}\log(p)/n_k,\\ &1/\alpha_1^3\kappa_{\mathbf{H}}^2 d^2(\log(p)/n_k)^{3/2}\}\Big),\end{aligned} \tag{2.3.5}$$

where under the above bounds on $\kappa_{\mathbf{\Sigma}}$, $\kappa_{\mathbf{H}}$ and $1/\alpha_1$, it simplifies to

$$\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty = O_p\Big(\max\{d\sqrt{\log(p)/n_k}, d^{3/2}\log(p)/n_k, d^2(\log(p)/n_k)^{3/2}\}\Big). \tag{2.3.6}$$

A consistent estimator for σ_{ab}^2 based on the k-th sub-sample can be also constructed using Lemma 2 of Jankova and van de Geer (2015) as $\hat{\sigma}_{ab,k}^2 = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}_{ab,k}^2$. This estimator will be further used in Section 2.4, where we leverage the asymptotic distribution of each $\hat{\mathbf{\Theta}}_k^d$ $(k = 1, \ldots, K)$ from (2.3.3) to construct the aggregated estimator using K estimators from the sub-samples.

2.4 Combined estimator across the sub-samples

The results in the previous section are provided at the level of the k-th sub-sample, $k = 1, \ldots, K$. To construct a more reliable estimator, we aggregate estimators drawn from different distributed locations. There are multiple ways to aggregate these K estimators into a combined estimator. In this chapter, due to the asymptotic normal distribution of the local estimators, we derive the final aggregated estimator by maximizing a pseudo log-likelihood function, which is constructed using the asymptotic normal density of local estimators constructed based on the sub-samples. Consider Δ as a general parameter of interest, for example $\boldsymbol{\Theta}_{ab}$ in this chapter, and $\hat{\Delta}_k$ as the k-th estimator based on the k-th sub-sample with variance σ^2 and denote its consistent estimator by $\hat{\sigma}_k^2$ which is also derived based on the k-th sub-sample. Consider the asymptotic normal density of $\hat{\Delta}_k$ at point ι_k as $\hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)$, where the variance σ^2 is replaced by its consistent estimator $\hat{\sigma}_k^2$. By maximizing the pseudo log-likelihood function constructed as

$$l(\Delta) = \log\Big(\prod_{k=1}^{K} \hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)\Big) \propto \sum_{k=1}^{K}(-n_k/2)(\iota_k - \Delta)^2/\hat{\sigma}_k^2,$$

with respect to Δ, the final aggregated estimator is of the form

$$\tilde{\Delta}_{\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2} \hat{\Delta}_k, \tag{2.4.1}$$

where the subscript "owAvg" stands for the optimally weighted average. We call this estimator an "optimally weighted average", as it is obtained using a maximization problem. As it is shown in Claeskens et al. (2016), when K is fixed, a convex combination of K estimators $\hat{\Delta}_1, \ldots, \hat{\Delta}_K$ is of the form $\tilde{\Delta}_c = \boldsymbol{w}^\top \hat{\boldsymbol{\Delta}}$, where $\hat{\boldsymbol{\Delta}} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_K)^\top$ and $\boldsymbol{w} = (w_1, \ldots, w_K)^\top$ is the vector of weights satisfying the constraint $\sum_{k=1}^{K} w_k = 1$. When the covariance between every two estimators is zero, the optimal weight for the k-th estimator in this convex combination is of the form $w_k = \frac{1/\sigma_k^2}{\sum_{k=1}^{K} 1/\sigma_k^2}$, where σ_k^2 is the variance of the k-th estimator. This result can be also encountered in portfolio theory, where the same weights are used to find the global minimum variance portfolio. The reader can refer to Kempf and Memmel (2006) and Golosnoy et al. (2022) among many other references for more details. In the distributed setting, when the sub-samples are equal (balanced case), our proposed estimator simplifies to the optimal weights convex combination, where we substitute the variance σ_k^2 with its consistent estimator $\hat{\sigma}_k^2$, as it is unknown in practice. Substituting the debiased estimator $\hat{\boldsymbol{\Theta}}_k^d$ in (2.4.1), one can aggregate distributed estimators of the precision matrix from the sub-samples into the final combined estimator $\tilde{\boldsymbol{\Theta}}_{\text{owAvg}}$. Note that if the variance σ^2 is known, then replacing $\hat{\sigma}_k^2$ by σ^2 simplifies the aggregated estimator $\tilde{\Delta}_{\text{owAvg}}$ to two special cases. In the

case of unbalanced sub-samples, $\tilde{\Delta}_{\text{owAvg}}$ is simplified to the sample size weighted average estimator

$$\tilde{\Delta}_{\text{wAvg}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\Delta}_k, \tag{2.4.2}$$

where "wAvg" stands for the weighted average proportional to the sub-sample sizes. Similarly to the optimally weighted average estimator, by substituting the debiased estimator $\hat{\mathbf{\Theta}}_k^d$ in (2.4.2), the weighted average aggregated estimator of $\mathbf{\Theta}$ can be constructed and is denoted by $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$. Moreover, if the sub-samples are also balanced (i.e., $n_1 = n_2 = \ldots = n_K$), $\tilde{\Delta}_{\text{owAvg}}$ further simplifies to the simple average estimator

$$\tilde{\Delta}_{\text{sAvg}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\Delta}_k, \tag{2.4.3}$$

where "sAvg" stands for the simple average. By the same argument as for the optimally weighted and sample size weighted averages, the simple average estimator of $\mathbf{\Theta}$ can be constructed and denoted by $\tilde{\mathbf{\Theta}}_{\text{sAvg}}$. As an example of applying a simple average estimator for penalized regression models, the reader can refer to Battey et al. (2018) which studied distributed parameter estimation methods for penalized regression models and established the asymptotic oracle property of the averaged debiased estimators. However, from a finite sample perspective, the simple average estimator tends to underperform severely for unbalanced settings, as it will be shown in Section 2.5. The weighted average proportional to the sub-sample sizes is slightly more adapted as it accounts for the unbalanced setting by weighting according to the size of the samples on each machine. Note that as the variances of the estimators are unknown in practice, in the sequel, the special cases (2.4.2) and (2.4.3) are considered as general simple and weighted average estimators with unknown variances and their asymptotic distribution as well as the asymptotic distribution of the optimally weighted average estimator are investigated in Theorem 2.4.2 as $n_\dagger$ and K both grow with the sample size n, where we remind the reader that $n_\dagger = \min_{1 \leq k \leq K} n_k$.

Using (2.3.2), by considering the consistent estimator $\hat{\sigma}^2_{ab,k}$ for σ^2_{ab}, the aggregated estimator for the (a,b)-th element of $\mathbf{\Theta}$, $(a,b) \in \mathcal{V} \times \mathcal{V}$, is of the form

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \hat{\mathbf{\Theta}}^d_{ab,k}. \tag{2.4.4}$$

This estimator behaves like a weighted average where the weights are a function of sub-sample size n_k and estimated variance $\hat{\sigma}^2_{ab,k}$, constructed as in Jankova and van de Geer (2015), based on each sub-sample. Using (2.3.3) and (2.4.4), it can be

shown that

$$\sqrt{\sum_{k=1}^{K}\frac{n_k}{\hat{\sigma}^2_{ab,k}}}\left(\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}}-\boldsymbol{\Theta}_{ab}\right)=\mathbf{W}_{ab,\dagger}+\mathbf{R}_{ab,\dagger}, \tag{2.4.5}$$

where

$$\mathbf{W}_{ab,\dagger}=\sqrt{\frac{\sum_{k=1}^{K}n_k/\sigma^2_{ab}}{\sum_{k=1}^{K}n_k/\hat{\sigma}^2_{ab,k}}}\times\sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab})$$

and

$$\mathbf{R}_{ab,\dagger}=\sqrt{\frac{\sum_{k=1}^{K}n_k/\sigma^2_{ab}}{\sum_{k=1}^{K}n_k/\hat{\sigma}^2_{ab,k}}}\sum_{k=1}^{K}\frac{n_k\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}\mathbf{R}_{ab,k},$$

where $\boldsymbol{\Theta}_a$ is the a-th column of $\boldsymbol{\Theta}$ and $\dot{\mathbf{Y}}_{l,k}=(Y^1_{l,k},\ldots,Y^p_{l,k})^\top$ as in Section 2.3. In words, it represents the p-dimensional vector of variables corresponding to the l-th unit from the k-th sub-sample. Moreover, for the wAvg estimator, denoted by $\tilde{\boldsymbol{\Theta}}_{ab,\text{wAvg}}$, and the sAvg estimator, denoted by $\tilde{\boldsymbol{\Theta}}_{ab,\text{sAvg}}$, it can be shown that

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k}}}(\tilde{\boldsymbol{\Theta}}_{ab,\text{wAvg}}-\boldsymbol{\Theta}_{ab})=\mathbf{W}'_{ab,\dagger}+\mathbf{R}'_{ab,\dagger}, \tag{2.4.6}$$

and

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)}}(\tilde{\boldsymbol{\Theta}}_{ab,\text{sAvg}}-\boldsymbol{\Theta}_{ab})=\mathbf{W}''_{ab,\dagger}+\mathbf{R}''_{ab,\dagger}, \tag{2.4.7}$$

where

$$\begin{aligned}
\mathbf{W}'_{ab,\dagger}&=\frac{\sqrt{K}}{\sqrt{n\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k}}}\sum_{k=1}^{K}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab}),\\
\mathbf{W}''_{ab,\dagger}&=\frac{\sqrt{K}}{\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)}}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{1}{n_k}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab}),\\
\mathbf{R}'_{ab,\dagger}&=\frac{\sqrt{K}}{\sqrt{n\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k}}}\sum_{k=1}^{K}n_k\mathbf{R}_{ab,k},\\
\mathbf{R}''_{ab,\dagger}&=\frac{\sqrt{K}}{\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)}}\sum_{k=1}^{K}\mathbf{R}_{ab,k}.
\end{aligned}$$

In the following, it is assumed that the precision matrix $\mathbf{\Theta}$ satisfies the irrepresentability condition (A1) with a constant $\alpha_1 \in (0, 1]$, where $1/\alpha_1 = O(1)$, and the bounded eigenvalues condition (A2). The quantities $\kappa_{\mathbf{\Sigma}}$, $\kappa_{\mathbf{H}}$ are also considered bounded as $\kappa_{\mathbf{\Sigma}} = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$, respectively. Lemma 2.4.1 shows that by considering upper bounds on K, the remainder terms $\mathbf{R}_{ab,\dagger}$, $\mathbf{R}'_{ab,\dagger}$ and $\mathbf{R}''_{ab,\dagger}$ are negligible. Moreover, Theorem 2.4.2 states that the terms $\mathbf{W}_{ab,\dagger}$, $\mathbf{W}'_{ab,\dagger}$ and $\mathbf{W}''_{ab,\dagger}$ converge weakly to $\mathcal{N}(0, 1)$ as K and $n_\dagger$ grow.

Lemma 2.4.1. Suppose that $\hat{\mathbf{\Theta}}_k$ $(k = 1, \ldots, K)$ is the solution to the optimization problem (2.3.1) with tuning parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$ and consider $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$, $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ and $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$ as the optimally weighted, sample size weighted and simple average estimators. Then for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$,

a) the remainder term $|\mathbf{R}_{ab,\dagger}|$ of the owAvg estimator grows at the rate

$$|\mathbf{R}_{ab,\dagger}| = O_p\left(K/\sqrt{n}\max\left\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\right\}\right), \tag{2.4.8}$$

which is $o_p(1)$ as long as K and the maximum node degree of $\mathbf{\Theta}$ grow at the rates of $K = O(n^{1/3}/(d\log(p)))$ and $d^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, respectively. Moreover, the remainder term $|\mathbf{R}'_{ab,\dagger}|$ of the wAvg estimator grows at the same rate as (2.4.8),

b) the remainder term $|\mathbf{R}''_{ab,\dagger}|$ of the sAvg estimator grows at the rate

$$|\mathbf{R}''_{ab,\dagger}| = O_p\left(\sqrt{Kn}\max\left\{d^{3/2}\log(p)/n_\dagger, d^2(\log(p)/n_\dagger)^{3/2}\right\}\right), \tag{2.4.9}$$

which is $o_p(1)$ as long as K and the maximum node degree of $\mathbf{\Theta}$ grow at the rates of $K = O(n_\dagger^{4/3}/n)$ and $d^{3/2} = o(n_\dagger^{1/3}/\log(p))$, respectively.

The proof is given in Appendix 2.8.1

Theorem 2.4.2 below shows the asymptotic unbiasedness and normality of the optimally weighted, sample size weighted, and simple averages estimators. As such, one can easily construct confidence intervals and statistical tests based on them.

Theorem 2.4.2. *Under the assumptions of* Lemma 2.4.1*,*

a) for the owAvg estimator in (2.4.5)*, we have* $\mathbf{W}_{ab,\dagger} \xrightarrow{d} \mathcal{N}(0, 1)$.

b) for the wAvg estimator in (2.4.6)*, we have* $\mathbf{W}'_{ab,\dagger} \xrightarrow{d} \mathcal{N}(0, 1)$.

c) for the sAvg estimator in (2.4.7)*, we have* $\mathbf{W}''_{ab,\dagger} \xrightarrow{d} \mathcal{N}(0, 1)$.

The proof is given in Appendix 2.8.2.

Remark 2.1. Based on the asymptotic distribution of the distributed estimators from Theorem 2.4.2, one can construct inferential procedures. Accordingly, the $(1-\alpha)100\%$ asymptotic confidence interval for $\mathbf{\Theta}_{ab}$ using the optimally weighted, sample size weighted and simple average estimators can be constructed as

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}},$$

$$\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})/(nK)},$$

$$\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K} 1/n_k)/K^3},$$

where $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution. Moreover, the rejection region at level α for the hypothesis test $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ (which indicates that there is no edge between nodes corresponding to Y^a and Y^b) vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$ can be constructed as

$$|\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}| > \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}},$$

$$|\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})/(nK)},$$

$$|\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K} 1/n_k)/K^3},$$

using the optimally weighted, sample size weighted and simple average estimators, respectively.

The coverage probabilities and the length of the confidence intervals are investigated via a simulation study further in Section 2.5.

Theorem 2.4.3 provides the convergence rate of the optimally weighted estimator $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ with respect to the elementwise ℓ_∞ norm.

Theorem 2.4.3. *Under assumptions* (A1) *and* (A2) *with* $1/\alpha_1 = O(1)$, $\kappa_{\mathbf{\Sigma}} = O(1)$, $\kappa_{\mathbf{H}} = O(1)$, *the maximal distance of the distributed estimator* $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ *from the true precision matrix* $\mathbf{\Theta}$ *is of order*

$$\begin{aligned}&\|\tilde{\mathbf{\Theta}}_{\text{owAvg}} - \mathbf{\Theta}\|_\infty \\ &= O_p\left(K \max\left\{d\sqrt{\log(p)/n}, d^{3/2}\log(p)/n, d^2(\log(p))^{3/2}/(n\sqrt{n_\dagger})\right\}\right). \quad (2.4.10)\end{aligned}$$

Moreover, under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ *and* $K = O(n^{1/3}/(d\log(p)))$*, the estimator* $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ *is consistent for* $\mathbf{\Theta}$*.*

The proof is given in Appendix 2.8.3. By considering the convergence rate (2.3.5) for the centralized full estimator with size n, and comparing it with (2.4.10), it is observed that as long as K is not too large, the distributed estimator exhibits a similar statistical error to the full one, and it can approximate the full estimator properly.

Remark 2.2. When the true precision matrix is block diagonal with M blocks, the quantities $\kappa_{\mathbf{\Sigma}}$ and $\kappa_{\mathbf{H}}$ are tractable. Consider d'_m as the size of block $\mathbf{\Theta}^m$, $m = 1, \ldots, M$, and the maximum row degree of $\mathbf{\Theta}$ as $d := \max_{m=1,\ldots,M} d'_m$. Then due to the relation between the matrix ℓ_∞ norm and the spectral norm, and the fact that $d'_m \leq d$,

$$\kappa_{\mathbf{\Sigma}} = \max_{m=1,\ldots,M} |||(\mathbf{\Theta}^m)^{-1}|||_\infty \leq \sqrt{d} \max_{m=1,\ldots,M} \|(\mathbf{\Theta}^m)^{-1}\|_2,$$

where $\|\cdot\|_2$ is the spectral norm of a matrix defined as $\|\mathbf{A}\|_2 = \sqrt{\Lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$ for an arbitrary real valued matrix $\mathbf{A}$ and $\Lambda_{\max}(\mathbf{A}^\top \mathbf{A})$ is the maximum eigenvalue of $\mathbf{A}^\top \mathbf{A}$. Based on the fact that the eigenvalues of $\mathbf{\Sigma}$ are the inverses of the eigenvalues of $\mathbf{\Theta}$ and using assumption (A2), it is deduced that $\kappa_{\mathbf{\Sigma}} = O(\sqrt{d})$. The same argument can be made for $\kappa_{\mathbf{H}}$ and it holds that $\kappa_{\mathbf{H}} = O(d)$. Then, using (2.3.5), the convergence rate of $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ simplifies to

$$\begin{aligned} &\|\tilde{\mathbf{\Theta}}_{\text{owAvg}} - \mathbf{\Theta}\|_\infty \\ &= O_p\left(K \max\left\{ d\sqrt{\log(p)/n}, d^{5/2}\log(p)/n, d^4(\log(p))^{3/2}/(n\sqrt{n_\dagger}) \right\} \right) \end{aligned} \tag{2.4.11}$$

and considering the sparsity condition $d^{5/2} = o(\sqrt{n_\dagger}/\log(p))$ and $K = O(n^{1/3}/(d\log(p)))$, the consistency of $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ follows directly.

Note that the convergence rate (2.4.11) is valid only for the entries of the block diagonals in the case when the block structures are known. In practice however, the block structures are unknown and as such, in the simulation part of Section 2.5 corresponding to the block diagonal data generating process, the errors and confidence intervals are computed for all elements of the matrix.

2.5 Simulation study

In this section, we illustrate the performance of the proposed distributed estimator with a simulation study. To conduct the simulation, we set the total sample size $n = 50000$ as fixed and changed $K = 5, 10$ and 20. To show the performance of the distributed estimator in the unbalanced setting, we considered the following data splitting procedure.

Suppose that among all available machines, two of them are powerful. The first one is the most powerful and $(55-K)\%$ of the dataset is distributed on this machine. The second one is less powerful than the first one and $(60-K)\%$ of the remaining dataset is distributed on this one. The remaining dataset is distributed roughly equally on the remaining machines. To compare the performance of the distributed estimator, we considered the following competitors:

1) (Full) A debiased estimator based on the non-distributed, full data given by the debiased graphical Lasso (the estimator of Jankova and van de Geer, 2015), denoted by $\hat{\mathbf{\Theta}}^F$.

2) (sAvg) An estimator based on splitting data and averaging directly the debiased graphical Lasso estimators from each machine i.e., $\bar{\mathbf{\Theta}}_{\text{sAvg}}$.

3) (wAvg) An estimator based on splitting data and taking the sample size weighted average of the debiased graphical Lasso estimators from each machine, i.e., $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$.

4) (Top1) The most powerful machine which takes $(55-K)\%$ of the dataset. Since estimation on each machine is consistent and asymptotically normal, investigating the performance of the first machine which takes most of the dataset is relevant.

5) (Thresholded I) The thresholded distributed estimator introduced by Arroyo and Hou (2016), which is a sparse estimator built by thresholding debiased estimators. To obtain this estimator, a hard threshold ρ is applied on the debiased estimator (2.3.2) on the k-th sub-sample, such that for every $(a,b) \in \mathcal{V}\times\mathcal{V}, a \neq b$,

$$\hat{\mathbf{\Theta}}^{d,\rho}_{ab,k} = \hat{\mathbf{\Theta}}^{d}_{ab,k}\mathbb{I}(|\hat{\mathbf{\Theta}}^{d}_{ab,k}| > \rho\hat{\sigma}_{ab,k}),$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k}$, the constructed estimator of Jankova and van de Geer (2015) based on the k-th sub-sample. The final aggregated estimator of Arroyo and Hou (2016) is proposed by taking a simple average over K machines, i.e., $\hat{\mathbf{\Theta}}^D_{ab} = \frac{1}{K}\sum_{k=1}^K \hat{\mathbf{\Theta}}^{d,\rho}_{ab,k}$, and then applying the final hard threshold τ on $\hat{\mathbf{\Theta}}^D_{ab}$ for every $(a,b) \in \mathcal{V}\times\mathcal{V}, a \neq b$, such that $\hat{\mathbf{\Theta}}^{D,\tau}_{ab} = \hat{\mathbf{\Theta}}^D_{ab}\mathbb{I}(|\hat{\mathbf{\Theta}}^D_{ab}| > \tau\hat{\sigma}_{ab})$, where $\hat{\sigma}^2_{ab} = \frac{1}{K}\sum_{k=1}^K \hat{\sigma}^2_{ab,k}$.

6) (Thresholded II) The thresholded estimator of Wang and Cui (2021), which is proposed for Transelliptical graphical models. The Transelliptical family, defined by Liu et al. (2012), is an extension of the nonparanormal family proposed by Liu et al. (2010). Just as the nonparanormal extends the normal by transforming the variables using univariate functions, the transelliptical extends the elliptical family in the same way. The estimator of Wang and Cui (2021) is constructed by obtaining the nonparametric Kendall's τ statistic

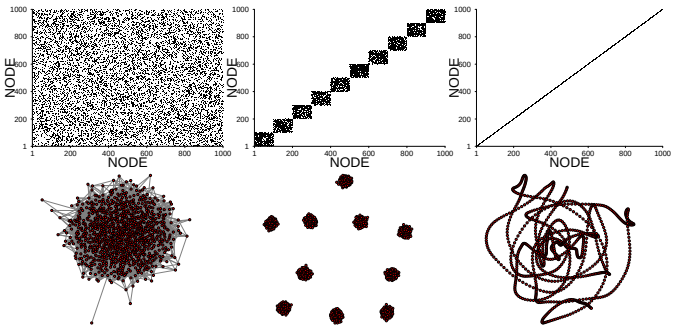


Figure 2.1: Visual representation of three $\boldsymbol{\Theta}$ matrices and their graph structures for $p = 1000$ nodes. From left to right: random, block diagonal with 10 blocks and tridiagonal.

as the correlation matrix estimator in each sub-sample and then plugging it into the Lasso D-trace optimization procedure (Zhang and Zou, 2014). By debiasing these estimators in the sub-samples and then taking a simple average over all debiased estimators, the aggregated estimator is constructed. A hard threshold is applied on this estimator to get the final aggregated sparse estimator.

A comparison with the full estimator reveals how much the performance deteriorates due to splitting the data, while a comparison with the weighted and simple average estimators has the purpose to evaluate if indeed the proposed estimator is better equipped to tackle unbalanced settings due to a more appropriate weighting. A comparison with the Top1 estimator has the purpose to evaluate if the remaining $(K-1)$ machines which account for $(45+K)\%$ of the original data are still able to produce informative estimators even though they receive low amounts of data. For the data generating process, we fixed the number of variables (nodes) to $p = 1000$ and the samples are generated from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ such that $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is graph structured and has the following sparse structures:

- random with probability of connection 0.05;
- block diagonal with 10 blocks and 0.2 as the probability of connection in each block;
- chain or tridiagonal with 1 on the diagonal and 0.2 on the upper and lower diagonals.

The corresponding graph structures are shown in Figure 2.1 for illustration. In this study, the tuning parameter λ_k in the graphical Lasso algorithm has been set to $\lambda_k = \lambda = \sqrt{\log(p)/n}$ for all simulation settings like in Jankova and van de

Geer (2015) except for the tridiagonal case. Due to the high sparsity of the precision matrix in the tridiagonal structure, we increased slightly the regularization parameter to $\lambda_k = \sqrt{\log(p)/n_k}$. Moreover, $\lambda_k = \sqrt{\log(p)/n_k}$ for the thresholded estimator of Wang and Cui (2021) was used to get comparable results with the other competitors. Alternatively, λ_k can be chosen by K-fold cross validation in each sub-sample. The final threshold τ for the estimator of Arroyo and Hou (2016) has been set to $\tau = \sqrt{\log(p)/n}$ as the authors proposed this value for the theoretical guarantees. Moreover, as they showed a communication bandwidth of size $B \asymp p^{2-c}$, where $0 < c \leq 1$ is a constant, is enough to select the correct set of entries, we considered $B = 10p$ and $50p$. To select the best threshold ρ, we considered 200 candidate values from the interval $(0.001, 0.25)$ and we chose the value of ρ for which the absolute difference between the number of active components estimated and the bandwidth B is minimum. Furthermore, the hard threshold of Wang and Cui (2021) has been set to $5\sqrt{\log(p)/n}$ to ensure comparable results with the other estimators. All simulation results are calculated as averages over $R = 500$ different repetitions.

To compare the performance of the estimators, we used the Frobenius norm, elementwise and matrix ℓ_∞ norms between the estimated precision matrix for each competitor and the true precision matrix. The results are presented in Table 2.1 for the Frobenius norm via the simulation setting explained above and it is observed that the performance of the proposed estimator is similar to that of the non-distributed, full estimator in terms of Frobenius norm. The results for the two other norms lead to similar conclusions and are presented in Appendix B. The proposed norm is computed for the active set S, which corresponds to the edges of graph, and the non-active set S^c, separately. By increasing K from 5 to 20, the norm of the distributed estimator does not increase much and the differences are negligible. This suggests that by splitting observations in combination with the proposed aggregation, one might not lose much information. On the other hand, for the sAvg estimator, by increasing K the norm increases substantially, which suggests that it is highly sensitive to K, as opposed to the distributed estimator we propose. The wAvg estimator is better performing than sAvg. However, its performance is not as satisfactory as that of the optimally weighted estimator and it gets far from the full estimator with increasing K. As observed from Table 2.1, the Frobenius norm of Top1 is much larger than that of the centralized full estimator. Moreover, it provides a worse performance compared to the proposed estimator. As such, by considering just the first machine with the largest amount of data and disregarding the remaining machines one loses more information as this strategy does not provide an accurate estimate. The results of the Thresholded I with bandwidth $B = 50p$ and the Thresholded II are also reported in this table. The Frobenius norm for these estimators on the active set increases considerably with increasing K. However, since these estimators are sparse, their errors on the non-active set are much smaller than those of the (non-sparse) proposed estimators. Moreover, it seems that the Thresholded I works better than the Thresholded II

Table 2.1: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets, for the proposed estimator and six competitors when $n = 50000$ and $p = 1000$. Three different graph structures are considered.

		Active set				Non-active set			
		K				K			
		1	5	10	20	1	5	10	20
Random	Full	1.83				6.14			
		(.01)				(.01)			
	owAvg		1.85	1.77	2.05		6.19	6.40	6.86
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Top1		2.43	2.55	2.86		8.67	9.14	10.37
			(.01)	(.01)	(.01)		(.01)	(.01)	(.02)
	sAvg		2.40	3.34	5.67		8.55	11.03	12.25
			(.01)	(.02)	(.02)		(.01)	(.02)	(.01)
	wAvg		1.85	1.78	2.47		6.19	6.46	7.19
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Thresholded I		3.48	5.50	6.39		0.20	0.58	0.31
			(.03)	(.04)	(.03)		(.02)	(.02)	(.02)
	Thresholded II		3.18	4.65	5.84		0.10	1.22	3.90
			(.01)	(.02)	(.02)		(.05)	(.04)	(.04)
Block diag.	Full	1.37				6.11			
		(.01)				(.01)			
	owAvg		1.34	1.26	1.84		6.16	6.37	6.82
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Top1		1.71	1.77	1.95		8.63	9.10	10.32
			(.02)	(.02)	(.02)		(.01)	(.01)	(.02)
	sAvg		1.62	2.89	6.14		8.51	10.99	12.20
			(.01)	(.02)	(.02)		(.01)	(.01)	(.01)
	wAvg		1.34	1.30	2.50		6.17	6.43	7.16
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Thresholded I		1.62	2.90	5.82		0.95	1.01	0.67
			(.01)	(.02)	(.02)		(.02)	(.02)	(.03)
	Thresholded II		3.41	5.29	6.83		0.52	2.48	5.02
			(.02)	(.02)	(.02)		(.04)	(.03)	(.02)
Tridiag.	Full	0.21				5.07			
		(.00)				(.01)			
	owAvg		0.22	0.21	0.21		4.85	4.73	4.54
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Top1		0.30	0.31	0.36		7.04	7.40	8.32
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	sAvg		0.29	0.31	0.33		6.21	6.63	5.92
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	wAvg		0.22	0.21	0.21		4.86	4.76	4.60
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Thresholded I		0.29	0.31	0.33		1.21	1.07	0.50
			(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Thresholded II		0.82	0.82	0.80		0.00	0.00	0.00
			(.01)	(.01)	(.01)		(.00)	(.00)	(.00)

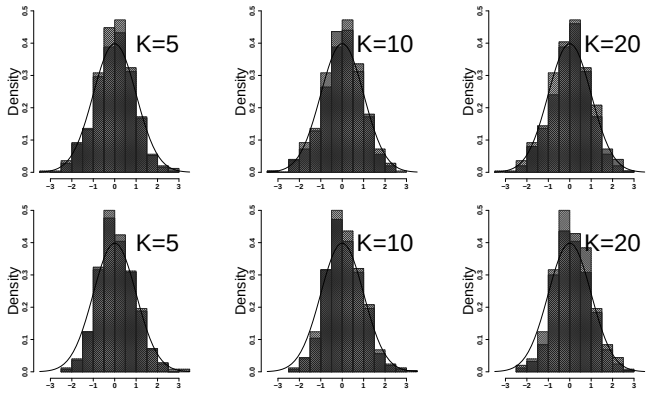


Figure 2.2: Histograms of the normalized full and distributed estimators corresponding to pair $(a, b) = (1, 3)$ on the top row and $(a, b) = (1, 10)$ on the bottom row for the random graph structure when $n = 50000$, $p = 1000$ and the number of machines is $K = 5,\ 10$ or 20. The light and dark gray histograms present the histograms of the normalized full and distributed estimators, respectively. The density of the standard normal distribution is also presented.

on the non-active set in random and block diagonal structures, especially with the smaller final threshold that we considered. However, the results for this estimator are dependent on the bandwidth to be used. The results with $B = 10p$ were much worse than the reported ones and they are not mentioned in the table. This behavior holds on both the active and the non-active sets for all three structures of $\mathbf{\Theta}$ and points to the importance of accurately choosing the extra tuning parameters for this competitor.

In addition to error norms, comparing the asymptotic distribution and inferential properties of the proposed estimator with those of the competitors is of interest. To this end, histograms of the normalized full and distributed estimators are reported in Figure 2.2 for the random structure. The light gray histograms correspond to the normalized full estimator and the dark ones correspond to the proposed estimator. The normalized form of the full estimator is obtained using the technique proposed in Jankova and van de Geer (2015) as $\sqrt{n}(\hat{\mathbf{\Theta}}^F_{ab} - \mathbf{\Theta}_{ab})/\hat{\sigma}_{ab}$, where $\hat{\mathbf{\Theta}}^F_{ab}$ is the (a, b)-th element of $\hat{\mathbf{\Theta}}^F$, while the normalized distributed estimator is obtained as $\sqrt{\sum_{k=1}^K n_k/\hat{\sigma}^2_{ab,k}}(\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} - \mathbf{\Theta}_{ab})$. As an illustrative example, the figures are reported for $(a, b) = (1, 3)$ and $(1, 10)$ although similar conclusions hold also for other couples $(a, b) \in \mathcal{V} \times \mathcal{V}$. We conclude that both distributions are close to the reference $\mathcal{N}(0, 1)$.

The coverage probability and the lengths of the confidence intervals are also obtained for the proposed estimator and competitors at significance level $\alpha = .05$

Table 2.2: Average coverage probability and average length of the confidence interval over 500 repetitions for the proposed estimator and competitors when $n = 50000$ and $p = 1000$ for random and block diagonal graphs.

				Avg.Cov				Avg.Len			
				K				K			
				1	5	10	20	1	5	10	20
Random	Active set		Full	.92				.03			
			owAvg		.91	.93	.90		.03	.03	.03
			Top1		.93	.93	.94		.04	.04	.05
			sAvg		.93	.89	.62		.04	.05	.05
			wAvg		.92	.95	.88		.03	.03	.03
	Non-active	set	Full	.97				.03			
			owAvg		.97	.97	.97		.03	.03	.03
			Top1		.97	.97	.97		.04	.04	.05
			sAvg		.97	.96	.95		.04	.05	.05
			wAvg		.97	.98	.98		.03	.03	.03
Block diag.	Active set		Full	.85				.03			
			owAvg		.86	.89	.72		.03	.03	.03
			Top1		.90	.90	.91		.04	.04	.05
			sAvg		.91	.72	.24		.04	.05	.05
			wAvg		.87	.90	.60		.03	.03	.03
	Non-active	set	Full	.97				.03			
			owAvg		.97	.97	.96		.03	.03	.03
			Top1		.97	.97	.97		.04	.04	.04
			sAvg		.96	.96	.95		.04	.04	.05
			wAvg		.97	.97	.98		.03	.03	.03

except for the thresholded estimators which are not asymptotically normally distributed. To this end, we estimated the coverage probability of the confidence intervals by computing their empirical version. The empirical probability that the true value $\boldsymbol{\Theta}_{ab}$ is included in the confidence interval is defined as $\hat{\mathbb{P}}_{ab} = \frac{\#\{\boldsymbol{\Theta}_{ab} \in \mathrm{CI}_{ab,r}\}}{R}$, where R is the number of simulation repetitions, $\mathrm{CI}_{ab,r}$ is the confidence interval for $\boldsymbol{\Theta}_{ab}$ at the r-th repetition, and $\#$ denotes the number of times that $\boldsymbol{\Theta}_{ab}$ belongs to the confidence interval. After obtaining $\hat{\mathbb{P}}_{ab}$ for all $(a, b) \in \mathcal{V} \times \mathcal{V}$, the average coverage probability on the active set S was obtained as $\mathrm{Avg.Cov}_S = \frac{1}{s} \sum_{(a,b)\in S} \hat{\mathbb{P}}_{ab}$, where s is the number of active components. Similar computations have been implemented for obtaining the estimated coverage probability over the non-active set S^c.

The obtained results are reported in Table 2.2 for the random and block diagonal structures. The results of tridiagonal were close to the random one and they are reported in Appendix B. The proposed estimator performs similarly to the

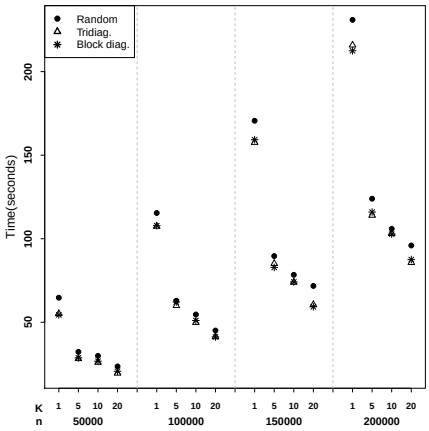


Figure 2.3: Running time in seconds for the full and proposed estimator for the random graph structure, when $p = 1000$. The regularization parameter for the distributed estimator is considered as $\lambda_k = \sqrt{\log(p)/n_k}$.

non-distributed one on both the active and the non-active sets, as the coverage probabilities are close to the nominal level of 95%. Moreover, the average lengths are relatively low and are stable with increasing K. The coverage probability of Top1 is also close to the nominal level, but its length is slightly larger than the length of the full and of the distributed estimator. The coverage probability for sAvg on the active set is low and it gets worse with increasing K. Moreover, the length of its confidence interval gets larger with increasing K. The performance of wAvg is generally better than that of sAvg, but it is still worse than the performance of the distributed estimator.

Another quantity which is important to keep track of, is the running time. In this chapter, the running time is defined as the sum of the maximum time among all parallel jobs and the time to combine the results. These results are shown in Figure 2.3 for the three graph structures. The symbols for the case $K = 1$ represent the full estimator for different sample sizes ranging from $n = 50000$ to 200000 and the remaining symbols represent the distributed estimator for different sample sizes and $K = 5, 10, 20$. For any fixed sample size, the computation time of the distributed estimator is less than the full one and as expected it decreases with increasing K. This behavior is the same for all three graph structures and shows the efficiency of the proposed estimator in terms of computation time.

In the last part of the simulation study, we investigated the effect of the machine with the smallest sample size on the aggregated estimator. The estimator that we have proposed is a weighted average of the local estimators, where the weight for the k-th estimator is of the form $w_{ab,k} = \frac{n_k/\hat{\sigma}^2_{ab,k}}{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}$, $k = 1, \ldots, K$, for every

Table 2.3: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets in three different scenarios.

	Active set		Non-active set	
	K		K	
	1	2	1	2
scenario I	.4521		1.3620	
	(.0269)		(.0183)	
scenario II		.4492		1.3603
		(.0260)		(.0180)
scenario III		.4418		1.3597
		(.0268)		(.0185)

pair $(a, b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$. If the sub-sample size is too small, the estimated weight assigned to the corresponding estimator will also be very small, and the estimator with the smallest sample size will not have a large effect on the final aggregated estimator. To see this, we considered a simulation study with three different scenarios:

I) a full dataset with sample size $n = 10000$ on one machine;

II) the same dataset as in scenario I, but with 10 additional data points on the second machine;

III) the same dataset as in scenario I, but with 100 additional data points on the second machine.

To generate the model, a random graph structure is considered with $p = 100$ nodes and 0.05 as the probability of connection between every pair of nodes. We expect that the performance of the aggregated estimator improves by increasing the number of data points. To this end, we compared the Frobenius norm between the estimated and the true precision matrices in three mentioned scenarios. The results are provided in Table 2.3 as the averages over 500 repetitions. From Table 2.3 it is observed that by increasing n the Frobenius norm decreases on both the active and non-active sets.

Moreover, to investigate the inferential properties, the coverage probability and the length of the confidence intervals are compared for the three mentioned scenarios and presented in Table 2.4. It is observed that the coverage probabilities and the length of the confidence intervals are the same in all three scenarios, which confirms the good performance of the proposed estimator.

To investigate more about the behavior of the distributed estimator when one adds the second machine with a few data points, we presented the estimated weights

Table 2.4: Average coverage probability and average length of the confidence interval over 500 repetitions for the proposed estimator in three different scenarios.

		Avg.Cov		Avg.Len	
		K		K	
		1	2	1	2
Active set	scenario I	.90		.07	
	scenario II		.90		.07
	scenario III		.91		.07
Non-active set	scenario I	.97		.06	
	scenario II		.97		.06
	scenario III		.97		.06

in scenarios II and III in Figure 2.4 for the first and the second machine. They are of the form

$$w_{ab,1} = \frac{n_1/\hat{\sigma}^2_{ab,1}}{n_1/\hat{\sigma}^2_{ab,1} + n_2/\hat{\sigma}^2_{ab,2}}, \quad w_{ab,2} = \frac{n_2/\hat{\sigma}^2_{ab,2}}{n_1/\hat{\sigma}^2_{ab,1} + n_2/\hat{\sigma}^2_{ab,2}},$$

for the (a,b)-th element, $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$. As it is observed in Figure 2.4 for scenario II, when we assign only 10 data points to the second machine, the estimated weights for the corresponding estimator are very small, almost zero, and the estimated weights corresponding to the first machine are close to one. As such, the estimated weights assigned to the second machine are too small and this estimator does not have a large effect on the final aggregated estimator. As it is observed in scenario III, by increasing the number of data points on the second machine, the estimated weights corresponding to this machine increase and so the corresponding estimator has a larger effect on the aggregated estimator. As such, even adding a machine with a few data points does not ruin the performance of the aggregated estimator.

2.6 Real data example

To explore the performance of our proposed estimator on real data, we used the publicly available "4 university web-pages" dataset which was collected in January 1997 by the World Wide Knowledge Base project of the CMU text learning group. The original dataset had been pre-processed by standard text mining methods and it is available in Cardoso-Cachopo (2007). This dataset contains web-pages collected from computer science departments of four US universities Cornell, Texas, Washington and Wisconsin for seven categories: student, faculty, staff, department, course, project and other. The "other" category is a collection of web-pages that

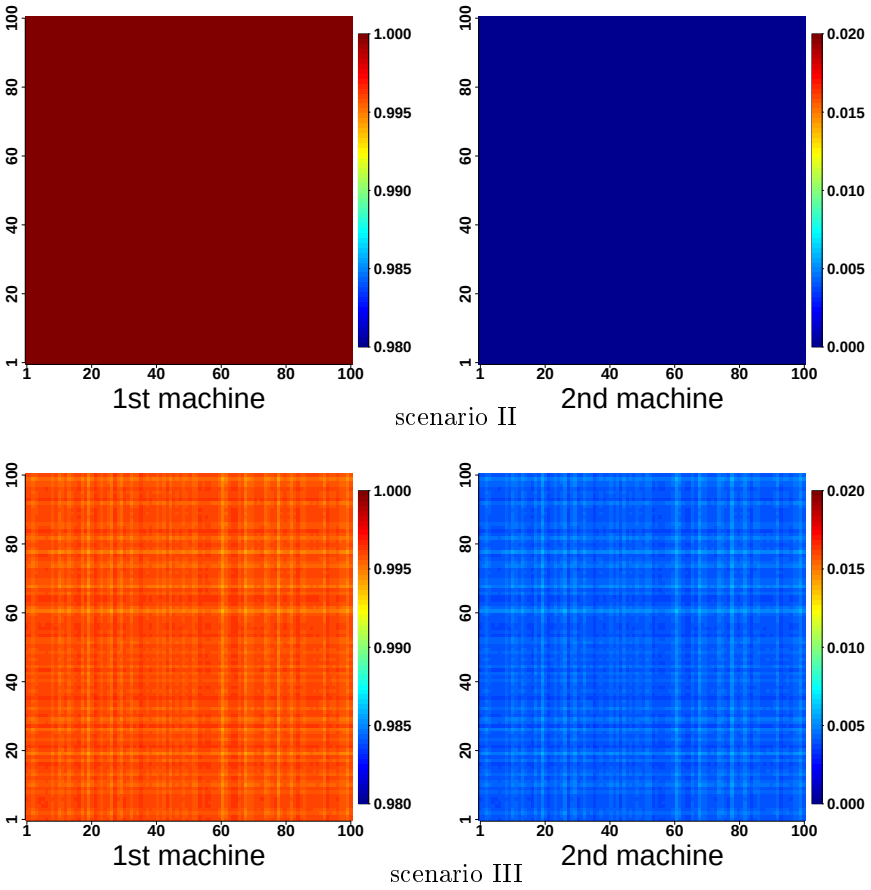


Figure 2.4: Estimated weights for the distributed estimator in scenario II (top row) and III (bottom row) for the first (left panels) and second (right panels) machines.

Table 2.5: Significant and common edges between four competitors and the estimator using the full dataset. For all methods a Bonferroni correction is applied.

	Significant edges				Common edges with Full		
	K				K		
	1	3	4	5	3	4	5
Full	1072						
owAvg		855	835	818	811	789	751
Top1		569	561	569	482	477	478
sAvg		720	547	418	646	478	360
wAvg		821	671	515	783	657	512

were not considered in the six main categories. In this study, we considered the four largest categories which are student, faculty, course and project containing 1641, 1124, 930 and 504 web-pages, respectively. To calculate the term-document matrix $\mathbf{Y}$, we used the log-entropy weighting method of Dumais (1991) which was also implemented in Guo et al. (2011). The final dataset contains $n = 4199$ web-pages as the observations and 7686 distinct terms as variables. We considered $p = 500$ terms with the highest log-entropy weights for the analysis.

We compared next Top1, sAvg, wAvg, the proposed distributed estimator with $K = 3, 4, 5$ and the debiased full estimator. The splitting setting is considered the same as for the simulation study. For all procedures $\lambda = \sqrt{\log(p)/n}$, $\alpha = 5\%$ and a Bonferroni correction is applied for multiple testing. The rejection region for the non-distributed estimator (that uses the full data) of Jankova and van de Geer (2015) is defined as $|\hat{\mathbf{\Theta}}^F_{ab}| > \Phi^{-1}(1 - \frac{\alpha}{p(p-1)})\hat{\sigma}_{ab}/\sqrt{n}$ with Bonferroni correction. The obtained results are presented in Table 2.5. Relative to the competitors, there are more common edges between the graphs estimated by the distributed estimator and the full one, while the least similar competitor is sAvg. By increasing K, the number of common edges tends to decrease for all competitors (except for Top1) due to the loss of information by splitting data on more machines.

The first ten terms with the highest node degrees based on the considered estimators are presented in Figure 2.5. The term "course" is identified as the term with the highest node degree by all estimators except Top1 and sAvg when $K = 5$. Moreover, some terms like "parallel" and "faculty" are common among all estimators. Additionally, the sparse graphical Lasso estimator is considered as a competitor and with this method we identified 18228 edges suggesting that many estimated edges might in fact be false positive edges.

Afterwards, to investigate the performance of the estimators for completely independent variables, we permuted randomly sample data for each variable, thus breaking up the correlation structure in order to construct a dataset with mutually independent variables. As such, all the off-diagonal elements of the precision

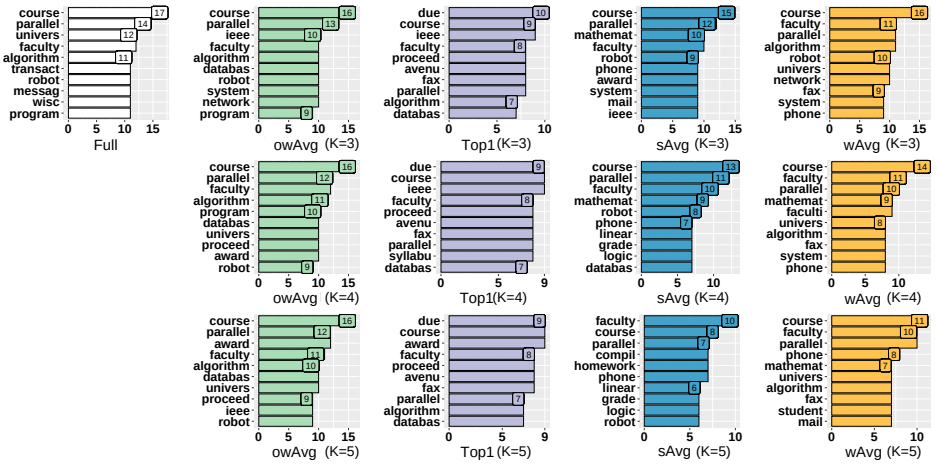


Figure 2.5: The first 10 terms with the highest node degrees, identified by different estimators.

matrix should be estimated as zero. By implementing separately the debiased estimator using the full data, the Top1, the wAvg and the distributed estimator with $K = 3, 4, 5$ machines on this randomly permuted dataset, we identified zero significant edges which confirms this assertion. However, by implementing the sparse graphical Lasso estimator on this independent dataset we identified 1647 edges which confirms the large amount of false positive edges that are identified by this procedure. Also, sAvg identified wrongly between 2 and 19 false positive edges, which confirms the weak performance of this estimator.

2.7 Conclusion and discussion

In this chapter, we proposed a new distributed estimator of the precision matrix in the Gaussian graphical models framework. The estimator is constructed to tackle an unbalanced split of the sample data as input. To improve misaligned selection of non-zero estimated entries on different sub-samples, a bias correction was performed in each sub-sample. This bias correction procedure provides asymptotic normality of the estimators and helps to perform inference. An alternative approach to the debiasing procedure is selective inference, where one first selects a model, and then conditioning on the selected model estimates the parameters of the model. The reader can refer to Candès and Tao (2007) and Javanmard and Montanari (2013) among many others for more details on this procedure. Each of these two procedures has its own pros and cons. The target for the debiased

graphical Lasso procedure is $\boldsymbol{\Theta}$ (the entries in the true model) rather than $\boldsymbol{\Theta}^{\hat{S}}$ (the entries in the selected model $\hat{S}$). These two models are not the same unless $\hat{S}$ happens to contain all non-zero entries of $\boldsymbol{\Theta}$. Although inference for $\boldsymbol{\Theta}$ is appealing, it requires assumptions about the correctness of the model and sparsity of $\boldsymbol{\Theta}$. Moreover, the debiased graphical Lasso does not perform any model selection, which is a disadvantage especially when there are thousands variables in high-dimensional settings. In contrast, the selective inference approach relies heavily on the specific selection methodology which is used before inference. This limits the applicability of post-selection inference in practice. All coverage probabilities and confidence intervals are obtained by conditioning on the event that $\hat{S}$ is selected, i.e., $\hat{S} = S$. On the other hand, it is shown in Ravikumar et al. (2011) that under the irrepresentability condition, the model $\hat{S}$ selected by the graphical Lasso satisfies $\hat{S} \subseteq S$ with high probability. Moreover, under a beta-min condition on the entries of the true precision matrix $\boldsymbol{\Theta}$, the graphical Lasso implies exact variable selection, i.e., $\hat{S} = S$, with high probability. As such, the post-selection inference based on the graphical Lasso needs a beta-min condition. However, the debiased approach of Jankova and van de Geer (2015) does not need any beta-min condition. The advantage of the debiased graphical Lasso to post-model selection is in the situations when the entries of the precision matrix are small but non-zero and thus the beta-min type condition, which guarantees exact variable selection, is violated.

As it was mentioned, the debiased estimators in the sub-samples enjoy asymptotic normality. To aggregate estimators from the sub-samples, a pseudo log-likelihood function was constructed using their asymptotic normal density, and the final distributed estimator was built by maximizing this function. Statistical guarantees for the distributed estimator were provided. Its tractable limit distribution was derived and it was shown that under the sparsity condition $d^{3/2} = o(\sqrt{n_+}/\log(p))$ and the upper bound $K = O(n^{1/3}/(d\log(p)))$, for the number of machines, it follows an asymptotic normal distribution. Moreover, the convergence rate of this consistent estimator was provided. The finite sample performance was evaluated with a simulation study where it was observed that the estimator produced competitive results relative to a non-distributed estimator that uses the entire data. Since we perform estimation by distributing the computational load across multiple machines, not surprisingly the computational time comparison favors our novel estimator. Moreover, this estimator performed substantially better than the naive, average-based estimators in terms of accuracy. It was also observed that the coverage probability of the distributed estimator is close to that of the non-distributed estimator. This points to the fact that in practice, performing distributed estimation across multiple machines in our unbalanced framework induces a minimal loss in performance relative to models using all the data in a centralized location, which might not be feasible in practice.

2.8 Appendix A: Proofs

2.8.1 Proof of Lemma 2.4.1

a) First, we show that $\sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}} \xrightarrow{p} 1$. Note that using Lemma 2 of Jankova and van de Geer (2015), under the multivariate Gaussian distribution, $1/\sigma_{ab} = O(1)$, and under $1/\alpha_1 = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$, we get $|\hat{\sigma}_{ab,k}^2 - \sigma_{ab}^2| = O_p(\sqrt{\log(p)/n_k})$ which is $o_p(1)$ as n_k grows faster than $\log(p)$. It can be shown that the inverse ratio also satisfies

$$|\frac{1}{\hat{\sigma}_{ab,k}^2} - \frac{1}{\sigma_{ab}^2}| = O_p(\sqrt{\log(p)/n_k}). \tag{2.8.1}$$

As such,

$$\begin{aligned}\Big|\sum_{k=1}^K \{\frac{n_k}{n} \times \frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2}\} - 1\Big| &= \sigma_{ab}^2 \Big|\sum_{k=1}^K \frac{n_k}{n}\{\frac{1}{\hat{\sigma}_{ab,k}^2} - \frac{1}{\sigma_{ab}^2}\}\Big| \\ &\leq \sigma_{ab}^2 \sum_{k=1}^K \frac{n_k}{n}|\frac{1}{\hat{\sigma}_{ab,k}^2} - \frac{1}{\sigma_{ab}^2}| = O_p(K\sqrt{\log(p)/n}) \end{aligned} \tag{2.8.2}$$

where the last equality holds using (2.8.1) and the fact that $n_k \leq n$, $k = 1, \ldots, K$. As such, $\Big|\sum_{k=1}^K \{\frac{n_k}{n} \times \frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2}\} - 1\Big| = o_p(1)$ as K grows at the rate $K = O(n^{1/3}/(d\log(p)))$, and as a result $\frac{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}{\sum_{k=1}^K n_k/\sigma_{ab}^2} \xrightarrow{p} 1$. By considering the continuous map $g(x) = 1/\sqrt{x}$, the sequence $\sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}}$ converges in probability to 1 as K and n_k, $k = 1, \ldots, K$, grow.

As such, due to the definition of $\mathbf{R}_{ab,\dagger}$, it is enough to show the bound (2.4.8) for the term $\sum_{k=1}^K \frac{n_k\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2}\mathbf{R}_{ab,k}$. Since the graphical Lasso estimator $\hat{\mathbf{\Theta}}_k$ is positive definite, there exists a positive constant L_k such that with high probability $\hat{\sigma}_{ab,k}^2 = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}_{ab,k}^2 \geq \Lambda_{\min}^2(\hat{\mathbf{\Theta}}_k) > L_k$, where $\Lambda_{\min}(\hat{\mathbf{\Theta}}_k)$ is the minimum eigenvalue of $\hat{\mathbf{\Theta}}_k$, and then $1/\hat{\sigma}_{ab,k}^2 = O_p(1)$. Thus, using (2.3.4), with high probability,

$$\begin{aligned}|\sum_{k=1}^K \frac{n_k\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2}\mathbf{R}_{ab,k}| &\leq \sum_{k=1}^K \frac{n_k}{\sqrt{n}}|\mathbf{R}_{ab,k}| \times O_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{k=1}^K O_p(\max\{d^{3/2}\log(p), d^2\frac{(\log(p))^{3/2}}{\sqrt{n_k}}\}) \\ &\leq \frac{K}{\sqrt{n}} O_p(\max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\}),\end{aligned}$$

and the claimed result in (2.4.8) follows. By considering $K = O(n^{1/3}/(d\log(p)))$ and the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, it can be shown that $|\mathbf{R}_{ab,\dagger}|= o_p(1)$.

Now, to show the result (2.4.8) for the remainder term $|\mathbf{R}'_{ab,\dagger}|$, we first show that $\frac{\sqrt{K}\sigma_{ab}}{\sqrt{\sum_{k=1}^K \hat{\sigma}^2_{ab,k}}} \xrightarrow{p} 1$. We have

$$\begin{aligned}\left|\frac{\sum_{k=1}^K \hat{\sigma}^2_{ab,k}}{K\sigma^2_{ab}} - 1\right| = \left|\frac{1}{K}\sum_{k=}^K \{\frac{\hat{\sigma}^2_{ab,k}}{\sigma^2_{ab}} - 1\}\right| &\leq \frac{1}{K\sigma^2_{ab}}\sum_{k=1}^K |\hat{\sigma}^2_{ab,k} - \sigma^2_{ab}| \\ &= \frac{1}{K}\sum_{k=1}^K O_p(\sqrt{\log(p)/n_k}) \\ &= O_p(\sqrt{\log(p)/n_\dagger}),\end{aligned}$$

which is $o_p(1)$ as $\log(p) = o(n_\dagger)$. Note that the second equality holds using Lemma 2 of Jankova and van de Geer (2015). As such, $\frac{\sum_{k=1}^K \hat{\sigma}^2_{ab,k}}{K\sigma^2_{ab}} \xrightarrow{p} 1$, and by considering the continuous map $g(x) = 1/\sqrt{x}$, we get $\frac{\sqrt{K}\sigma_{ab}}{\sqrt{\sum_{k=1}^K \hat{\sigma}^2_{ab,k}}} \xrightarrow{p} 1$. As such, to show the negligibility of the remainder term $\mathbf{R}'_{ab,\dagger}$, it is enough to show the presented bound in the Lemma for $\sum_{k=1}^K \frac{n_k}{\sqrt{n}\sigma_{ab}}\mathbf{R}_{ab,k}$. By a similar argument as in part a), it can be shown that

$$|\mathbf{R}'_{ab,\dagger}|= O_p\bigg(K/\sqrt{n}\max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\}\bigg).$$

b) As it is shown in part a), $\frac{\sqrt{K}\sigma_{ab}}{\sqrt{\sum_{k=1}^K \hat{\sigma}^2_{ab,k}}} \xrightarrow{p} 1$. As such, to show the negligibility of the remainder term $\mathbf{R}''_{ab,\dagger}$, it is enough to show the bound (2.4.9) for $\frac{1}{\sqrt{\sum_{k=1}^K 1/n_k}}\sum_{k=1}^K \mathbf{R}_{ab,k}/\sigma_{ab}$.
As $n_k < n$, for every $k = 1, \ldots, K$, we have $\frac{1}{\sqrt{\sum_{k=1}^K 1/n_k}} < \sqrt{n/K}$, and we get

$$\begin{aligned}\left|\frac{1}{\sqrt{\sum_{k=1}^K 1/n_k}}\sum_{k=1}^K \mathbf{R}_{ab,k}/\sigma_{ab}\right| &\leq \sqrt{\frac{n}{K}}\sum_{k=1}^K |\mathbf{R}_{ab,k}| \\ &= O_p\bigg(\sqrt{Kn}\max\{d^{3/2}\log(p)/n_\dagger, d^2(\log(p)/n_\dagger)^{3/2}\}\bigg),\end{aligned}$$

which is $o_p(1)$ under the mentioned conditions on K and d.

2.8.2 Proof of Theorem 2.4.2

a) Define

$$\xi_{ab,\dagger} := \sum_{k=1}^{K}\sum_{l=1}^{n_k} \frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} \times \frac{1}{\sqrt{n}\sigma_{ab}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}).$$

As $\sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}} \xrightarrow{p} 1$, using Slutsky's theorem, it is enough to prove that $\xi_{ab,\dagger}$ converges in distribution to $\mathcal{N}(0,1)$. Defining

$$\xi'_{ab,\dagger} := \sum_{k=1}^{K}\sum_{l=1}^{n_k} \frac{1}{\sqrt{n}\sigma_{ab}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

we have

$$\xi'_{ab,\dagger} = \frac{1}{\sigma_{ab}\sqrt{n}}\sum_{l=1}^{n}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l}\dot{\mathbf{Y}}_{l}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

where $\dot{\mathbf{Y}}_l$ is a column vector of length p and is the l-th sample, $l = 1, \ldots, n$, of $\dot{\mathbf{Y}}$. The random variable $\xi'_{ab,\dagger}$ converges to $\mathcal{N}(0,1)$ as it is shown in Theorem 2 of Jankova and van de Geer (2015). As such, we only need to show that $|\xi_{ab,\dagger} - \xi'_{ab,\dagger}| \xrightarrow{p} 0$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, and then convergence in distribution of $\xi'_{ab,\dagger}$ results from the convergence in distribution of $\xi_{ab,\dagger}$. We have

$$\begin{aligned}|\xi_{ab,\dagger} - \xi'_{ab,\dagger}| &= \left|\sum_{k=1}^{K}\{(\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} - 1) \times \frac{1}{\sqrt{n}\sigma_{ab}}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\}\right| \\ &\leq \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\left\{|\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} - 1| \times |\frac{1}{\sigma_{ab}}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})|\right\}. \quad (2.8.3)\end{aligned}$$

Using the definition of $\mathbf{W}_k$ and since $1/\sigma_{ab} = O(1)$, we get

$$|\frac{1}{\sigma_{ab}}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})| = \frac{n_k}{\sigma_{ab}}|\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab}| \leq n_k\|\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\|_\infty.$$

Under assumption (A2) and using Lemma 8 of Jankova and van de Geer (2015), it holds that $\|\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\|_\infty = O_p(d\sqrt{\log(p)/n_k})$. Moreover, by (2.8.1), $|\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} - 1| = O_p(\sqrt{\log(p)/n_k})$. Substituting $K = O(n^{1/3}/(d\log(p)))$ and the mentioned bounds in (2.8.3), we get

$$|\xi_{ab,\dagger} - \xi'_{ab,\dagger}| = O_p(Kd\log(p)/\sqrt{n}) = O_p(1/n^{1/6}),$$

which is $o_p(1)$ as n grows.

b) Similarly to a).

c) Define

$$\xi_{ab,\dagger} := \frac{1}{\sqrt{\sum_{k=1}^{K} 1/n_k}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \frac{1}{n_k \sigma_{ab}} (\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k} \dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}).$$

As $\frac{\sqrt{K}\sigma_{ab}}{\sqrt{\sum_{k=1}^{K} \hat{\sigma}_{ab,k}^2}} \xrightarrow{p} 1$, using Slutsky's theorem, it is enough to show that $\xi_{ab,\dagger}$ converges in distribution to $\mathcal{N}(0,1)$. To show the asymptotic normal distribution of $\xi_{ab,\dagger}$, we use the Lindeberg-Feller condition (see for example, Theorem 4.12 of Kallenberg, 1997). We have $\mathbb{E}(\mathbf{Z}_{ab,l,k}) = 0$, where $\mathbf{Z}_{ab,l,k} := \boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k} \dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$. Moreover,

$$\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \frac{1}{\sigma_{ab}^2 n_k^2 (\sum_{j=1}^{K} 1/n_j)} \mathbb{E}(\mathbf{Z}_{ab,l,k}^2)$$
$$= \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{1}{n_k (\sum_{j=1}^{K} 1/n_j)} = 1,$$

where the first equality is deduced due to the fact that $\mathbb{V}\text{ar}(\mathbf{Z}_{ab,l,k}) = \sigma_{ab}^2$. The reader can refer to Jankova and van de Geer (2015), pp. 1225 for more details. To show the third condition of the Lindeberg-Feller Theorem, since $\mathbf{Z}_{ab,l,k}$ are identical for $l = 1, \ldots, n_k$ and $k = 1, \ldots, K$, for every $\epsilon > 0$, we can write

$$\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \frac{1}{\sigma_{ab}^2 n_k^2 (\sum_{j=1}^{K} 1/n_j)} \mathbb{E}\Big\{ \mathbf{Z}_{ab,l,k}^2 \mathbb{I}\Big(|\mathbf{Z}_{ab,l,k}| > \epsilon \sigma_{ab} n_k \sqrt{\sum_{j=1}^{K} 1/n_j} \Big) \Big\}$$
$$= \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{1}{\sigma_{ab}^2 n_k (\sum_{j=1}^{K} 1/n_j)} \mathbb{E}\Big\{ \mathbf{Z}_{ab,1,1}^2 \mathbb{I}\Big(|\mathbf{Z}_{ab,1,1}| > \epsilon \sigma_{ab} n_k \sqrt{\sum_{j=1}^{K} 1/n_j} \Big) \Big\}, \tag{2.8.4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Following Jankova and van de Geer (2015), pp. 1223, for a random variable X and a positive constant a, one can write

$$\mathbb{E}(X^2 \mathbb{I}(|X| > a)) = a^2 \mathbb{P}(|X| > a) + 2 \int_a^\infty u \mathbb{P}(|X| > u)\, du.$$

Applying this equality to the expectation in (2.8.4), we get

$$\begin{aligned}
&\mathbb{E}\Big\{\mathbf{Z}_{ab,1,1}^2\mathbb{I}\Big(|\mathbf{Z}_{ab,1,1}|> \epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)\Big\}\\
&\quad= \epsilon^2\sigma_{ab}^2 n_k^2(\sum_{j=1}^K 1/n_j)\mathbb{P}\Big(|\mathbf{Z}_{ab,1,1}|> \epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)\\
&\quad+ 2\int_{\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}}^{\infty} u\mathbb{P}(|\mathbf{Z}_{ab,1,1}|> u)\, du. \qquad (2.8.5)
\end{aligned}$$

Using Lemma 4 of Jankova and van de Geer (2015), for a constant $c_1 > 0$, we have

$$\mathbb{P}\Big(|\mathbf{Z}_{ab,1,1}|> \epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big) \le 4d\exp\Big\{-\Big(\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)/(c_1 d)\Big\}, \qquad (2.8.6)$$

where d is the maximal node degree of $\mathbf{\Theta}$. In the integral of (2.8.5), by changing the variable $t := \frac{u}{\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}}$, we get

$$\begin{aligned}
&\int_{\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}}^{\infty} u\mathbb{P}(|\mathbf{Z}_{ab,1,1}|> u)\, du\\
&= \epsilon^2\sigma_{ab}^2 n_k^2(\sum_{j=1}^K 1/n_j)\int_1^{\infty} t\mathbb{P}\Big(|\mathbf{Z}_{ab,1,1}|> t\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)\, dt\\
&\le 4d\epsilon^2\sigma_{ab}^2 n_k^2(\sum_{j=1}^K 1/n_j)\int_1^{\infty} t\exp\Big\{-\Big(t\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)/(c_1 d)\Big\}\, dt\\
&= \exp\Big\{-\Big(\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j}\Big)/(c_1 d)\Big\}\Big\{4c_1 d^2\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^K 1/n_j} + 4c_1^2 d^3\Big\}. \qquad (2.8.7)
\end{aligned}$$

Substituting (2.8.6) and (2.8.7) in (2.8.5) and then in (2.8.4), we get

$$\begin{aligned}
&\lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{1}{\sigma_{ab}^2 n_k^2(\sum_{j=1}^{K}1/n_j)}\mathbb{E}\Big\{\mathbf{Z}_{ab,l,k}^2\mathbb{I}\Big(|\mathbf{Z}_{ab,l,k}|>\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^{K}1/n_j}\Big)\Big\}\\
&\leq \lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}\Big\{4d\epsilon^2 n_k+\frac{4c_1d^2\epsilon}{\sigma_{ab}\sqrt{\sum_{j=1}^{K}1/n_j}}+\frac{4c_1^2d^3}{\sigma_{ab}^2 n_k(\sum_{j=1}^{K}1/n_j)}\Big\}\\
&\qquad\qquad\times\exp\Big\{-\Big(\epsilon\sigma_{ab}n_k\sqrt{\sum_{j=1}^{K}1/n_j}\Big)/(c_1d)\Big\}\\
&\leq \lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}\Big\{4d\epsilon^2 n_k+\frac{4c_1d^2\epsilon}{\sigma_{ab}\sqrt{\sum_{j=1}^{K}1/n_j}}+\frac{4c_1^2d^3}{\sigma_{ab}^2 n_k(\sum_{j=1}^{K}1/n_j)}\Big\}\\
&\qquad\qquad\times\exp\Big\{\frac{-\epsilon\sigma_{ab}}{c_1d}\times\frac{n_\dagger}{n}\times\sqrt{nK}\Big\}\\
&\leq \lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\Big\{4d\epsilon^2 n+\frac{4c_1d^2\epsilon\sqrt{nK}}{\sigma_{ab}}+\frac{4c_1^2d^3}{\sigma_{ab}^2}\Big\}\exp\Big\{\frac{-\epsilon\sigma_{ab}}{c_1d}\times\frac{n_\dagger}{n}\times\sqrt{nK}\Big\}=0,
\end{aligned}$$

where equality to zero is deduced using l'Hôpital's rule and the fact that the terms in the exponential function go faster to $-\infty$. As such, the requirements of the Lindeberg-Feller Theorem are fulfilled and $\xi_{ab,\dagger}$ converges in distribution to $\mathcal{N}(0,1)$.

2.8.3 Proof of Theorem 2.4.3

Based on the definition of the elementwise ℓ_∞ norm, for every $(a,b)\in\mathcal{V}\times\mathcal{V}$ we have

$$\|\tilde{\boldsymbol{\Theta}}_{\text{owAvg}}-\boldsymbol{\Theta}\|_\infty=\max_{a,b}|\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}}-\boldsymbol{\Theta}_{ab}|.$$

It can be shown that

$$\begin{aligned}
|\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}}-\boldsymbol{\Theta}_{ab}|\leq&\Big|\frac{n/\sigma_{ab}^2}{\sum_{k=1}^{K}n_k/\hat{\sigma}_{ab,k}^2}\times\frac{1}{n}\sum_{k=1}^{K}\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab})\Big|\\
&+\Big|\frac{n/\sigma_{ab}^2}{\sum_{k=1}^{K}n_k/\hat{\sigma}_{ab,k}^2}\sum_{k=1}^{K}\frac{n_k\sigma_{ab}^2}{n\hat{\sigma}_{ab,k}^2}\mathbf{R}_{ab,k}\Big|.
\end{aligned}\tag{2.8.8}$$

Since $\frac{n/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2} \xrightarrow{p} 1$, and $1/\hat{\sigma}_{ab,k}^2 = O_p(1)$, as it was shown in the proof of Lemma 2.4.1, for the first term on the right hand side of (2.8.8), we have

$$\begin{aligned}
&\left| \frac{n/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2} \times \frac{1}{n} \sum_{k=1}^K \frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k} \dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}) \right| \\
&\leq O_p(1) \frac{1}{n} \sum_{k=1}^K \left| \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k} \dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}) \right| \\
&\leq O_p(1) \frac{1}{n} \sum_{k=1}^K O_p(d\sqrt{n_k \log(p)}) = O_p(Kd\sqrt{\log(p)/n}), \qquad (2.8.9)
\end{aligned}$$

as it was shown in the proof of Theorem 2.4.2. Moreover, for the second term on the right hand side of (2.8.8), we have

$$\begin{aligned}
&\left| \frac{n/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2} \sum_{k=1}^K \frac{n_k \sigma_{ab}^2}{n\hat{\sigma}_{ab,k}^2} \mathbf{R}_{ab,k} \right| \\
&\quad \leq O_p(1) \sum_{k=1}^K \frac{n_k}{n} O_p\Big(\max\{d^{3/2}\log(p)/n_k, d^2(\log(p)/n_k)^{3/2}\} \Big) \\
&\quad \leq O_p\left(\frac{K}{n} \max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\} \right). \qquad (2.8.10)
\end{aligned}$$

Substituting (2.8.9) and (2.8.10) in (2.8.8), the result of (2.4.10) follows directly.

2.9 Appendix B: Supporting tables

Simulation results to complement Section 2.5 are presented in this section.

Table 2.6: Average and standard deviation (between parentheses) of the elementwise ℓ_∞ norm over 500 repetitions on the active and non-active sets, for the proposed estimator and six competitors when $n = 50000$ and $p = 1000$.

		Active set				Non-active set			
		K				K			
		1	5	10	20	1	5	10	20
Random	Full	.04 (.00)				.03 (.00)			
	owAvg		.04 (.00)	.04 (.00)	.04 (.00)		.03 (.00)	.03 (.00)	.04 (.00)
	Top1		.05 (.00)	.05 (.00)	.06 (.00)		.05 (.00)	.05 (.00)	.06 (.00)
	sAvg		.05 (.00)	.06 (.00)	.08 (.00)		.05 (.00)	.06 (.00)	.07 (.00)
	wAvg		.04 (.00)	.04 (.00)	.04 (.00)		.03 (.00)	.03 (.00)	.04 (.00)
	Thresholded I		.10 (.01)	.12 (.01)	.14 (.01)		.03 (.03)	.04 (.00)	.04 (.00)
	Thresholded II		.07 (.01)	.14 (.01)	.16 (.00)		.06 (.02)	.11 (.01)	.13 (.01)
Block diag.	Full	.06 (.01)				.04 (.01)			
	owAvg		.06 (.01)	.06 (.01)	.06 (.01)		.04 (.01)	.05 (.01)	.06 (.01)
	Top1		.07 (.01)	.07 (.01)	.07 (.01)		.05 (.01)	.06 (.01)	.06 (.01)
	sAvg		.07 (.01)	.08 (.01)	.12 (.01)		.06 (.01)	.08 (.01)	.12 (.02)
	wAvg		.06 (.01)	.06 (.01)	.06 (.01)		.04 (.01)	.05 (.01)	.07 (.01)
	Thresholded I		.07 (.01)	.08 (.01)	.12 (.01)		.05 (.01)	.05 (.01)	.07 (.01)
	Thresholded II		.12 (.01)	.18 (.01)	.24 (.02)		.11 (.02)	.17 (.02)	.23 (.02)
Tridiag.	Full	.02 (.00)				.02 (.00)			
	owAvg		.02 (.00)	.02 (.00)	.02 (.00)		.02 (.00)	.02 (.00)	.02 (.00)
	Top1		.02 (.00)	.02 (.00)	.03 (.00)		.03 (.00)	.03 (.00)	.04 (.00)
	sAvg		.02 (.00)	.02 (.00)	.02 (.00)		.03 (.00)	.03 (.00)	.03 (.00)
	wAvg		.02 (.00)	.02 (.00)	.02 (.00)		.02 (.00)	.02 (.00)	.02 (.00)
	Thresholded I		.02 (.00)	.02 (.00)	.02 (.00)		.03 (.00)	.03 (.00)	.02 (.00)
	Thresholded II		.04 (.00)	.04 (.00)	.04 (.00)		.00 (.00)	.00 (.00)	.02 (.00)

Table 2.7: Average and standard deviation (between parentheses) of the matrix ℓ_∞ norm over 500 repetitions on the active and non-active sets, for the proposed estimator and six competitors when $n = 50000$ and $p = 1000$. Three different structures of the precision matrix are considered.

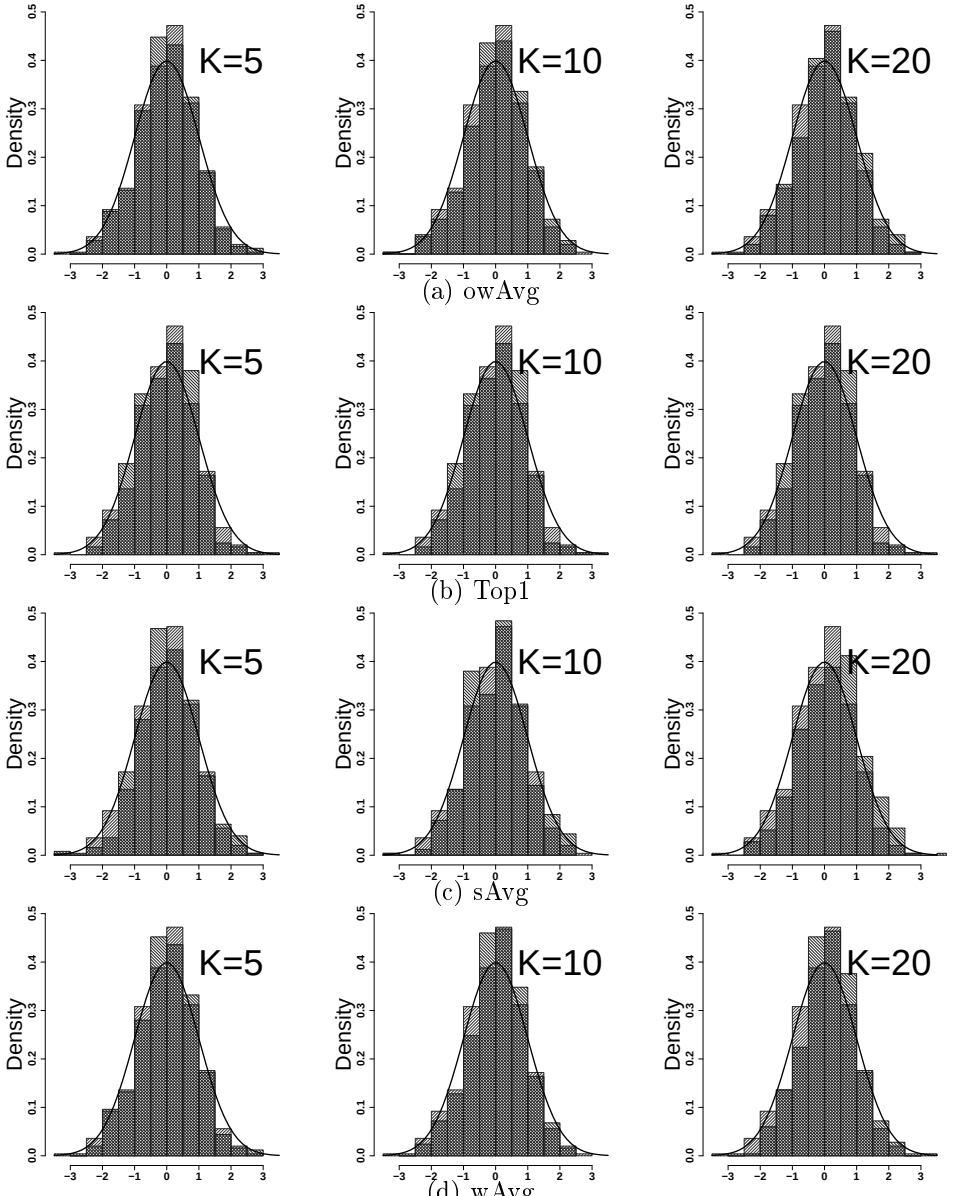
		Active set				Non-active set			
		K				K			
		1	5	10	20	1	5	10	20
Random	Full	.65				5.93			
		(.04)				(.15)			
	owAvg		.69	.63	.63		5.91	6.05	6.42
			(.04)	(.04)	(.03)		(.15)	(.14)	(.14)
	Top1		.84	.88	.97		8.34	8.79	9.93
			(.05)	(.05)	(.06)		(.20)	(.21)	(.30)
	sAvg		.84	1.60	4.87		10.89	13.61	21.92
			(.05)	(.05)	(.06)		(.20)	(.21)	(.20)
	wAvg		.68	.62	.71		5.91	6.08	6.61
			(.04)	(.04)	(.03)		(.15)	(.14)	(.14)
	Thresholded I		1.07	1.87	2.37		.05	.11	.07
			(.07)	(.11)	(.12)		(.01)	(.01)	(.01)
	Thresholded II		1.19	1.77	2.20		.06	.43	1.81
			(.07)	(.07)	(.07)		(.03)	(.08)	(.16)
Block diag.	Full	.70				7.23			
		(.07)				(.19)			
	owAvg		.73	.66	.57		7.22	7.35	7.70
			(.07)	(.07)	(.06)		(.18)	(.19)	(.19)
	Top1		.76	.77	.81		10.17	10.71	12.13
			(.08)	(.09)	(.09)		(.27)	(.28)	(.32)
	sAvg		.77	.72	1.27		9.66	11.71	12.26
			(.08)	(.07)	(.05)		(.26)	(.29)	(.27)
	wAvg		.73	.65	.57		7.22	7.36	7.81
			(.07)	(.07)	(.05)		(.18)	(.18)	(.19)
	Thresholded I		.77	.73	1.22		.17	.18	.13
			(.08)	(.07)	(.05)		(.02)	(.02)	(.02)
	Thresholded II		1.33	2.06	2.66		.29	.93	1.80
			(.08)	(.09)	(.09)		(.06)	(.11)	(.11)
Tridiag.	Full	.02				.4.73			
		(.00)				(.04)			
	owAvg		.03	.02	.02		4.18	4.07	3.91
			(.00)	(.00)	(.00)		(.04)	(.03)	(.03)
	Top1		.03	.04	.04		6.08	6.38	7.18
			(.00)	(.00)	(.00)		(.05)	(.05)	(.06)
	sAvg		.03	.04	.04		5.35	5.71	5.09
			(.00)	(.00)	(.00)		(.04)	(.05)	(.04)
	wAvg		.03	.02	.02		4.19	4.10	3.96
			(.00)	(.00)	(.00)		(.04)	(.03)	(.03)
	Thresholded I		.03	.04	.04		.25	.22	.09
			(.00)	(.00)	(.00)		(.02)	(.02)	(.01)
	Thresholded II		.07	.07	.07		.00	.00	.00
			(.00)	(.00)	(.00)		(.00)	(.00)	(.00)

Table 2.8: Average coverage probability and average length of the confidence interval over 500 repetitions when $n = 50000$ and $p = 1000$ for the tridiagonal graph structure.

		Avg.Cov				Avg.Len			
		K				K			
		1	5	10	20	1	5	10	20
Active set	Full	.96				.02			
	owAvg		.94	.95	.94		.02	.02	.02
	Top1		.94	.94	.95		.03	.03	.04
	sAvg		.95	.95	.92		.03	.03	.03
	wAvg		.93	.93	.93		.02	.02	.02
Non-active set	Full	.93				.02			
	owAvg		.94	.94	.95		.02	.02	.02
	Top1		.94	.94	.94		.03	.03	.03
	sAvg		.95	.96	.96		.03	.03	.02
	wAvg		.94	.93	.93		.02	.02	.02

2.10 Appendix C: Extra figures

In this appendix, some extra figures are provided to present the performance of the proposed estimator. Figures 2.6-2.8 show the normalized distribution of the proposed estimator and competitors. The light gray histograms correspond to the reference Full estimator and the dark gray ones correspond to the distributed estimators.



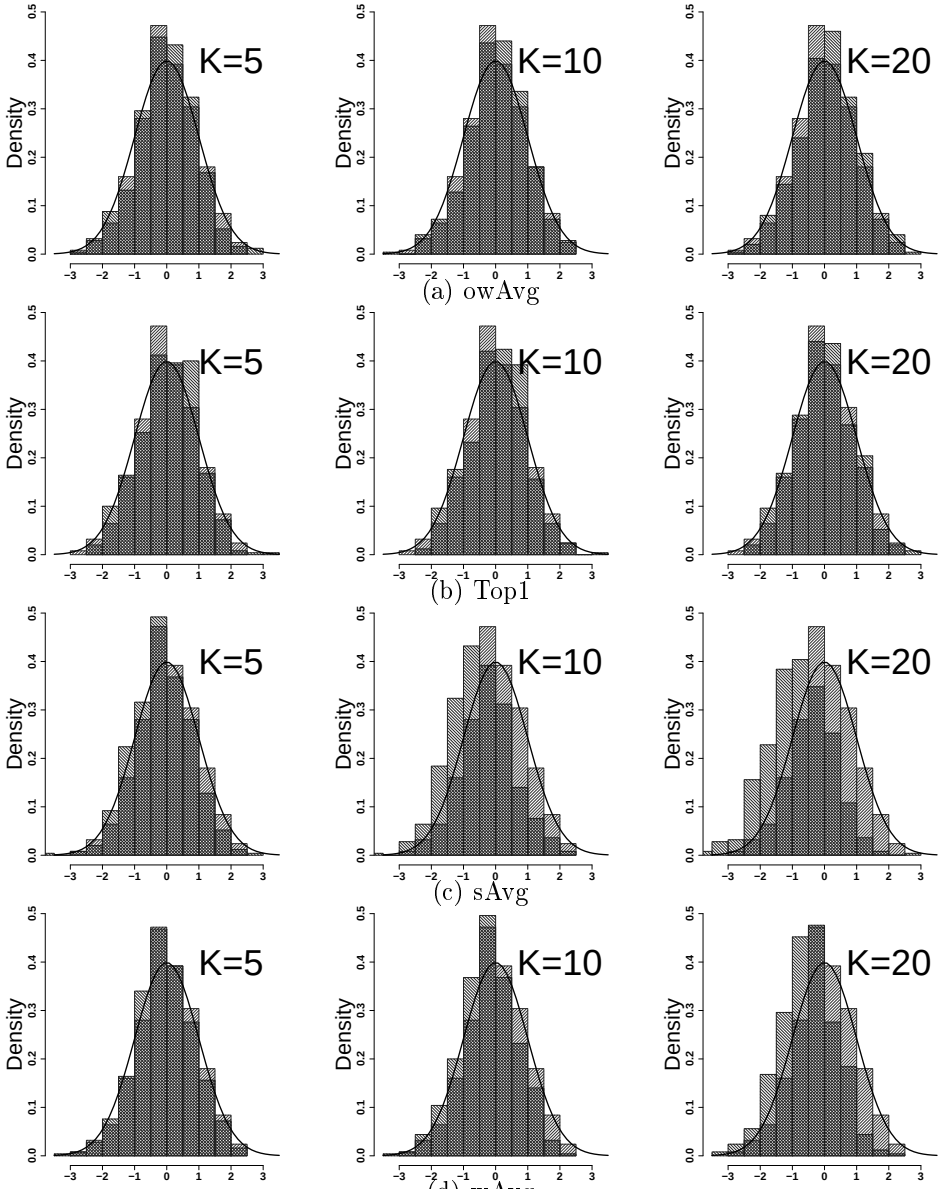
(a) owAvg

(b) Top1

(c) sAvg

(d) wAvg

Figure 2.6: Histogram of the normalized estimators for the pair $(a, b) = (1, 3)$ for the random graph structure when $n = 50000$ and $p = 1000$.



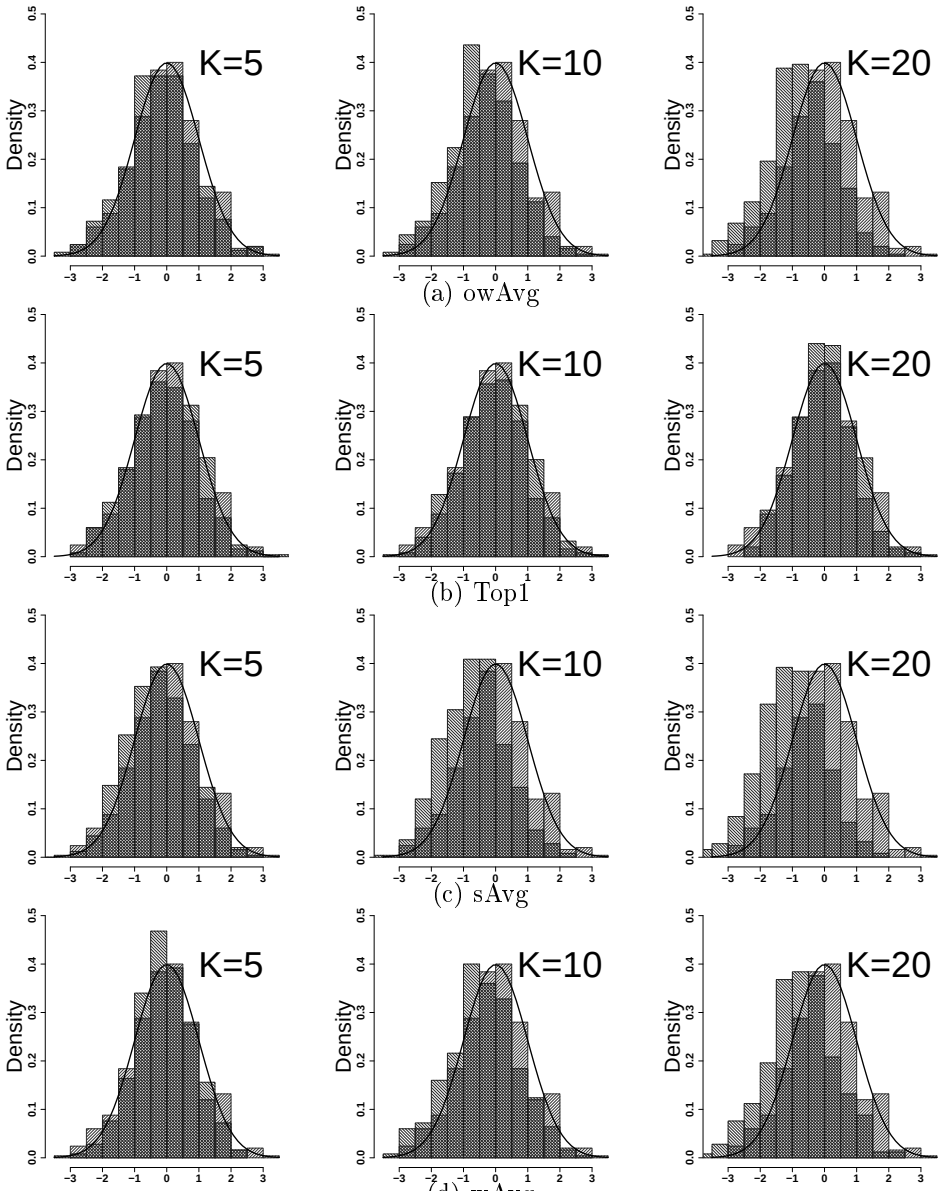
(a) owAvg

(b) Top1

(c) sAvg

(d) wAvg

Figure 2.7: Histogram of the normalized estimators for the pair $(a, b) = (1, 3)$ for the block diagonal graph structure when $n = 50000$ and $p = 1000$.



(a) owAvg

(b) Top1

(c) sAvg

(d) wAvg

Figure 2.8: Histogram of the normalized estimators for the pair $(a, b) = (1, 3)$ for the tridiagonal graph structure when $n = 50000$ and $p = 1000$.

CHAPTER 3

Directional false discovery rate control via debiased and distributed procedures in Gaussian graphical models

In this chapter, a multiple testing procedure for the conditional dependence in Gaussian graphical models is proposed. In practice, it is important to determine whether the conditional dependence between variables is positive or negative. However, there are several challenges to building statistics for testing using sample data. For instance, due to privacy concerns, one is not able to aggregate different datasets from several locations at one single location. In this chapter, different test statistics are constructed using debiased and distributed estimators to address this problem in a multiple testing framework. It is shown that, under mild conditions, the proposed procedure can control asymptotically the directional false discovery rate, which focuses on the sign of the estimation, at a pre-specified level. An asymptotic power equal to one is also attainable under mild conditions on the non-zero entries of the precision matrix. Different simulation scenarios and real data examples are used to investigate the performance of the proposed procedure and to confirm the theoretical results. The content of this chapter is based on the paper

> Nezakati, E. and E. Pircalabelu (2023a). Directional false discovery rate control via debiased and distributed procedures in Gaussian graphical models. pp. 1–34, *Submitted*.

3.1 Introduction

Gaussian graphical models (GGMs), which represent the conditional dependence relationships among a set of random variables through a graph structure, have many applications in the real world, including brain connectivity based on FMRI measurements (Ng et al., 2013), gene regulatory network discovery (Schäfer and

Strimmer, 2005), social network analysis, and many more. A large body of literature studied the estimation of GGMs, such as Yuan and Lin (2007), Friedman et al. (2008), Cai et al. (2011) and Ravikumar et al. (2011) among many others.

Considering $\mathbf{\Theta}_{ab} \neq 0$ as soon as the graphical Lasso estimates $\hat{\mathbf{\Theta}}_{ab} \neq 0$, i.e., considering naively just the estimated graph and precision matrix, is a deficient strategy leading to a high false discovery rate (FDR) value. Remember that on completely independent data in Section 2.6 the estimated precision matrix from the graphical Lasso contained more than 1600 falsely estimated edges. Generally this has to do with the presence of high-dimensional data and small sample sizes in many practical applications. In this chapter, we consider creating test statistics for the couple $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$, where $\mathbf{\Theta}_{ab}$ is the (a, b)-th entry of the inverse covariance matrix $\mathbf{\Theta}$ in GGMs, based on the transformations of the simple graphical Lasso estimators. The challenge is the fact that the distribution of the Lasso estimators is untractable. Liu (2013) considered inference for GGMs based on a nodewise regression approach and introduced a multiple testing procedure on partial correlations with false discovery proportion (FDP) and FDR control. Xia et al. (2015) studied simultaneous testing for differential networks. Recently, Li et al. (2022) constructed a multiple testing procedure based on a debiased CLIME estimator for edge detection with FDR control. They have shown theoretically and numerically that this procedure controls the FDR in multiple testing under a pre-specified level. In this work, we propose a procedure based on the debiased graphical Lasso estimator which controls the directional FDR as well as the FDR at a pre-specified level.

The directional FDR, which was first introduced by Gelman and Tuerlinckx (2000), is a useful metric for identifying type s errors (where "s" stands for sign). A type s error (also known as type III error) occurs when one confidently claims that a comparison goes one way, but it actually goes the other way. This metric is relevant for GGMs since by identifying the connections between the nodes, one is able to also state with confidence that their conditional dependence is positive or negative via the sign of the estimated partial correlation. In fact, partial correlation in GGMs measures the relationship between two variables while controlling for all other variables in the model. The sign of the partial correlation can provide valuable insights into the nature of the relationship between two variables. A positive partial correlation suggests a positive relationship between two variables, while a negative partial correlation suggests a negative relationship. However, estimating the sign of the partial correlation incorrectly can have important consequences for the interpretation of the model.

For example, suppose that one is modeling the relationship between various financial market indicators, including stock prices, interest rates, and inflation rates with GGMs. One estimates a positive sign edge between stock prices and inflation rates, while in reality, they have a negative relationship due to the well-known concept of inflation eroding the real value of stocks. This incorrect estimation could lead to a wrong investment decision, such that the investors might allocate

more funds to stocks during periods of high inflation, thinking it is a positive influence, leading to potential losses. Moreover, in genomics, researchers often use GGMs to model the relationships between genes based on their expression levels. If one estimates a negative sign edge between two genes involved in a biological pathway, while in reality, they have a positive regulatory relationship, this could lead to a wrong drug development, for which the pharmaceutical companies might target the wrong genes for drug development, thinking that inhibiting a particular gene will reduce the expression of another, while in fact, it increases it. These applications motivate us to investigate the control of type s error in GGMs.

By controlling the type s error in graphical models, one can control type I error as well as the error that occurs when estimating the sign. In the framework of regression, Zhao and Yu (2006) studied the sign consistency of Lasso based on the Irrepresentable condition. Jia et al. (2013) studied sign consistency of a sparse Poisson-like heteroscedastic model. Barber and Candès (2019) introduced a procedure using knockoff filters to control the directional FDR for high-dimensional linear models. Recently, Javanmard and Javadi (2019) and Cui et al. (2021) proposed different procedures to control the directional FDR under a pre-specified level using a debiased estimator for high-dimensional linear regression models with a random design and high-dimensional generalized linear models, respectively. In this chapter, our goal is to keep under control the directional FDR at a pre-specified level for Gaussian graphical models using a debiased estimator when data are distributed on multiple machines.

The process of collecting and analyzing all sample data at a single location for estimation and inference purposes is becoming challenging these days due to the growing size of the datasets, limited capacity of machines and security and privacy concerns. For example, suppose that one wants to estimate a genetic network for a specific disease using datasets that come from different hospitals. To maintain the privacy of patients' information, one cannot aggregate all the datasets at one location. To address these issues, distributed statistical approaches, which involve aggregating multiple estimators from different locations, have gained popularity in recent years for solving various statistical problems. The reader can refer to Chen and Xie (2014), Battey et al. (2018), Jordan et al. (2018) and Tang et al. (2020) for some relevant references on this topic. Nezakati and Pircalabelu (2023c) developed a new aggregated estimator for GGMs based on debiased local estimators which use non-overlapping parts of the dataset with unbalanced sizes. This paper has been explored extensively in Chapter 2 of this thesis. A similar approach has been explored in Nezakati and Pircalabelu (2023b) for covariate adjusted GGMs, and will be presented later in Chapter 4. However, to the best of our knowledge, the control of the directional FDR in the distributed setting has not yet been investigated.

The rest of the chapter is organized as follows. Notation and preliminaries are presented in Section 3.2. The debiased estimator using the full dataset and distributed estimators are reviewed in Section 3.3 and their properties are presented. A procedure for controlling the directional FDR is proposed in Section 3.4 based

on the debiased estimator using the full dataset and distributed estimators and their statistical guarantees are investigated. The performance of this procedure is evaluated via a simulation study and real data examples in Sections 3.5 and 3.6, respectively. We close with a discussion on the method and possible extensions in Section 3.7. Proofs of the supporting lemmas and theorems can be found in the Appendix.

3.2 Preliminaries

Consider the Gaussian graphical models framework, where a p-dimensional random vector $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top$ follows a multivariate Gaussian distribution with mean vector zero and covariance matrix $\mathbf{\Sigma}$. The components of $\dot{\mathbf{Y}}$ are mapped to the node set $\mathcal{V} = \{1, \ldots, p\}$ of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the edge set $\mathcal{E}$ describes the conditional dependence between every pair of components $Y^1, \ldots, Y^p$. A pair (a, b) is not included in the edge set $\mathcal{E}$ if and only if the variables Y^a and Y^b are independent given all remaining variables. Under the multivariate Gaussian distribution with positive definite covariance matrix, a pair of variables is conditionally independent given all remaining variables if and only if the corresponding entry in the precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is equal to zero. Elements of the precision matrix may thus be interpreted as the edge weights. Denote the maximum number of non-zero (active) upper off-diagonal entries of $\mathbf{\Theta}$ per row, i.e., the maximal node degree, with d and the index set of the non-zero upper diagonal entries of $\mathbf{\Theta}$ with S. Formally,

$$S = \{(a, b) \in \mathcal{V} \times \mathcal{V}, a < b : \quad \mathbf{\Theta}_{ab} \neq 0\},$$

having cardinality $s = \#S$, where $\mathbf{\Theta}_{ab}$ is the (a, b)-th element of $\mathbf{\Theta}$. The set of non-active components is denoted by S^c.

In this chapter, we propose a framework to select a set $\hat{S}$ of pairs of nodes, while controlling the directional false discovery rate ($\mathrm{FDR}_{\mathrm{dir}}$) for the selected pairs. We define $\mathrm{FDR}_{\mathrm{dir}}$ for a selected set $\hat{S}$ of pairs of nodes with estimated $\widehat{\mathrm{sign}}_{ab} \in \{-1, 1\}$ for the sign of $\mathbf{\Theta}_{ab}$ as

$$\mathrm{FDR}_{\mathrm{dir}} = \mathbb{E}[\mathrm{FDP}_{\mathrm{dir}}], \quad \mathrm{FDP}_{\mathrm{dir}} = \frac{\#\{(a, b) \in \hat{S} : \quad \widehat{\mathrm{sign}}_{ab} \neq \mathrm{sign}_{ab}\}}{\max\{\hat{s}, 1\}}, \tag{3.2.1}$$

where $\hat{s}$ denotes the cardinality of $\hat{S}$, i.e., $\hat{s} = \#\hat{S}$, $\mathrm{sign}_{ab} = \mathrm{sign}(\mathbf{\Theta}_{ab})$ and by convention $\mathrm{sign}(0) = 0$. Based on this definition, $\mathrm{FDR}_{\mathrm{dir}}$ represents the expected fraction of false discoveries among the selected edges, where a false discovery is measured with respect to type s and type I errors. For example, if $\widehat{\mathrm{sign}}_{ab} = +1$ while $\mathbf{\Theta}_{ab} = 0$, we consider that the procedure which selects $\hat{S}$ made a type I error. On the other hand, if $\widehat{\mathrm{sign}}_{ab} = +1$ while $\mathbf{\Theta}_{ab} < 0$, we consider that the procedure made a type s error. In both cases, it is considered as a false discovery. From (3.2.1) and

based on the definition of the classical FDR, i.e., $\text{FDR} = \mathbb{E}[\frac{\#\{(a,b)\in\hat{S}:\quad \boldsymbol{\Theta}_{ab}=0\}}{\max\{\hat{s},1\}}]$, it holds that $\text{FDR}_{\text{dir}} \geq \text{FDR}$. As such, by controlling the FDR_{dir}, one can provide a control on the FDR, too. It is noteworthy that controlling the FDR_{dir} is more challenging. The reason is that when one works with type I errors, the false discovery rate is built under the null hypothesis $H_{0,ab} : \boldsymbol{\Theta}_{ab} = 0$. In contrast, the directional FDR is not built under the null hypothesis and the element $\boldsymbol{\Theta}_{ab}$ might be nonzero with a positive or negative sign. Similarly, we define the directional power of a selected set $\hat{S}$ as

$$\text{Power}_{\text{dir}} = \mathbb{E}\bigg[\frac{\#\{(a,b)\in\hat{S}:\quad \widehat{\text{sign}}_{ab} = \text{sign}_{ab}\}}{\max\{s,1\}}\bigg]. \tag{3.2.2}$$

In this chapter, we provide a theoretical procedure, where we control the directional false discovery rate and we achieve asymptotic power equal to one using a debiased distributed estimator.

3.3 Debiased Lasso estimator

Consider i.i.d. samples $\dot{\mathbf{Y}}_1, \ldots, \dot{\mathbf{Y}}_n$ from $\dot{\mathbf{Y}}$, each of length p, and arrange them in matrix form denoted as $\mathbf{Y}$ of size $n \times p$. The goal is to control the false discovery rate for the procedure estimating the structure of the graph $\mathcal{G}$, or equivalently, finding the zero pattern of $\boldsymbol{\Theta}$. The usual assumption in this context is that the underlying graph is sparse, which is translated in imposing a certain bound on the maximal node degree (d) of the precision matrix. To impose this sparsity condition on the estimation, one common approach is to add an ℓ_1 penalty to the negative log-likelihood function. There exists a wide variety of methods making use of ℓ_1 regularization, see for example Friedman et al. (2008), Hsieh et al. (2014), Cai et al. (2016), among others. The graphical Lasso estimator $\hat{\boldsymbol{\Theta}}$, defined by Friedman et al. (2008), is the solution to the optimization problem

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}\in S^p_{++}} \Big\{\text{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_{1,\text{off}}\Big\}, \tag{3.3.1}$$

where $\hat{\boldsymbol{\Sigma}} = \mathbf{Y}^\top\mathbf{Y}/n$ is the sample covariance, $\lambda > 0$ is the penalty term, $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal norm of the matrix, defined as $\|\boldsymbol{\Theta}\|_{1,\text{off}} = \sum_{a\neq b}|\boldsymbol{\Theta}_{ab}|$ and S^p_{++} is the space of positive definite matrices of dimension $p \times p$. The graphical Lasso estimator (3.3.1) is biased due to the ℓ_1 penalty which is added to the negative log-likelihood function. Using the Karush-Kuhn-Tucker (KKT) condition (see for example, Ravikumar et al., 2011), and certain regularity assumptions (assumptions (A1)-(A2) in Chapter 2), Jankova and van de Geer (2015) proposed the debiased graphical Lasso estimator as

$$\hat{\boldsymbol{\Theta}}^d = 2\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}. \tag{3.3.2}$$

We call this estimator as "debiased full", since it uses the full dataset of size $n \times p$. Under regularity assumptions, (A1) and (A2) in Chapter 2, with $\lambda \asymp \sqrt{\log(p)/n}$ and the sparsity condition $d^{3/2} = o(\sqrt{n}/\log(p))$, it has been shown in Jankova and van de Geer (2015) that for every $(a, b) \in \mathcal{V} \times \mathcal{V}$,

$$\frac{\sqrt{n}}{\sigma_{ab}}(\hat{\mathbf{\Theta}}^d_{ab} - \mathbf{\Theta}_{ab}) = \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n} \mathbf{Z}_{ab,l} + \frac{\sqrt{n}}{\sigma_{ab}} \mathbf{R}_{ab}, \tag{3.3.3}$$

where $\frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n} \mathbf{Z}_{ab,l}$ converges weakly to $\mathcal{N}(0,1)$, and $|\mathbf{R}_{ab}| = o_p(1/\sqrt{n})$. Note that $\hat{\mathbf{\Theta}}^d_{ab}$ and $\mathbf{R}_{ab}$ are the (a,b)-th element of $\hat{\mathbf{\Theta}}^d$ and of a remainder term matrix, called $\mathbf{R}$, respectively. Moreover, $\sigma^2_{ab} = \mathbf{\Theta}_{aa}\mathbf{\Theta}_{bb} + \mathbf{\Theta}^2_{ab}$ and $\mathbf{Z}_{ab,l} = \mathbf{\Theta}_a^\top \dot{\mathbf{Y}}_l \dot{\mathbf{Y}}_l^\top \mathbf{\Theta}_b - \mathbf{\Theta}_{ab}$, where $\mathbf{\Theta}_a$ is the a-th column of $\mathbf{\Theta}$ and $\dot{\mathbf{Y}}_l = (Y_l^1, \ldots, Y_l^p)^\top$ is the p-dimensional vector of variables corresponding to the l-th sample, $l = 1, \ldots, n$, of $\dot{\mathbf{Y}}$. In Jankova and van de Geer (2015), it has been shown that

$$\|\hat{\mathbf{\Theta}}^d - \mathbf{\Theta}\|_\infty = O_p\Big(\max\{d\sqrt{\log(p)/n}, d^{3/2}\log(p)/n, d^2(\log(p)/n)^{3/2}\}\Big),$$

which is an $o_p(1)$ quantity under the mentioned sparsity assumption on d.

Now, consider K i.i.d. sub-samples from $\mathbf{Y}$ and denote the k-th sub-sample arranged in matrix form by $\mathbf{Y}_k$ with dimensions $n_k \times p$, $k = 1, \ldots, K$, such that $n_k/n \to c_k \in (0,1)$ as $n_k \to \infty$, and $\lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1$. Denote $n_\dagger = \min_{1\le k\le K} n_k$ as the minimum sub-sample size between all sub-samples. For any fixed K, one can use each sample in parallel, and using (3.3.2), we have access to K debiased estimators $\hat{\mathbf{\Theta}}^d_{ab,k}$, $k = 1, \ldots, K$, for every element $(a, b) \in \mathcal{V} \times \mathcal{V}$. Due to the asymptotic normal distribution of the local estimators, a final aggregated estimator has been derived in Nezakati and Pircalabelu (2023b) and Nezakati and Pircalabelu (2023c) by maximizing a pseudo log-likelihood function. This function is constructed using the asymptotic normal density of the local estimators $\hat{\mathbf{\Theta}}^d_{ab,k}$, $k = 1, \ldots, K$, and the aggregated estimator is of the form

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \hat{\mathbf{\Theta}}^d_{ab,k}. \tag{3.3.4}$$

The reader can refer to the mentioned references above and Chapter 2 in this manuscript for more details on the derivations and theoretical properties of the estimator in (3.3.4). Note that $\hat{\sigma}^2_{ab,k}$ in (3.3.4) is a consistent estimator of σ^2_{ab} obtained based on the k-th sample and is of the form $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k}$, where $\hat{\mathbf{\Theta}}_{ab,k}$ is the (a,b)-th element of the graphical Lasso estimator in (3.3.1). It has been shown in Nezakati and Pircalabelu (2023c) that for every $(a, b) \in \mathcal{V} \times \mathcal{V}$,

$$\sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \Big(\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} - \mathbf{\Theta}_{ab} \Big) = \mathbf{W}_{ab,\dagger} + \mathbf{R}_{ab,\dagger},$$

where $\mathbf{W}_{ab,\dagger} \xrightarrow{d} \mathcal{N}(0,1)$ and $|\mathbf{R}_{ab,\dagger}| = o_p(1)$, as K grows at the rate $K = O(n^{1/3}/(d\log(p)))$, under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $\lambda_k \asymp \sqrt{\log(p)/n_k}$, where $n_\dagger = \min_{1\leq k\leq K} n_k$. Note that, $\mathbf{R}_{ab,\dagger}$ is the (a,b)-th element of a remainder term matrix, called $\mathbf{R}_\dagger$. It is shown in Chapter 2 that

$$\|\tilde{\mathbf{\Theta}}_{\text{owAvg}} - \mathbf{\Theta}\|_\infty = O_p\left(K\max\left\{d\sqrt{\log(p)/n}, d^{3/2}\log(p)/n, d^2(\log(p))^{3/2}/(n\sqrt{n_\dagger})\right\}\right),$$

where $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ is the matrix form of the aggregated estimator with elements $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$, $(a,b) \in \mathcal{V}\times\mathcal{V}$. Using the triangle inequality, it is deduced that also $\|\tilde{\mathbf{\Theta}}_{\text{owAvg}} - \hat{\mathbf{\Theta}}^d\|_\infty$ is equal to the bound above. This implies that the convergence rate at which the distributed estimator approaches the full estimator $\hat{\mathbf{\Theta}}^d$ is the same convergence rate at which it approaches the true matrix $\mathbf{\Theta}$. As such, it is noteworthy that $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ approximates the debiased full estimator $\hat{\mathbf{\Theta}}^d$ well, and it exhibits a similar statistical error as $\hat{\mathbf{\Theta}}^d$ as long as the number of machines K is not too large.

In the next section, we provide a procedure using the debiased full and distributed estimators to control the directional false discovery rate.

3.4 Controlling the directional false discovery rate

In order to perform model selection for high-dimensional GGMs, we consider the multiple testing problem

$$H_{0,ab} : \mathbf{\Theta}_{ab} = 0 \quad \text{vs} \quad H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0, \quad 1 \leq a < b \leq p.$$

Here the $p(p-1)/2$ hypotheses are tested at the same time and it is equivalent to selecting the set $\hat{S}$ of estimated non-zero entries of $\mathbf{\Theta}$. By controlling the directional FDR, one not only cares about type I error, but also cares about the sign of the entries of the selected set $\hat{S}$ and whether they match with the sign of the true entries.

3.4.1 Directional FDR control via the debiased full estimator

To control the directional FDR using the debiased full estimator from (3.3.2), for every $1 \leq a < b \leq p$, we define the test statistic

$$\hat{T}_{ab} := \frac{\sqrt{n}\hat{\mathbf{\Theta}}^d_{ab}}{\hat{\sigma}_{ab}}, \tag{3.4.1}$$

where $\hat{\sigma}_{ab}$ is a consistent estimator for σ_{ab} defined as $\hat{\sigma}^2_{ab} = \hat{\mathbf{\Theta}}_{aa}\hat{\mathbf{\Theta}}_{bb} + \hat{\mathbf{\Theta}}^2_{ab}$ (for more details, see Lemma 2 of Jankova and van de Geer, 2015). For a given threshold $t > 0$, we reject $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ in favor of $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$, if $|\hat{T}_{ab}| \geq t$ and we

return the sign of $\hat{T}_{ab}$ as the estimated sign of $\mathbf{\Theta}_{ab}$. With this test statistic, the directional FDP at a given threshold t is defined as

$$\mathrm{FDP}_{\mathrm{dir}}(t) = \frac{\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\hat{T}_{ab} \leq -t) + \sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t)}{\max\{R(t), 1\}}, \tag{3.4.2}$$

where

$$\begin{aligned} S_{\leq 0} &= \{(a,b) \in \mathcal{V} \times \mathcal{V}, a < b : \quad \mathbf{\Theta}_{ab} \leq 0\}, \\ S_{\geq 0} &= \{(a,b) \in \mathcal{V} \times \mathcal{V}, a < b : \quad \mathbf{\Theta}_{ab} \geq 0\}. \end{aligned}$$

Moreover, $R(t) := \sum_{1\leq a<b\leq p} \mathbb{I}(|\hat{T}_{ab}| \geq t)$ is the total number of rejections at threshold t, with $\mathbb{I}(\cdot)$ being the indicator function. In words, the term $\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\hat{T}_{ab} \leq -t)$ in (3.4.2) counts the number of times that $\mathbf{\Theta}_{ab}$ is truly non-negative, but one mistakenly estimates its sign as negative. A similar argument holds for $\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t)$.

Define next the random variable

$$\tilde{T}_{ab} := \frac{1}{\sqrt{n}\hat{\sigma}_{ab}} \sum_{l=1}^{n} \mathbf{Z}_{ab,l} + \frac{\sqrt{n}}{\hat{\sigma}_{ab}} \mathbf{R}_{ab}. \tag{3.4.3}$$

Using the definition of $\hat{T}_{ab}$ in (3.4.1), it can be shown that $\hat{T}_{ab} = \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}} + \frac{\sqrt{n}}{\hat{\sigma}_{ab}}\mathbf{R}_{ab} + \frac{1}{\sqrt{n}\hat{\sigma}_{ab}} \sum_{l=1}^{n} \mathbf{Z}_{ab,l}$, with $\mathbf{R}_{ab}$ and $\mathbf{Z}_{ab,l}$ the same quantities as in (3.3.3). Then on the set $S_{\leq 0}$ we get $\hat{T}_{ab} \leq \tilde{T}_{ab}$ and as such, $\mathbb{I}(\hat{T}_{ab} \geq t) \leq \mathbb{I}(\tilde{T}_{ab} \geq t)$. On the other hand, on $S_{\geq 0}$, we have $\hat{T}_{ab} \geq \tilde{T}_{ab}$ and as such $\mathbb{I}(\hat{T}_{ab} \leq -t) \leq \mathbb{I}(\tilde{T}_{ab} \leq -t)$. As a result, in (3.4.2), we can write

$$\mathrm{FDP}_{\mathrm{dir}}(t) \leq \frac{\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\tilde{T}_{ab} \leq -t) + \sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\tilde{T}_{ab} \geq t)}{\max\{R(t), 1\}}. \tag{3.4.4}$$

In the sequel, we propose a procedure for which one can control the directional FDR by bounding the right hand side of (3.4.4) at a pre-specified level $0 < \alpha < 1$.

As it is shown in Lemma 3.4.1, the tail probability $\mathbb{P}(|\tilde{T}_{ab}| > t)$ can be approximated by $G(t)$ uniformly in $1 \leq a < b \leq p$, where $G(t) = 2(1 - \Phi(t))$ and $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. As such, the expected value of the numerator in (3.4.4) can be approximated by $p_1 G(t)$, where $p_1 := p(p-1)/2$. To find the threshold t for a pre-specified level $0 < \alpha < 1$, let $t_p = (2\log(p_1) - 2\log(\log(p_1)))^{1/2}$ and define

$$t_0 := \inf\{0 \leq t \leq t_p : \quad \frac{p_1 G(t)}{\max\{R(t), 1\}} \leq \alpha\}. \tag{3.4.5}$$

If the value in (3.4.5) does not exist, let $t_0 = \sqrt{2\log(p_1)}$. For every $1 \leq a < b \leq p$, we reject $H_{0,ab}$ if $|\hat{T}_{ab}| \geq t_0$ and we return $\widehat{\text{sign}}_{ab} = \text{sign}(\hat{T}_{ab})$ as the estimate for $\text{sign}(\boldsymbol{\Theta}_{ab})$.

To provide statistical guarantees on controlling the directional FDR using the defined procedure, we need to introduce two technical sets. For every $\gamma > 0$ and some constant $c_0 > 0$, define first the set

$$\Gamma(\gamma, c_0) := \{1 \leq a < b \leq p : \quad |\boldsymbol{\Theta}^0_{ab}| \geq c_0(\log(p))^{-1-\gamma/2}\}, \tag{3.4.6}$$

with cardinality $\Upsilon = \#\Gamma(\gamma, c_0)$, and define next the set

$$\Psi(\rho) := \{1 \leq a < b \leq p : \quad |\boldsymbol{\Theta}^0_{ab}| \geq \frac{1}{\sqrt{2}}(\frac{1-\rho}{1+\rho})\},$$

for every $\rho \in (0, 1)$, with cardinality $\psi = \#\Psi(\rho)$, where $\boldsymbol{\Theta}^0_{ab}$ is the normalized form of $\boldsymbol{\Theta}_{ab}$, defined as $\boldsymbol{\Theta}^0_{ab} = \boldsymbol{\Theta}_{ab}/\sqrt{\boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}}$. We call the set $\Gamma(\gamma, c_0)$ as the set of "highly correlated pair nodes", and its complement set, which is denoted by $\Gamma^c(\gamma, c_0)$, as the set of "weakly correlated pair nodes". These two sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$, involving the elements of the precision matrix $\boldsymbol{\Theta}$, are similar to the sets defined in Javanmard and Javadi (2019), are needed to guarantee the control of the directional FDR in GGMs, and are used explicitly in the proofs of Lemma 3.4.1 and Theorem 3.4.2. More precisely, the proof of the main theorems in this chapter are constructed using Lemmas 6.1 and 6.2 of Liu (2013), wherein the covariance between specific random vectors should be bounded. To fulfill those conditions, we need the two technical sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$ with specific bounds on the entries $\boldsymbol{\Theta}^0_{ab}$. Note that these two sets are not disjoint, and at some point, we need to define a set difference between these two sets. For more details, the reader can refer to the proof of Lemma 3.9.1 in the Appendix B.

Since the threshold t_0 tends typically to infinity with a growing number of nodes p, to control the directional FDP in (3.4.4), the elementwise asymptotic normality result in (3.3.3) is not enough to provide statistical guarantees for the procedure. As such we use a stronger, Cramér-type moderate deviation result similar to Liu (2013) which provides guarantees uniformly over pairs (a, b). This is provided in Lemma 3.4.1.

Lemma 3.4.1. Let $p \leq n^r$ for some $r > 0$, and suppose that assumptions (A1)-(A2) in Chapter 2 hold. Consider the random variable $\tilde{T}_{ab}$ defined in (3.4.3), and denote by

$$T_{ab} := \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n} \mathbf{Z}_{ab,l}, \tag{3.4.7}$$

for every $1 \leq a < b \leq p$, where $\mathbf{Z}_{ab,l}$ is defined in (3.3.3). Under the sparsity condition $d^{3/2} = o(\sqrt{n}/(\log(p))^{3/2})$, and considering the set $\Gamma(\gamma, c_0)$ with cardinality

$\Upsilon = o(p^{\delta\rho})$, $\delta \in (0, 1/2]$, $\rho \in (0, 1)$, we have

$$\max_{a,b} |\tilde{T}_{ab} - T_{ab}| = o_p(1/\sqrt{\log(p)}) \tag{3.4.8}$$

and

$$\max_{a,b} \sup_{0 \le t \le t_p} \left| \frac{\mathbb{P}(|T_{ab}| > t)}{G(t)} - 1 \right| \to 0.$$

The proof is given in Appendix 3.8.1.

Theorem 3.4.2 next provides statistical guarantees for the control of the directional FDR using the threshold introduced in (3.4.5).

Theorem 3.4.2. *Let $p \le n^r$ for some $r > 0$, and suppose that assumptions (A1)-(A2) in Chapter 2 hold. Consider the test statistic $\hat{T}_{ab}$ defined in (3.4.1) based on the debiased estimator from (3.3.2). Under the sparsity condition $d^{3/2} = o(\sqrt{n}/(\log(p))^{3/2})$, and considering the sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$ with cardinalities $\Upsilon = o(p^{\delta\rho})$, $\rho \in (0, 1)$, $\delta \in (0, 1/2]$ and $\psi = O(p^{\delta\rho}/\sqrt{\log(p)})$, respectively, for the threshold t_0 defined in (3.4.5) and every $\epsilon > 0$, we have*

$$\limsup_{(n,p)\to\infty} \mathrm{FDR}_{\mathrm{dir}} \le \alpha, \quad \textit{and} \quad \lim_{(n,p)\to\infty} \mathbb{P}(\mathrm{FDP}_{\mathrm{dir}} \le \alpha + \epsilon) = 1. \tag{3.4.9}$$

The proof is given in Appendix 3.8.2.

Comparing the sparsity condition of Theorem 3.4.2 with that of Jankova and van de Geer (2015) for the pointwise asymptotic normality of the test statistic, it is observed that the number of non-zero entries per row of the precision matrix $\boldsymbol{\Theta}$ should grow at a slower rate in this case, such that the matrix be sparser for the proposed procedure to provide control on the directional FDR. Moreover, the conditions on the cardinality of two sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$ need that the test statistics should have mild correlations for the directional FDR to be under control. In the sequel, Theorem 3.4.3 provides guarantees for the statistical power of the proposed procedure.

Theorem 3.4.3. *Suppose that the conditions of Theorem 3.4.2 hold. If $|\boldsymbol{\Theta}_{ab}| > \sqrt{\frac{2\sigma_{ab}^2 \log(p_1/s)}{n}}$ for $(a, b) \in S$, where $p_1 = p(p-1)/2$, $s = \#S$ the cardinality of the support of $\boldsymbol{\Theta}$, and $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$, then we have*

$$\liminf_{n\to\infty} \frac{\mathrm{Power}_{\mathrm{dir}}}{\frac{1}{s}\sum_{(a,b)\in S} \left(1 - \Phi(\Phi^{-1}(1 - \frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\boldsymbol{\Theta}_{ab}|}{\sigma_{ab}})\right)} \ge 1, \tag{3.4.10}$$

where α is the pre-specified level in the hypotheses testing procedure. Moreover, if $\frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{\max}} - \Phi^{-1}(1 - \frac{\alpha s}{2p_1}) \to \infty$, where $\boldsymbol{\Theta}_{\min} = \min_{(a,b)\in S}|\boldsymbol{\Theta}_{ab}|$ and $\sigma_{\max} = \max_{(a,b)\in S}\sigma_{ab}$, then $\mathrm{Power}_{\mathrm{dir}} \to 1$.

The proof is given in Appendix 3.8.3.

In the next section, we show that the directional FDR can still be maintained under control using the distributed estimator.

3.4.2 Directional FDR control via the distributed estimator

Similarly to (3.4.1), one can construct the test statistic

$$\hat{U}_{ab} = \sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}}\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}, \tag{3.4.11}$$

for every $1 \leq a < b \leq p$, based on the distributed estimator. This statistic will be used for the hypothesis test $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$. Using this test statistic, one can reject $H_{0,ab}$ in favor of $H_{1,ab}$ if $|\hat{U}_{ab}| \geq t$ for a given threshold t. As such, the directional FDP at a given threshold t based on the statistic defined in (3.4.11) will be similar to (3.4.2), by replacing $\hat{T}_{ab}$ with $\hat{U}_{ab}$, that is

$$\text{FDP}_{\text{dir}}(t) = \frac{\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\hat{U}_{ab} \leq -t) + \sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{U}_{ab} \geq t)}{\max\{R'(t), 1\}},$$

with $R'(t) := \sum_{1\leq a<b\leq p} \mathbb{I}(|\hat{U}_{ab}| \geq t)$. Similarly to (3.4.3), by defining the random variable

$$\tilde{U}_{ab} = \mathbf{W}_{ab,\dagger} + \mathbf{R}_{ab,\dagger}, \tag{3.4.12}$$

one can show that

$$\text{FDP}_{\text{dir}}(t) \leq \frac{\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\tilde{U}_{ab} \leq -t) + \sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\tilde{U}_{ab} \geq t)}{\max\{R'(t), 1\}}.$$

Lemma 3.4.4 shows that the difference between $\tilde{U}_{ab}$ and T_{ab} is negligible, and as such, using Lemma 3.4.1, the tail probability $\mathbb{P}(|\tilde{U}_{ab}| > t)$ can be approximated by $G(t)$. As a result, for a pre-specified level $0 < \alpha < 1$, the threshold t'_0 can be defined as

$$t'_0 := \inf\{0 \leq t \leq t_p : \quad \frac{p_1 G(t)}{\max\{R'(t), 1\}} \leq \alpha\}, \tag{3.4.13}$$

where relative to (3.4.5), we substitute the total number of rejections $R(t)$ by $R'(t)$ based on the distributed test statistic $\hat{U}_{ab}$ from (3.4.11).

Lemma 3.4.4. Let $p \leq n^r$ for some $r > 0$, and suppose that assumptions (A1)-(A2) in Chapter 2 hold. Consider the random variable $\tilde{U}_{ab}$ defined in (3.4.12) for every $1 \leq a < b \leq p$. Then

$$\max_{a,b} |\tilde{U}_{ab} - T_{ab}| = O_p\bigg(\frac{K}{\sqrt{n}} \max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}, Kd\log(p)\}\bigg),$$

which is $o_p(1)$ as K grows at the rate $K = O(n^{1/4}/(d\log(p)))$ with sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/(\log(p))^{3/2})$.

The proof is given in Appendix 3.8.4.

Theorem 3.4.5 provides guarantees for the control over the directional FDR using the threshold introduced in (3.4.13).

Theorem 3.4.5. *Let $p \leq n^r$ for some $r > 0$, and suppose that assumptions* (A1)-(A2) *in Chapter 2 hold. Consider the test statistic $\hat{U}_{ab}$ defined in* (3.4.11) *based on the distributed estimator. Suppose that the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/(\log(p))^{3/2})$ holds, and consider the sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$ with cardinalities $\Upsilon = o(p^{\delta\rho})$, $\rho \in (0,1)$, $\delta \in (0, 1/2]$ and $\psi = O(p^{\delta\rho}/\sqrt{\log(p)})$, respectively. As long as K grows at the rate $K = O(n^{1/4}/(d\log(p)))$, the threshold defined in* (3.4.13) *based on the test statistic $\hat{U}_{ab}$ controls the directional* FDR *at level α, i.e., for every $\epsilon > 0$,*

$$\limsup_{(n,p)\to\infty} \mathrm{FDR}_{\mathrm{dir}} \leq \alpha, \quad \textit{and} \quad \lim_{(n,p)\to\infty} \mathbb{P}(\mathrm{FDP}_{\mathrm{dir}} \leq \alpha + \epsilon) = 1.$$

The proof is given in Appendix 3.8.5.

Comparing the upper bound provided for the growing number of machines K in Theorem 3.4.4 with that of Nezakati and Pircalabelu (2023c) for the pointwise asymptotic normality of the distributed test statistic, it is observed that K should grow at a slower rate here, in order for the procedure to provide a control on the directional FDR. The rate in Chapter 2 is too large for our purposes and using that rate does not guarantee $o_p(\cdot)$ results, reason for which we use a slightly smaller term in the bound for K. Furthermore, the conditions on the cardinality of the two sets $\Gamma(\gamma, c_0)$ and $\Psi(\rho)$ are the same as in Theorem 3.4.2. These conditions are different, but similar, to those of Javanmard and Javadi (2019) for controlling the directional FDR for hypothesis testing in high-dimensional regression models.

Remark 3.1. By the same technique as in Theorem 3.4.3 and the provided assumptions on the entries of $\boldsymbol{\Theta}_{ab}$, under the conditions of Theorem 3.4.5, it can be shown that the directional power of the proposed distributed procedure tends to 1 with increasing n.

In the next section, we investigate the performance of the proposed procedure via a simulation study.

3.5 Simulation study

In this section, we analyze the performance of the proposed procedure in controlling the FDR by conducting a simulation study. Two simulation settings are considered for this purpose. In the first setting, we compare the performance of the statistic defined in (3.4.1), which is based on the debiased full estimator, with that of the other existing methods. In the second setting, we compare the performance of the distributed test statistic defined in (3.4.11) with that of the debiased full estimator.

To generate a dataset, we simulated n samples from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, such that $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is graph-structured and has the following sparse structures

(i) random with probability of connection 0.05;

(ii) block diagonal with 10 blocks and 0.2 as the probability of connection in each block;

(iii) chain or tridiagonal with 1 on the diagonal and 0.2 on the upper and lower diagonals.

All the simulation results are empirical averages over 100 replications from the specified DGPs. Consider $\hat{S}_r = \{(a,b) \in \mathcal{V} \times \mathcal{V}, a < b : \quad |\hat{T}_{ab,r}| \geq t_{0,r}\}$ as the set of selected pairs (a,b) at the r-th replication, using the proposed procedure, where $\hat{T}_{ab,r}$ and $t_{0,r}$ are the test statistic defined in (3.4.1) and the threshold defined in (3.4.5), respectively, for the r-th replication. The empirical FDP for the r-th replication, $r = 1, \ldots, 100$, is defined as

$$\text{FDP}_r = \frac{\sum_{(a,b)\in \hat{S}_r} \mathbb{I}(\mathbf{\Theta}_{ab} = 0)}{\max\{\hat{s}_r, 1\}},$$

where $\hat{s}_r$ is the cardinality of the selected set $\hat{S}_r$, i.e., $\hat{s}_r = \#\hat{S}_r$. Then, the empirical FDR over 100 replications, is of the form $\text{FDR} = (1/100)\sum_{r=1}^{100} \text{FDP}_r$. Moreover, the empirical power is calculated as

$$\text{Power} = \frac{1}{100}\sum_{r=1}^{100}\left\{\frac{\sum_{(a,b)\in \hat{S}_r} \mathbb{I}(\mathbf{\Theta}_{ab} \neq 0)}{\max\{s, 1\}}\right\},$$

where s is the cardinality of the set S, which is the set of active components of $\mathbf{\Theta}$. Similar quantities can be defined based on different test statistics such as the distributed one.

3.5.1 Performance of the debiased full test statistic

To compare the performance of the threshold introduced in (3.4.5) based on the debiased full estimator defined in (3.3.2) with other competitors, we considered the procedures of Li et al. (2022) and Liu (2013) based on the debiased Clime estimator and the nodewise Lasso procedure, respectively. We set the number of variables (nodes) to $p = 200$ and the sample size to $n = 150, 200, 250$ and 2000. The regularization parameter is considered equal to $\lambda = \sqrt{\log(p)/n}$ for all procedures and the results are presented in Table 3.1. The proposed statistic based on the debiased graphical Lasso and the ones from Li et al. (2022) and Liu (2013) have been denoted by "dGL", "dCl" and "GFC", respectively in this table.

It is observed that dGL and dCl can control FDR at the pre-specified levels $\alpha = 0.1$ and 0.2 for all three scenarios. However, the results for dGL are slightly

better than those based on the dCl test statistic. The FDR results for the GFC procedure are high, especially for the random graph structure. For this structure, when n is small, GFC is not able to control FDR at the pre-specified levels, but as n increases, it reaches eventually the pre-specified levels 0.1 and 0.2. However, its power is the highest compared to the two other methods. Moreover, it is observed that the powers when using model (ii) as DGP are the highest relative to those obtained when using models (i) and (iii). The reason is that the non-zero entries in the true precision matrix $\mathbf{\Theta}$ have the largest magnitude in model (ii). The results for the directional FDR corresponding to the dGL test statistic were also similar to the ones presented in Table 3.1, just slightly higher than the reported ones. However, since dCl and GFC do not provide any procedure for controlling the directional FDR, we did not report these results here. The FDPs over 100 replications are also plotted in Figure 3.1 for the random graph structure, with different values of n with $p = 200$ at the pre-specified level $\alpha = 0.1$. It is observed that most FDPs corresponding to the dGL and dCl procedures are concentrated around the FDRs and under the pre-specified level. Moreover, the median of the dGL is smaller than that of the dCl. The FDP values corresponding to the GFC are larger and are not under the pre-specified level 0.1. However, with increasing n, they are getting closer to 0.1.

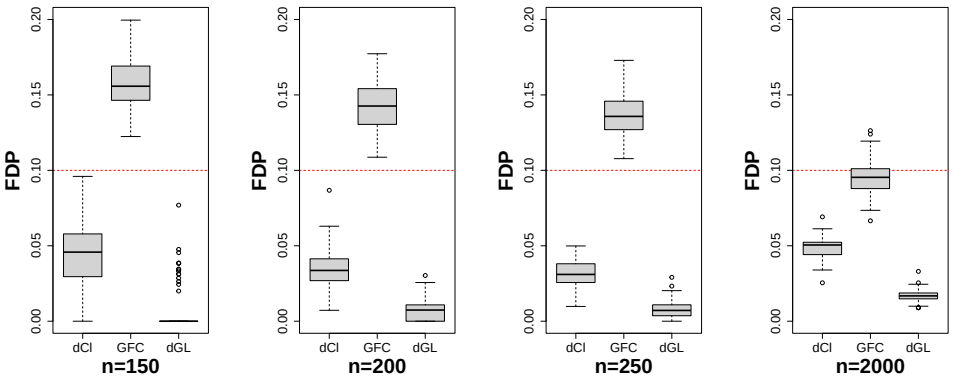


Figure 3.1: Boxplots of FDPs over 100 replications for the random graph structure with $p = 200$, at the pre-specified level $\alpha = 0.1$.

3.5.2 Performance of the distributed test statistic

In this section, we investigate the performance of the distributed test statistic and compare it with that of the full ones in controlling the FDR. To this end, we set the total sample size to $n = 50000$, the number of nodes to $p = 1000$ and changed the number of machines from $K = 5, 10$ to 20. To show the performance

Table 3.1: Empirical false discovery rate and power for the debiased full test statistic and two competitors, when $p = 200$.

			$n = 150$			$n = 200$			$n = 250$			$n = 2000$		
			dGL	dCl	GFC	dGL	dCl	GFC	dGL	dCl	GFC	dGL	dCl	GFC
$\alpha = 0.1$	Random	FDR	.00	.05	.16	.01	.04	.14	.01	.03	.14	.02	.05	.10
		Power	.03	.11	.37	.13	.25	.54	.29	.40	.67	1	1	1
	Block diag.	FDR	.01	.04	.10	.01	.04	.09	.01	.04	.09	.02	.06	.09
		Power	.38	.49	.65	.62	.71	.82	.78	.85	.91	1	1	1
	Tridiag.	FDR	.03	.05	.11	.02	.08	.10	.03	.08	.10	.06	.08	.09
		Power	.01	.05	.10	.12	.23	.25	.28	.40	.42	1	1	1
$\alpha = 0.2$	Random	FDR	.01	.09	.28	.02	.07	.26	.02	.07	.26	.04	.10	.20
		Power	.06	.19	.51	.22	.35	.66	.39	.51	.78	1	1	1
	Block diag.	FDR	.02	.08	.20	.02	.08	.19	.02	.08	.19	.05	.13	.19
		Power	.46	.58	.74	.69	.78	.87	.83	.89	.94	1	1	1
	Tridiag.	FDR	.04	.16	.23	.06	.16	.20	.06	.16	.19	.12	.17	.19
		Power	.05	.15	.18	.19	.33	.36	.37	.51	.53	1	1	1

of the distributed estimator in the unbalanced setting, we considered the following unbalanced data splitting procedure similar to Nezakati and Pircalabelu (2023c).

Suppose that among all available machines, two of them are powerful. The first one is the most powerful one and $(55 - K)\%$ of the dataset is distributed on this machine. The second one is less powerful than the first one and $(60 - K)\%$ of the remaining dataset is distributed on this one. The remaining dataset is distributed roughly equally on the remaining machines.

In this study, the regularization parameter λ_k in the k-th graphical Lasso problem has been set to $\lambda_k = \lambda = \sqrt{\log(p)/n}$ for all simulation settings except for the tridiagonal case. Due to the high sparsity of the precision matrix in the tridiagonal structure, we increased slightly the regularization parameter to $\lambda_k = \sqrt{\log(p)/n_k}$. The empirical false discovery rate results are shown in Table 3.2. All the powers were equal to 1 and they have not been presented in the table. Similarly to the previous simulation study, the results for the directional FDR corresponding to the distributed test statistic were slightly higher than the false discovery rates presented in Table 3.2 and have not been shown in this table.

From Table 3.2, it is observed that all the false discovery rates in the distributed case are under control at the pre-specified levels $\alpha = 0.1$ and 0.2. Moreover, the results are close to those of the debiased full one and they confirm that one does not lose much information when the dataset is distributed on different locations and estimators are aggregated. The results concerning the FDPs for the three graph structures with $\alpha = 0.1$ over 100 replications are shown in Figure 3.2 for dCl, GFC, dGL and the distributed procedure when $K = 20$. It is observed that all empirical FDPs for dGL and $K = 20$ are under the pre-specified level $\alpha = 0.1$. However, in the tridiagonal structure, GFC and dCl have values larger than 0.1 for some replications.

As it is mentioned in Theorems 3.4.2 and 3.4.4, the theoretical guarantee on controlling the directional FDR depends on several assumptions e.g. the sparsity and the magnitude of the entries of $\mathbf{\Theta}$. To investigate the effect of the sparsity on the directional FDR, we considered the random graph structure and changed the probability of connection between every pair of nodes from 0.01 to 0.1. The results are presented in the left panel of Figure 3.3 for the dGL and the distributed procedure with $K = 5, 10$ and 20. It is observed that the results are not sensitive to the probability of connection, as they are stable in the interval $(0.02, 0.06)$ and under the pre-specified level $\alpha = 0.1$.

Moreover, to investigate the effect of the magnitude of the off-diagonal entries on the directional FDR, we considered different random graph structures, such that the magnitude of the largest off-diagonal entry changes from 0.2 to 6.5. The results are presented in the right panel of Figure 3.3. It is observed that as the magnitude of the off-diagonal entries of $\mathbf{\Theta}$ increases, the directional FDR increases for both debiased full and distributed procedures. This confirms that the cardinality of the highly correlated pair nodes should be under control, such that most of the test statistics have mild correlation as only in this case it provides a guarantee on

Table 3.2: Empirical false discovery rate for the distributed and debiased full test statistics and two competitors, when $n = 50000$ and $p = 1000$.

		$\alpha = 0.1$				$\alpha = 0.2$			
				K				K	
		1	5	10	20	1	5	10	20
Random	dGL	.03				.08			
	dCl	.07				.14			
	GFC	.09				.19			
	distributed		.04	.04	.05		.08	.09	.11
Block diag.	dGL	.04				.09			
	dCl	.07				.15			
	GFC	.10				.19			
	distributed		.04	.05	.07		.09	.10	.14
Tridiag.	dGL	.08				.16			
	dCl	.09				.18			
	GFC	.10				.19			
	distributed		.05	.04	.03		.11	.09	.07

controlling the directional FDR at the pre-specified level α.

3.6 Real data

In this section, we present two real data examples that illustrate the performance of our methodology.

3.6.1 Text dataset

The publicly available "4 university web-pages" dataset used in Chapter 2 has also been analyzed here. For a description we refer to Section 2.6 of Chapter 2. The dataset contained $n = 4199$ and $p = 500$ terms. We applied dGL, dCl, GFC, and the distributed procedure with $K = 3, 4$ and 5 to identify the connections between 500 distinct terms. The splitting setting is considered the same as in the simulation study. The regularization parameter and the pre-specified level are set to $\lambda = \sqrt{\log(p)/n}$ and $\alpha = 0.1$, respectively, for all procedures. The proportion of detected edges, which is obtained as the percentage of the number of detected edges divided by $p(p-1)/2$, is presented in the left panel of Figure 3.4. It is observed that the proportion of detected edges is low, meaning that the estimated graph is sparse. The distributed procedure identified the sparsest graph, while the graph identified by GFC is quite dense. The proportion of detected edges using dGL and the distributed procedure are close to each other, which confirms the similarity of

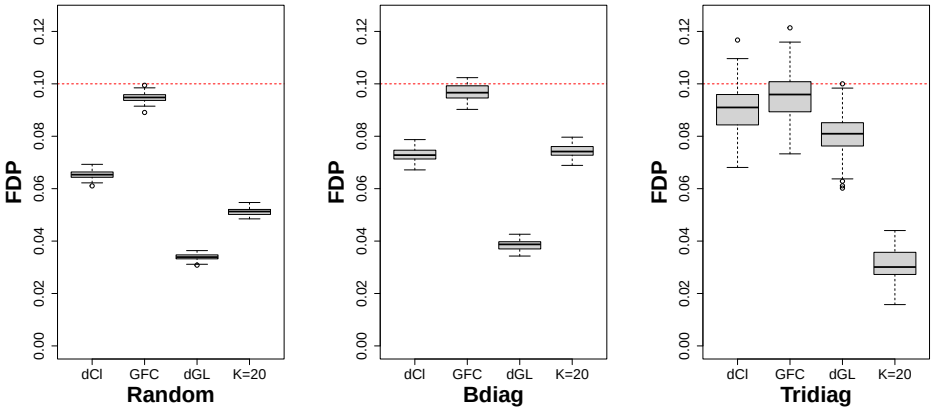


Figure 3.2: Boxplots of FDPs over 100 replications for three graph structures with $p = 1000$ and $n = 50000$, at the pre-specified level $\alpha = 0.1$.

these two procedures.

To evaluate the performance of the estimated graphs, we computed the out-of-sample prediction error using the negative log-likelihood. To this end, we sampled 10 percent of the dataset as the test set and the remaining dataset was used as the training set. After estimating the precision matrix using the training dataset by the proposed hypothesis testing procedures, denoted as $\check{\mathbf{\Theta}}$, we symmetrized it and computed the positive definite projection of $\check{\mathbf{\Theta}}$, denoted as $\check{\mathbf{\Theta}}_+$. Similarly to Li et al. (2022), the out-of-sample prediction error of $\check{\mathbf{\Theta}}_+$ is evaluated via

$$\text{PredErr}(\check{\mathbf{\Theta}}_+) = \frac{1}{2p}\Big\{\text{trace}(\hat{\mathbf{\Sigma}}_{\text{test}}\check{\mathbf{\Theta}}_+) - \log\det(\check{\mathbf{\Theta}}_+)\Big\},$$

where $\hat{\mathbf{\Sigma}}_{\text{test}}$ is the sample covariance using the test dataset. The average over 10 replications of the out-of-sample prediction errors is presented in the upper panel in Table 3.3. It is observed that all prediction errors are close to each other, indicating that the estimated graphs and estimated precision matrices are relatively similar. As the GFC procedure does not provide any estimates for the diagonal entries of the precision matrix, it is not possible to compute this prediction error for it.

Afterwards, to investigate the performance of the estimators for completely independent variables, we permuted randomly sample data for each variable, thus breaking up the correlation structure in order to construct a dataset with mutually independent variables. As such, we expect that there are no edges in the estimated graph. By applying the mentioned procedures on this dataset, we identified no edges in the estimated graph which confirms this assertion.

The first ten terms with the highest node degrees, identified by the proposed procedures, are presented in Figure 3.5. Some terms such as “project” and “memori”

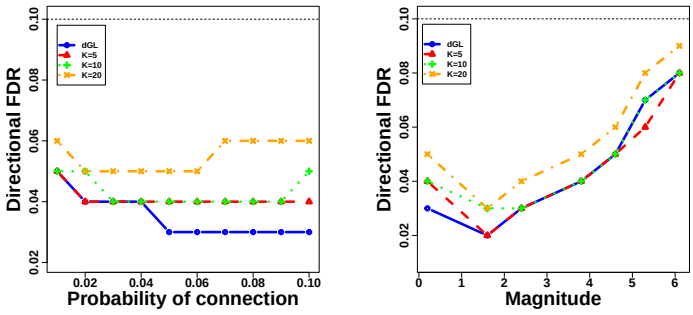


Figure 3.3: The left panel presents FDR_{dir} for the debiased full procedure labeled dGL and the distributed procedure with $K = 5, 10, 20$ as a function of the probability of connection in the random graph structure. The right panel presents FDR_{dir} for the mentioned procedures as a function of the magnitude of the off-diagonal entries of $\boldsymbol{\Theta}$. For both panels, $n = 50000$, $p = 1000$ and the pre-specified level $\alpha = 0.1$ are considered.

are identified by almost all the procedures. Based on the node degrees provided in this figure, it is confirmed that the estimated graph by GFC is quite dense compared to the estimated graph by the other procedures. Comparing the result of this figure for $K = 5$ with the one from Figure 2.5 of Chapter 2, it is observed that some words such as "course", "ieee", "award" and "algorithm" are common between two graphs. However, there are some differences in the node degree of two graphs. The difference can be attributed to the fact that the graphical Lasso with Bonferroni correction is much sparser than the graphical Lasso with FDR control procedure.

3.6.2 Pan-Cancer dataset

The process of collecting the Pan-Cancer dataset from The Cancer Genome Atlas (TCGA) project (available at https://xenabrowser.net/datapages/) was started in 2006 and in 10 years time, TCGA network investigators had characterized the molecular landscape of tumors from more than 11000 patients across 33 cancer types. This particular TCGA molecular dataset has been studied in multiple works to understand the cancer biology, including those of glioblastoma multiforme (GBM), ovarian, breast, lung, prostate, bladder and others (see for instance, Akbani et al., 2015; Cancer Genome Atlas Research Network, 2017; Liu et al., 2018).

We applied the proposed methodology on 743 microRNA mature strand expressions to estimate the underlying graph among the microRNAs. The final dataset

Table 3.3: Out-of-sample prediction error of the proposed procedure

	Text dataset					
	dCl	GFC	dGL	$K = 3$	$K = 4$	$K = 5$
PredErr	.43	/	.42	.44	.45	.46
	Pan-cancer dataset					
	dCl	GFC	dGL	$K = 3$	$K = 5$	$K = 10$
PredErr	.27	/	-.16	-.15	-.14	-.11

consists of $n = 9986$ subjects having $p = 743$ nodes. The splitting setting is considered the same as in the simulation study and the number of machines is set to $K = 3, 5$ and 10. The regularization parameter and the pre-specified level are set to $\lambda = \sqrt{\log(p)/n}$ and $\alpha = 0.05$, respectively, for all procedures. The proportion of detected edges is presented in the right panel of Figure 3.4. It is observed that the estimated graph by the distributed, dGL and dCl methods are sparse, however the distributed procedure identified the sparsest graph. The estimated graph by the GFC method is much denser than the other presented methods. Moreover, it is observed in the lower panel of Table 3.3 that the dGL has the smallest prediction error, while dCl has the largest one. The prediction errors of the distributed method are also close to those of the debiased full one which confirms the similarity of these two procedures. Note that, after applying the mentioned procedures on the randomly permuted sample data, we identified no edges between the genes.

As it is mentioned in Wang (2015), the genes hsa-miR-136, hsa-miR-376a and hsa-miR-377 are the most important genes in identifying GBM cancer. Figure 3.6 presents the estimated sub-graph between these genes using the proposed procedures at the pre-specified level $\alpha = 0.1$. As it is observed, dGL and the distributed procedure with $K = 3, 5$ and 10 identified the same edges between these genes, while dCl and GFC identified a few more edges.

3.7 Conclusion

In this chapter, we studied the directional false discovery rate in a multiple testing framework for GGMs. First, a test statistic has been built using the debiased graphical Lasso estimator and it has been shown that as long as the maximum node degree of the precision matrix grows at the rate $d^{3/2} = o(\sqrt{n}/(\log(p))^{3/2})$, and the cardinality of the highly correlated pair nodes is under control, the directional FDR will be asymptotically below a pre-specified level.

Furthermore, related to modern problems such as data privacy, heterogeneous different datasets and parallelization, a distributed test statistic has been proposed to control the directional FDR in multiple testing. It has been shown that as

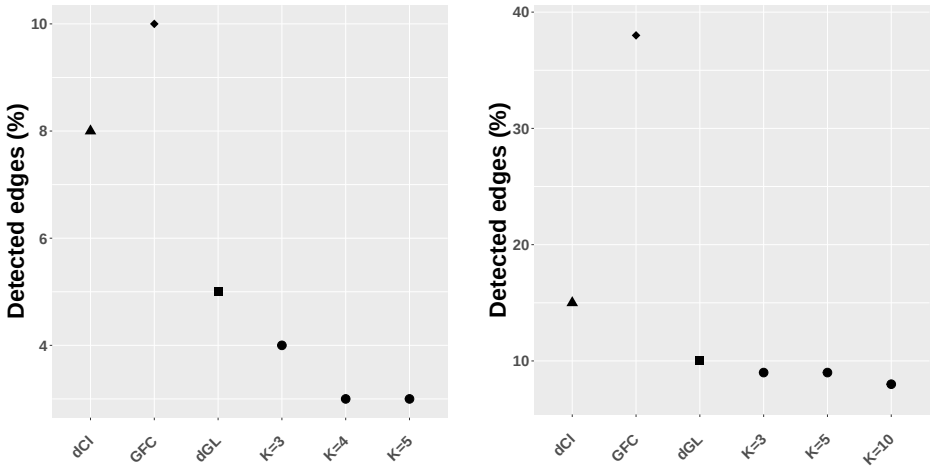


Figure 3.4: The left panel presents the proportion of detected edges for the text dataset using dCl, GFC, dGL and the distributed procedure with $K = 3, 4$ and 5, at the pre-specified level $\alpha = 0.1$. The right panel presents this proportion for the Pan-Cancer dataset, at the pre-specified level $\alpha = 0.05$.

long as the number of multiple datasets grows at the rate $K = O(n^{1/4}/(d\log(p)))$ with the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/(\log(p))^{3/2})$ and the cardinality of highly correlated pair nodes is under control, the directional FDR in the multiple testing of GGMs is asymptotically under a pre-specified level.

The finite sample performance of the proposed procedures has also been evaluated with a simulation study and real data examples. It is observed that both the debiased and distributed procedures have competitive results relative to the state-of-the-art methods. The FDR, as well as the directional FDR for both procedures are under the pre-specified level. This confirms the assertion that in practice, having datasets distributed across multiple machines, in our aggregation framework one does not lose much information and the performance is close to that of the debiased full procedure which would use the entire dataset, but which is unattainable in practice.

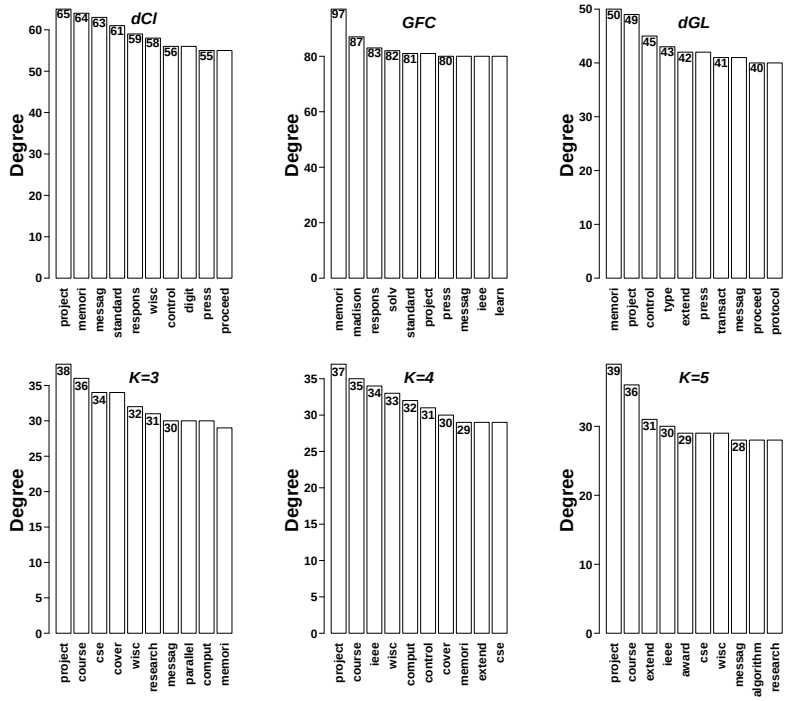


Figure 3.5: The first 10 terms with the highest node degrees, identified by different procedures, at the pre-specified level $\alpha = 0.1$.

3.8 Appendix A: Proofs of the main theorems and lemmas

3.8.1 Proof of Lemma 3.4.1

Consider $T_{ab} = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \frac{\mathbf{Z}_{ab,l}}{\sigma_{ab}}$, such that $\mathbf{Z}_{ab,l} = \boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_l \dot{\mathbf{Y}}_l^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$. It is shown in Jankova and van de Geer (2015) that under the multivariate Gaussian distribution, $\mathbb{V}\mathrm{ar}(\mathbf{Z}_{ab,l}) = \sigma_{ab}^2$, where $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$. As such, on the weakly correlated pair nodes, i.e., on the set $\Gamma^c(\gamma, c_0)$, the condition of Lemma 6.1 of Liu (2013) is fulfilled, which treats the boundedness of $|\mathbb{V}\mathrm{ar}(\mathbf{Z}_{ab,l}/\sigma_{ab}^2) - 1|$ in this case, and since

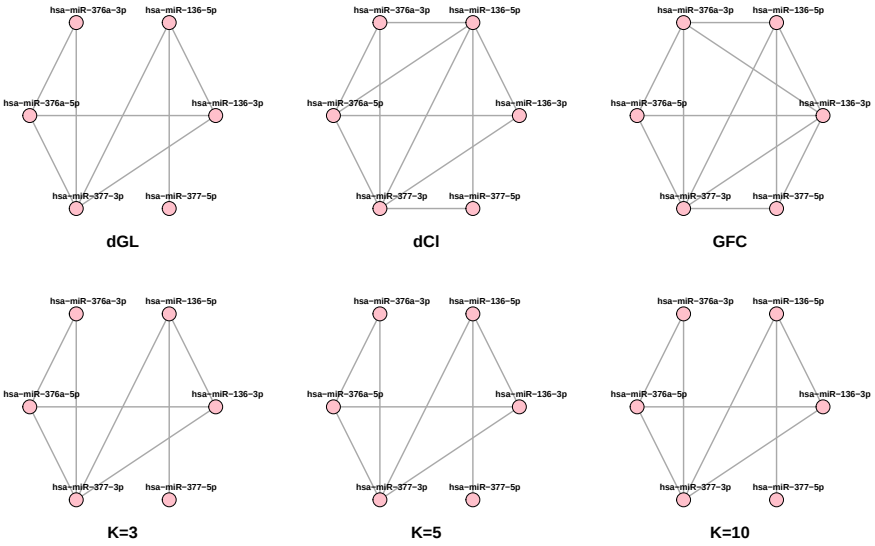


Figure 3.6: The estimated sub-graph between the genes for the GBM cancer, at the pre-specified level $\alpha = 0.05$.

$t_p < 2\sqrt{\log(p)}$, for some constant $C > 0$ we have

$$\max_{a,b} \sup_{0\le t\le t_p} \left| \frac{\mathbb{P}(|T_{ab}| > t)}{G(t)} - 1 \right| \le \max_{a,b} \sup_{0\le t\le 2\sqrt{\log(p)}} \left| \frac{\mathbb{P}(|\sum_{l=1}^{n} \mathbf{Z}_{ab,l}/\sigma_{ab}| > t\sqrt{n})}{G(t)} - 1 \right| \le C(\log(p))^{-1-\gamma_1},$$

where $\gamma_1 = \min\{1/2, \gamma\}$ and $\gamma > 0$ is defined in (3.4.6).

To show (3.4.8), we have

$$\max_{a,b} |\tilde{T}_{ab} - T_{ab}| = \max_{a,b} \left| \frac{\sigma_{ab}}{\hat{\sigma}_{ab}} T_{ab} + \frac{\sqrt{n}\mathbf{R}_{ab}}{\hat{\sigma}_{ab}} - T_{ab} \right| \le \max_{a,b} \left\{ |T_{ab}| \times \sigma_{ab} | \frac{1}{\hat{\sigma}_{ab}} - \frac{1}{\sigma_{ab}} | \right\} + \sqrt{n} \max_{a,b} \frac{|\mathbf{R}_{ab}|}{\hat{\sigma}_{ab}}. \quad (3.8.1)$$

Using Lemma 2 of Jankova and van de Geer (2015), we have

$$|\hat{\sigma}_{ab}^2 - \sigma_{ab}^2| = O_p(\frac{1}{\alpha_1} \kappa_{\mathbf{H}} \sqrt{\log(p)/n}), \quad (3.8.2)$$

where $\kappa_{\mathbf{H}} = |||(\mathbf{H}_{\mathbb{SS}})^{-1}|||_\infty$ and α_1 are defined in Chapter 2. As such, it can be

shown that the inverse ratio also satisfies

$$|(\sigma_{ab}/\hat{\sigma}_{ab}) - 1| = O_p(\sqrt{\log(p)/n}), \tag{3.8.3}$$

by letting $1/\alpha_1 = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$. Under assumption (A2), using Lemma 1 of Jankova and van de Geer (2015), it holds that

$$|\mathbf{R}_{ab}| = O_p\bigg(\frac{1}{\alpha_1^2}\kappa_{\mathbf{H}} \max\{d^{3/2}\log(p)/n, \frac{1}{\alpha_1}\kappa_{\mathbf{H}} d^2(\log(p)/n)^{3/2}\}\bigg), \tag{3.8.4}$$

which is $o_p(1/\sqrt{n})$, under the sparsity condition $d^{3/2} = o(\sqrt{n}/\log(p))$, and by letting $1/\alpha_1 = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$.

Moreover, under assumption (A2), by Lemma 8 of Jankova and van de Geer (2015), it also holds that

$$|T_{ab}| = O_p(d\sqrt{\log(p)}). \tag{3.8.5}$$

Combining (3.8.3), (3.8.4) with (3.8.5), and then substituting in (3.8.1), under the sparsity condition $d^{3/2} = o(\sqrt{n}/(\log(p))^{3/2})$, the $o_p(1/\sqrt{\log(p)})$ result follows directly. Note that due to the small cardinality of the highly correlated set, i.e., $\Upsilon = o(p^{\delta\rho})$, $\delta \in (0, 1/2]$, $\rho \in (0, 1)$, relative to the cardinality of the set $\Gamma^c(\gamma, c_0)$, the impact of the highly correlated statistics T_{ab} on the final rates can be neglected following the arguments of Liu (2013) and Li et al. (2022).

3.8.2 Proof of Theorem 3.4.2

Recall that $p_1 = p(p-1)/2$, $G(t) = 2(1 - \Phi(t))$ and let $Q(t) = G(t)/2$. First suppose that t_0 in (3.4.5) exists. By (3.4.4), we have

$$\begin{aligned}
\mathrm{FDP}_{\mathrm{dir}}(t_0) &\leq \frac{\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\tilde{T}_{ab} \geq t_0) + \sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\tilde{T}_{ab} \leq -t_0)}{\max\{R(t_0), 1\}} \\
&= \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(\tilde{T}_{ab} \geq t_0) - Q(t_0)\} + \sum_{(a,b)\in S_{\leq 0}} Q(t_0) + \sum_{(a,b)\in S_{\geq 0}} Q(t_0)}{\max\{R(t_0), 1\}} \\
&\quad + \frac{\sum_{(a,b)\in S_{\geq 0}} \{\mathbb{I}(\tilde{T}_{ab} \leq -t_0) - Q(t_0)\}}{\max\{R(t_0), 1\}} \\
&= \frac{p_1 G(t_0)}{\max\{R(t_0), 1\}} \times \bigg\{ \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(\tilde{T}_{ab} \geq t_0) - Q(t_0)\}}{p_1 G(t_0)} \\
&\quad + \frac{\sum_{(a,b)\in S_{\geq 0}} \{\mathbb{I}(\tilde{T}_{ab} \leq -t_0) - Q(t_0)\}}{p_1 G(t_0)} \bigg\} + \frac{Q(t_0)(2p_1 - s)}{\max\{R(t_0), 1\}} \\
&\leq \frac{p_1 G(t_0) A_p + p_1 G(t_0)}{\max\{R(t_0), 1\}} \leq \alpha(A_p + 1),
\end{aligned}$$

where the first equality is obtained by adding and subtracting $\sum_{(a,b)\in S_{\leq 0}} Q(t_0)$ and $\sum_{(a,b)\in S_{\geq 0}} Q(t_0)$, while the second equality is obtained by multiplying and dividing $p_1 G(t_0)$ to the first and third terms and the fact that $\sum_{(a,b)\in S_{\leq 0}} Q(t_0) + \sum_{(a,b)\in S_{\geq 0}} Q(t_0) = Q(t_0)(2p_1 - s)$. The last inequality is deduced using (3.4.5) and with $\alpha = \frac{p_1 G(t_0)}{\max\{R(t_0),1\}}$, where the term A_p is defined as

$$
\begin{aligned}
A_p &:= \sup_{0\leq t\leq t_p} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(\tilde{T}_{ab} \geq t) - Q(t)\} + \sum_{(a,b)\in S_{\geq 0}} \{\mathbb{I}(\tilde{T}_{ab} \leq -t) - Q(t)\}}{p_1 G(t)} \right| \\
&\leq \sup_{0\leq t\leq t_p} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(\tilde{T}_{ab} \geq t) - Q(t)\}}{p_1 G(t)} \right| \\
&\quad + \sup_{0\leq t\leq t_p} \left| \frac{\sum_{(a,b)\in S_{\geq 0}} \{\mathbb{I}(\tilde{T}_{ab} \leq -t) - Q(t)\}}{p_1 G(t)} \right|.
\end{aligned}
$$

As such, to show (3.4.9), it is enough to show that

$$
\sup_{0\leq t\leq t_p} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(\tilde{T}_{ab} \geq t) - Q(t)\}}{p_1 G(t)} \right| \xrightarrow{p} 0, \tag{3.8.6}
$$

and by similar techniques, one can show that the second term in the upper bound for A_p also tends to zero in probability. Using Lemma 3.4.1, the maximal difference over pairs (a,b) between $\tilde{T}_{ab}$ and T_{ab} is of order $o_p(1/\sqrt{\log(p)})$. As such, to show (3.8.6), we only need to show that

$$
\sup_{0\leq t\leq t_p} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 G(t)} \right| \xrightarrow{p} 0,
$$

and to this end, we need to show that for every $\upsilon > 0$,

$$
\sup_{0\leq t\leq t_p} \mathbb{P}\left\{ \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq \upsilon \right\} = o(1) \tag{3.8.7}
$$

and

$$
\int_0^{t_p} \mathbb{P}\left\{ \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq \upsilon \right\} dt = o(\nu_p), \tag{3.8.8}
$$

where $\nu_p = 1/\sqrt{\log(p)}$. The reader can refer to Lemma 7.3 of Javanmard and Javadi (2019) and Lemma 6.3 of Liu (2013) for more details on the motivation behind (3.8.7) and (3.8.8). We show that these two conditions hold in this situation and for the reader's convenience we first briefly explain the intuition behind.

By discretization of the interval $[0, t_p]$ to $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_b = t_p$, such that $\tau_j - \tau_{j-1} = \nu_p$ for $1 \leq j \leq b-1$ and $\tau_b - \tau_{b-1} \leq \nu_p$, where $\nu_p = 1/\sqrt{\log(p)}$, we have $b \approx t_p/\nu_p$. For any $t \in [\tau_{j-1}, \tau_j]$, it holds that

$$\begin{aligned}\frac{\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(T_{ab} \geq \tau_j)}{p_1 Q(\tau_j)} \times \frac{Q(\tau_j)}{Q(\tau_{j-1})} &\leq \sum_{(a,b)\in S_{\leq 0}} \frac{\mathbb{I}(T_{ab} \geq t)}{p_1 Q(t)} \\ &\leq \frac{\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(T_{ab} \geq \tau_{j-1})}{p_1 Q(\tau_{j-1})} \times \frac{Q(\tau_{j-1})}{Q(\tau_j)}.\end{aligned}$$

As such, it suffices to show that

$$\max_{0\leq j\leq b} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq \tau_j) - Q(\tau_j)\}}{p_1 Q(\tau_j)} \right| \xrightarrow{p} 0. \tag{3.8.9}$$

We have that

$$\begin{aligned}&\mathbb{P}\left(\max_{0\leq j\leq b} \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq \tau_j) - Q(\tau_j)\}}{p_1 Q(\tau_j)} \right| \geq v \right) \\ &\quad \leq \sum_{j=0}^{b} \mathbb{P}\left(\left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq \tau_j) - Q(\tau_j)\}}{p_1 Q(\tau_j)} \right| \geq v \right) \\ &\quad \leq \frac{1}{\nu_p} \int_0^{t_p} \mathbb{P}\left(\left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq v \right) dt \\ &\qquad + \sum_{j=b-1}^{b} \mathbb{P}\left(\left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq \tau_j) - Q(\tau_j)\}}{p_1 Q(\tau_j)} \right| \geq v \right).\end{aligned}$$

As such, to show (3.8.9), we need to show that

$$\frac{1}{\nu_p} \int_0^{t_p} \mathbb{P}\left(\left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq v \right) dt = o(1),$$

which is equivalent to (3.8.8), and

$$\sum_{j=b-1}^{b} \mathbb{P}\left(\left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq \tau_j) - Q(\tau_j)\}}{p_1 Q(\tau_j)} \right| \geq v \right) = o(1).$$

By showing (3.8.7), one can easily conclude that the later result holds.

Before proving (3.8.7), to simplify notation, define $\Psi_{\leq 0}(\rho) = \Psi(\rho) \bigcap S_{\leq 0}$. By dividing the set $S_{\leq 0}$ into two disjoint sets $\Psi_{\leq 0}(\rho)$ and $S_{\leq 0} \setminus \Psi_{\leq 0}(\rho)$, in (3.8.7) we

have

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\frac{\sum_{(a,b)\in S_{\leq 0}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}}{p_1Q(t)}\Big|\geq \upsilon\Big\}\\
&\leq \mathbb{P}\Big\{\Big|\sum_{(a,b)\in\Psi_{\leq 0}(\rho)}\{\frac{\mathbb{I}(T_{ab}\geq t)-Q(t)}{p_1Q(t)}\}\Big|\geq \upsilon/2\Big\}\\
&\quad+\mathbb{P}\Big\{\Big|\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\frac{\mathbb{I}(T_{ab}\geq t)-Q(t)}{p_1Q(t)}\}\Big|\geq \upsilon/2\Big\}, \qquad (3.8.10)
\end{aligned}$$

where we have used that $\mathbb{P}(|X+Y|\geq \upsilon)\leq \mathbb{P}(|X|\geq \upsilon/2)+\mathbb{P}(|Y|\geq \upsilon/2)$. Using Markov's inequality for the first probability on the right hand side of (3.8.10), we get

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\sum_{(a,b)\in\Psi_{\leq 0}(\rho)}\{\frac{\mathbb{I}(T_{ab}\geq t)-Q(t)}{p_1Q(t)}\}\Big|\geq \upsilon/2\Big\}\\
&\leq \frac{2}{\upsilon p_1Q(t)}\mathbb{E}\Big\{\Big|\sum_{(a,b)\in\Psi_{\leq 0}(\rho)}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}\Big|\Big\}\\
&\leq \frac{2}{\upsilon p_1Q(t)}\mathbb{E}\Big\{\sum_{(a,b)\in\Psi_{\leq 0}(\rho)}|\mathbb{I}(T_{ab}\geq t)|+Q(t)\Big\}\\
&\leq \frac{2}{\upsilon p_1Q(t)}\sum_{(a,b)\in\Psi_{\leq 0}(\rho)}\{\mathbb{E}(\mathbb{I}(T_{ab}\geq t))+Q(t)\}\\
&\leq \frac{4\psi}{\upsilon p_1}=O(\frac{p^{\delta\rho}}{p_1\sqrt{\log(p)}}), \qquad (3.8.11)
\end{aligned}$$

where the last inequality holds using the uniform convergence of the tail probability of T_{ab} to the tail probability of the normal distribution provided in Lemma 3.4.1, and the fact that $\Psi_{\leq 0}(\rho)\subseteq\Psi(\rho)$ with cardinality of $\Psi(\rho)$ satisfying $\psi=O(p^{\delta\rho}/\sqrt{\log(p)})$. Note that the probability in (3.8.11) becomes $o(1)$ under $0<\delta\leq 1/2$ and $0<\rho<1$, reason for which these two conditions are presented in the statement of Theorem 3.4.2. Using Chebyshev's inequality for the second probability on the right hand side of (3.8.10), we have

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\frac{\mathbb{I}(T_{ab}\geq t)-Q(t)}{p_1Q(t)}\}\Big|\geq \upsilon/2\Big\}\\
&\leq \frac{4}{p_1^2Q^2(t)\upsilon^2}\mathbb{E}\Big\{\Big(\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}\Big)^2\Big\}. \qquad (3.8.12)
\end{aligned}$$

Substituting (3.9.1) of Lemma 3.9.1 from Appendix B into (3.8.12) and then together with (3.8.11), the $o(1)$ result of (3.8.7) follows.

To show (3.8.8), by substituting (3.9.7) from the proof of Lemma 3.9.1 into (3.8.12), then together with (3.8.11), for some constants $C > 0$ and $C' > 0$ we have

$$\int_0^{t_p} \mathbb{P}\left\{ \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq \upsilon \right\} dt$$
$$\leq \int_0^{t_p} \left\{ \frac{4\psi}{\upsilon p_1} + \frac{4C' o(p^{2\delta\rho})}{v^2 p_1^2} \exp\left((1-\rho)(3+\rho)t_p^2/4\right) + \frac{4C}{v^2 (\log(p))^{1+\gamma_1}} \right\} dt.$$

By taking the integral we get

$$\int_0^{t_p} \mathbb{P}\left\{ \left| \frac{\sum_{(a,b)\in S_{\leq 0}} \{\mathbb{I}(T_{ab} \geq t) - Q(t)\}}{p_1 Q(t)} \right| \geq \upsilon \right\} dt$$
$$\leq \frac{4\psi t_p}{\upsilon p_1} + \frac{4C' o(p^{2\delta\rho}) t_p}{v^2 p_1^2} \exp\left((1-\rho)(3+\rho)t_p^2/4\right) + \frac{4C t_p}{v^2 (\log(p))^{1+\gamma_1}}$$
$$= O(\frac{p^{\delta\rho}}{p_1}) + \frac{4C' o(p^{2\delta\rho}) O(\sqrt{\log(p)})}{v^2 p_1^2} \times (\frac{p_1}{\log(p_1)})^{(1-\rho)(3+\rho)/2} + \frac{4CO(\sqrt{\log(p)})}{v^2 \log(p)^{1+\gamma_1}}$$
$$= o(1/\sqrt{\log(p)}),$$

where the first equality is deduced by substituting $\psi = O(p^{\delta\rho}/\sqrt{\log(p)})$ and $t_p = \sqrt{2\log(p_1) - 2\log(\log(p_1))}$, which is of order $O(\sqrt{\log(p)})$, and the last equality follows since the last term on the third line is the slowest term and it is of order $o(1/\sqrt{\log(p)})$.

Now, suppose that t_0 in (3.4.5) does not exist, and consider $t_0 = \sqrt{2\log(p_1)}$, our alternative value for the threshold that was proposed in Section 3.4.1. We have

$$\mathbb{P}\Big(\sum_{\substack{(a,b)\in\mathcal{V}\times\mathcal{V} \\ a<b}} \mathbb{I}(\widehat{\text{sign}}_{ab} \neq \text{sign}(\boldsymbol{\Theta}_{ab})) \geq 1\Big) \leq \mathbb{P}\Big(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1\Big)$$
$$+ \mathbb{P}\Big(\sum_{(a,b)\in S_{\geq 0}} \mathbb{I}(\hat{T}_{ab} \leq -t_0) \geq 1\Big). \quad (3.8.13)$$

Since on $S_{\leq 0}$, $\hat{T}_{ab} \leq \tilde{T}_{ab}$ (as was argued in Section 3.4.1), for the first probability on the right hand side of (3.8.13) we can write

$$\mathbb{P}\Big(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1\Big) \leq \mathbb{P}\Big(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\tilde{T}_{ab} \geq t_0) \geq 1\Big) = \mathbb{P}\Big(\bigcup_{(a,b)\in S_{\leq 0}} \{\tilde{T}_{ab} \geq t_0\}\Big).$$

Using the law of total probability, by conditioning on the event $\{\max_{a,b}|\tilde{T}_{ab} - T_{ab}| \leq$

$\frac{1}{\sqrt{\log(p)}}\}$ and its complement set, we get

$$\begin{aligned}\mathbb{P}(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1) &\leq \mathbb{P}\Big(\bigcup_{(a,b)\in S_{\leq 0}} \{\tilde{T}_{ab} \geq t_0\} \Big| \max_{a,b}|\tilde{T}_{ab} - T_{ab}| \leq \frac{1}{\sqrt{\log(p)}}\Big) \\ &\quad + \mathbb{P}(\max_{a,b}|\tilde{T}_{ab} - T_{ab}| > \frac{1}{\sqrt{\log(p)}}) \\ &\leq p_1 \max_{(a,b)\in S_{\leq 0}} \mathbb{P}\Big(\tilde{T}_{ab} \geq t_0 \Big| \max_{a,b}|\tilde{T}_{ab} - T_{ab}| \leq \frac{1}{\sqrt{\log(p)}}\Big) \\ &\quad + o(1) \\ &\leq p_1 \max_{(a,b)\in S_{\leq 0}} \mathbb{P}(T_{ab} \geq t_0 - \frac{1}{\sqrt{\log(p)}}) + o(1),\end{aligned} \tag{3.8.14}$$

where the $o(1)$ result is deduced using the first result of Lemma 3.4.1. Using the Cramér result from Lemma 3.4.1, we can approximate $\mathbb{P}(T_{ab} \geq t_0 - \frac{1}{\sqrt{\log(p)}})$ with $Q(t_0 - \frac{1}{\sqrt{\log(p)}})$. Then using the upper bound in Lemma 7.1 from Javanmard and Javadi (2019), which implies that $Q(t) < \phi(t)/t$, for all $t \geq 0$, with $\phi(t) = \exp(-t^2/2)/\sqrt{2\pi}$ as the density of the standard normal distribution at point t and π as the mathematical constant, and substituting $t_0 = \sqrt{2\log(p_1)}$ into (3.8.14), we get

$$\begin{aligned}&\mathbb{P}(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1) \\ &\leq \frac{p_1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - \frac{1}{\sqrt{\log(p)}})} \exp\big(-(\sqrt{2\log(p_1)} - \frac{1}{\sqrt{\log(p)}})^2/2\big) + o(1) \\ &= \frac{1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - \frac{1}{\sqrt{\log(p)}})} \exp\big(\sqrt{\frac{2\log(p_1)}{\log(p)}} - \frac{1}{2\log(p)}\big) + o(1) \\ &= \frac{p_1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - \frac{1}{\sqrt{\log(p)}})} \times \frac{1}{p_1} \\ &\quad \times \exp\Big(-\frac{1}{2\log(p)} + \sqrt{2 + \frac{2\log(p-1)}{\log(p)} - \frac{2\log(2)}{\log(p)}}\Big) + o(1),\end{aligned}$$

which tends to zero with increasing p. By similar techniques, we can show that the second probability on the right hand side of (3.8.13) also tends to zero with increasing p, and the proof follows directly.

3.8.3 Proof of Theorem 3.4.3

Since the denominator of the threshold defined in (3.4.5) is data-dependent and different from the denominator of the directional power in (3.2.2), which does not depend on data, we first upper bound the threshold t_0 by a data-independent threshold $t_\star$ defined as

$$t_\star := \Phi^{-1}(1 - \frac{\alpha s}{2p_1}(1 - o(1))).$$

By the same technique as in the proof of Lemma E.12 from Cui et al. (2021), it can be shown that $t_0 \leq t_\star$ with high probability.

For every $\epsilon_1 > 0$ and $\epsilon_2 > 0$, define next the event

$$D(\epsilon_1, \epsilon_2) := \{ \max_{a,b} |\tilde{T}_{ab} - T_{ab}| \leq \epsilon_1, \quad \max_{a,b} |\hat{\sigma}_{ab} - \sigma_{ab}| \leq \epsilon_2 \}.$$

Let next $S_{>0} := \{1 \leq a < b \leq p : \mathbf{\Theta}_{ab} > 0\}$ and $S_{<0} := \{1 \leq a < b \leq p : \mathbf{\Theta}_{ab} < 0\}$. Then $S = S_{>0} \bigcup S_{<0}$ and we have

$$\begin{aligned}
\text{Power}_{\text{dir}} &= \frac{1}{s} \sum_{(a,b) \in S_{>0}} \mathbb{P}(\hat{T}_{ab} \geq t_0) + \frac{1}{s} \sum_{(a,b) \in S_{<0}} \mathbb{P}(\hat{T}_{ab} \leq -t_0) \\
&\geq \frac{1}{s} \sum_{(a,b) \in S_{>0}} \mathbb{P}(\hat{T}_{ab} \geq t_\star) + \frac{1}{s} \sum_{(a,b) \in S_{<0}} \mathbb{P}(\hat{T}_{ab} \leq -t_\star) \\
&\geq \frac{1}{s} \sum_{(a,b) \in S_{>0}} \mathbb{P}(\{\hat{T}_{ab} \geq t_\star\} \bigcap D) + \frac{1}{s} \sum_{(a,b) \in S_{<0}} \mathbb{P}(\{\hat{T}_{ab} \leq -t_\star\} \bigcap D) \\
&\quad - \mathbb{P}(D^c) \\
&= \frac{1}{s} \sum_{(a,b) \in S_{>0}} \mathbb{P}\bigg(\{\tilde{T}_{ab} \geq t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}}\} \bigcap D\bigg) - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s} \sum_{(a,b) \in S_{<0}} \mathbb{P}\bigg(\{\tilde{T}_{ab} \leq -t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}}\} \bigcap D\bigg),
\end{aligned}$$

where the last equality is deduced due to the fact that $\hat{T}_{ab} = \tilde{T}_{ab} + \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}}$. On event D, we have $|\hat{\sigma}_{ab} - \sigma_{ab}| \leq \epsilon_2$ for every $1 \leq a < b \leq p$, which gives $\sigma_{ab} - \epsilon_2 \leq \hat{\sigma}_{ab} \leq \sigma_{ab} + \epsilon_2$. Therefore, on $S_{>0}$, we get $\frac{-\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}} \leq \frac{-\sqrt{n}\mathbf{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}$, and as a result

$$\bigg\{\tilde{T}_{ab} \geq t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\sigma_{ab} + \epsilon_2}\bigg\} \subseteq \bigg\{\tilde{T}_{ab} \geq t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}}\bigg\}.$$

By a similar argument, one can conclude that on $S_{<0}$,

$$\bigg\{\tilde{T}_{ab} \leq -t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\sigma_{ab} + \epsilon_2}\bigg\} \subseteq \bigg\{\tilde{T}_{ab} \leq -t_\star - \frac{\sqrt{n}\mathbf{\Theta}_{ab}}{\hat{\sigma}_{ab}}\bigg\}.$$

As such,

$$\begin{aligned}
\text{Power}_{\text{dir}} &\geq \frac{1}{s}\sum_{(a,b)\in S_{>0}} \mathbb{P}\bigg(\{\tilde{T}_{ab} \geq t_\star - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\}\bigcap D\bigg) - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{<0}} \mathbb{P}\bigg(\{\tilde{T}_{ab} \leq -t_\star - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\}\bigcap D\bigg) \\
&\geq \frac{1}{s}\sum_{(a,b)\in S_{>0}} \mathbb{P}\bigg(T_{ab} \geq t_\star + \epsilon_1 - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\bigg) - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{<0}} \mathbb{P}\bigg(T_{ab} \leq -t_\star - \epsilon_1 - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\bigg) \\
&= \frac{1}{s}\sum_{(a,b)\in S_{>0}} \mathbb{P}\bigg(T_{ab} \geq t_\star + \epsilon_1 - \frac{\sqrt{n}|\boldsymbol{\Theta}_{ab}|}{\sigma_{ab}+\epsilon_2}\bigg) - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{<0}} \mathbb{P}\bigg(T_{ab} \leq -(t_\star + \epsilon_1 - \frac{\sqrt{n}|\boldsymbol{\Theta}_{ab}|}{\sigma_{ab}+\epsilon_2})\bigg), \qquad (3.8.15)
\end{aligned}$$

where the second inequality follows because on event D, we have $|\tilde{T}_{ab} - T_{ab}| \leq \epsilon_1$ for every $1 \leq a < b \leq p$, which gives $T_{ab} - \epsilon_1 \leq \tilde{T}_{ab} \leq T_{ab} + \epsilon_1$. From this inequality, one can conclude that

$$\begin{aligned}
\bigg\{T_{ab} - \epsilon_1 \geq t_\star - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\bigg\} &\subseteq \bigg\{\tilde{T}_{ab} \geq t_\star - \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2}\bigg\}, \\
\bigg\{T_{ab} + \epsilon_1 \leq -(t_\star + \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2})\bigg\} &\subseteq \bigg\{\tilde{T}_{ab} \leq -(t_\star + \frac{\sqrt{n}\boldsymbol{\Theta}_{ab}}{\sigma_{ab}+\epsilon_2})\bigg\}.
\end{aligned}$$

Using Lemma 7.1 of Javanmard and Javadi (2019), we have

$$G(\sqrt{2\log(p_1/s)}) \leq \frac{1}{\sqrt{\pi\log(p_1/s)}}\exp\{-\log(p_1/s)\} = \frac{s}{p_1\sqrt{\pi\log(p_1/s)}}. \qquad (3.8.16)$$

We show next $t_\star < \sqrt{2\log(p_1/s)}$. Assuming otherwise, i.e., $t_\star \geq \sqrt{2\log(p_1/s)}$, we have $G(t_\star) \leq G(\sqrt{2\log(p_1/s)})$, since $G(\cdot)$ is a decreasing function. Combined with (3.8.16), we get $\frac{\alpha s}{p_1}(1-o(1)) = G(t_\star) \leq \frac{s}{p_1\sqrt{\pi\log(p_1/s)}}$. As such,

$$\alpha(1-o(1)) \leq \frac{1}{\sqrt{\pi\log(p_1/s)}}. \qquad (3.8.17)$$

Since $s \leq pd$ and $d = o(\frac{n^{1/3}}{\log(p)})$, then $s/p_1 = o(\frac{n^{1/3}}{(p-1)\log(p)})$ which is $o(1)$ if $(p-1)\log(p) > n^{1/3}$. Therefore, under this condition, s/p_1 tends to zero with increasing n, and as a result, $1/\sqrt{\pi\log(p_1/s)}$ tends to zero with increasing n, which is in

contradiction with (3.8.17) in the limit, since α is a constant larger than zero. Therefore, $t_\star < \sqrt{2\log(p_1/s)}$. By assuming $|\mathbf{\Theta}_{ab}| > \sqrt{\frac{2\sigma_{ab}^2 \log(p_1/s)}{n}}$, we have

$$\frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}} > \sqrt{2\log(p_1/s)} > t_\star > \Phi^{-1}(1 - \frac{\alpha s}{2p_1}).$$

As such, $\Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}}) < 1/2$ and the denominator of (3.4.10) is lower bounded by a constant. Thus, under the conditions of Lemma 3.4.1 and Lemma 2 of Jankova and van de Geer (2015), we have

$$\frac{\mathbb{P}(D^c)}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)} \to 0,$$

as $n \to \infty$. As (3.8.15) holds for all positive small values ϵ_1 and ϵ_2, we can write

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{\text{Power}_{\text{dir}}}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)} \\
&\geq \liminf_{n\to\infty} \frac{1}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)}\Bigg\{ - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{>0}} \mathbb{P}\left(T_{ab} \geq t_\star - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}}\right) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{<0}} \mathbb{P}\left(T_{ab} \leq -(t_\star - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)\Bigg\} \\
&= \liminf_{n\to\infty} \frac{1}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)}\Bigg\{ - \mathbb{P}(D^c) \\
&\quad + \frac{1}{s}\sum_{(a,b)\in S_{>0}} \left(1 - \Phi(t_\star - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right) + \frac{1}{s}\sum_{(a,b)\in S_{<0}} \Phi\left(- (t_\star - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)\Bigg\}
\end{aligned}$$

$$\begin{aligned}
&= \liminf_{n\to\infty}\Bigg\{ \frac{1}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1-\frac{\alpha s}{2p_1}) - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)} \\
&\quad \times \frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(t_\star - \frac{\sqrt{n}|\mathbf{\Theta}_{ab}|}{\sigma_{ab}})\right)\Bigg\} = 1,
\end{aligned}$$

where the first equality is deduced using the Cramér result from Lemma 3.4.1, and the last equality is deduced by substituting $t_\star = \Phi^{-1}(1 - \frac{\alpha s}{2p_1}(1 - o(1)))$.

Note that $1 - \Phi(\Phi^{-1}(1 - \frac{\alpha s}{2p_1}) - u)$ is an increasing function of u. As such, by considering $\boldsymbol{\Theta}_{\min} = \min_{(a,b)\in S}|\boldsymbol{\Theta}_{ab}|$, we get

$$\liminf_{n\to\infty} \frac{\text{Power}_{\text{dir}}}{\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1 - \frac{\alpha s}{2p_1}) - \frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}})\right)} \geq 1.$$

Thus, to show $\text{Power}_{\text{dir}} \to 1$, it is enough to show that

$$\frac{1}{s}\sum_{(a,b)\in S}\left(1 - \Phi(\Phi^{-1}(1 - \frac{\alpha s}{2p_1}) - \frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}})\right) \to 1,$$

or, equivalently, it is enough to show that $\frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}} - \Phi^{-1}(1 - \frac{\alpha s}{2p_1}) \to \infty$. Using Lemma 7.1 of Javanmard and Javadi (2019), we have that $G(\sqrt{2\log(\frac{2p_1}{\alpha s})}) \leq \frac{\alpha s}{2p_1\sqrt{\pi\log(\frac{2p_1}{\alpha s})}} < \frac{\alpha s}{p_1}$, from which we deduce that $\sqrt{2\log(\frac{2p_1}{\alpha s})} \geq \Phi^{-1}(1 - \frac{\alpha s}{2p_1})$. As such,

$$\frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}} - \Phi^{-1}(1 - \frac{\alpha s}{2p_1}) \geq \frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}} - \sqrt{2\log(\frac{2p_1}{\alpha s})}.$$

Assuming $\sqrt{n}\boldsymbol{\Theta}_{\min} - \sqrt{2\max_{(a,b)\in S}\sigma_{ab}^2\log(\frac{2p_1}{\alpha s})} \to \infty$, it can be easily shown that $\frac{\sqrt{n}\boldsymbol{\Theta}_{\min}}{\sigma_{ab}} - \Phi^{-1}(1 - \frac{\alpha s}{2p_1}) \to \infty$, and the proof is completed.

3.8.4 Proof of Lemma 3.4.4

Define the random variable

$$U_{ab} = \sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

where we remind the reader that $\dot{\mathbf{Y}}_{l,k} = (Y_{l,k}^1, \ldots, Y_{l,k}^p)^\top$ is the l-th vector of variables, $l = 1, \ldots, n_k$, in the k-th sub-sample, $k = 1, \ldots, K$. We have that $\max_{a,b}|\tilde{U}_{ab} - T_{ab}| \leq \max_{a,b}|\tilde{U}_{ab} - U_{ab}| + \max_{a,b}|U_{ab} - T_{ab}|$. First, to show the negligibility of the term $\max_{a,b}|U_{ab} - T_{ab}|$, we have

$$\max_{a,b}|U_{ab} - T_{ab}| \leq \max_{a,b}\left\{\sum_{k=1}^{K}\frac{\sigma_{ab}}{\sqrt{n}}|\frac{1}{\hat{\sigma}_{ab,k}^2} - \frac{1}{\sigma_{ab}^2}| \times \left|\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right|\right\}. \tag{3.8.18}$$

Rewriting (2.8.1) from Chapter 2 for the reader's convenience, we have

$$|\frac{1}{\hat{\sigma}^2_{ab,k}} - \frac{1}{\sigma^2_{ab}}| = O_p(\sqrt{\log(p)/n_k}), \quad k = 1, \ldots, K, \tag{3.8.19}$$

under the conditions $1/\alpha_1 = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$. Moreover, under assumption (A2) from Chapter 2 and using Lemma 8 of Jankova and van de Geer (2015), it holds that

$$|\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})| = O_p(d\sqrt{n_k \log(p)}). \tag{3.8.20}$$

Substituting (3.8.19) and (3.8.20) into (3.8.18), we get

$$\max_{a,b}|U_{ab} - T_{ab}| = O_p(Kd\log(p)/\sqrt{n}). \tag{3.8.21}$$

To bound the term $\max_{a,b}|\tilde{U}_{ab} - U_{ab}|$, we have

$$\begin{aligned}\max_{a,b}|\tilde{U}_{ab} - U_{ab}| \leq & \max_{a,b}|\mathbf{R}_{ab,\dagger}| \\ & + \max_{a,b}\left\{\left|\sqrt{\frac{\sum_{k=1}^K n_k/\sigma^2_{ab}}{\sum_{k=1}^K n_k/\hat{\sigma}^2_{ab,k}}} - 1\right| \times \left|\sum_{k=1}^K\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right|\right\}.\end{aligned} \tag{3.8.22}$$

It has been shown in Nezakati and Pircalabelu (2023c) that

$$|\mathbf{R}_{ab,\dagger}| = O_p(\frac{K}{\sqrt{n}}\max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\}). \tag{3.8.23}$$

To bound $\max_{a,b}\left|\sqrt{\frac{\sum_{k=1}^K n_k/\sigma^2_{ab}}{\sum_{k=1}^K n_k/\hat{\sigma}^2_{ab,k}}} - 1\right|$, using (2.8.2) from Chapter 2, we have

$$\max_{a,b}\left|\sqrt{\frac{\sum_{k=1}^K n_k/\sigma^2_{ab}}{\sum_{k=1}^K n_k/\hat{\sigma}^2_{ab,k}}} - 1\right| = O_p(K\sqrt{\log(p)/n}). \tag{3.8.24}$$

Moreover, to bound $\max_{a,b}\left|\sum_{k=1}^K\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right|$, using the triangle inequality, we have

$$\begin{aligned}&\max_{a,b}\left|\sum_{k=1}^K\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right| \\ &\leq \max_{a,b}\left|\sum_{k=1}^K\sum_{l=1}^{n_k}\frac{1}{\sqrt{n}\sigma_{ab}}(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right| \\ &\quad + \max_{a,b}\left|\sum_{k=1}^K\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}}(\frac{1}{\hat{\sigma}^2_{ab,k}} - \frac{1}{\sigma^2_{ab}})(\boldsymbol{\Theta}_a^\top \dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right|.\end{aligned}$$

Using the bounds from (3.8.19) and (3.8.20), we get

$$\begin{aligned}\max_{a,b}\left|\sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}}(\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_{l,k}\dot{\mathbf{Y}}_{l,k}^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab})\right| &= O_p(Kd\sqrt{\log(p)})\\ &\quad + O_p(Kd\log(p)/\sqrt{n})\\ &= O_p(Kd\sqrt{\log(p)}). \qquad (3.8.25)\end{aligned}$$

Substituting (3.8.23), (3.8.24) and (3.8.25) into (3.8.22), we have

$$\begin{aligned}\max_{a,b}|\tilde{U}_{ab}-U_{ab}| &= O_p(\frac{K}{\sqrt{n}}\max\{d^{3/2}\log(p),d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\})\\ &\quad + O_p(K^2d\log(p)/\sqrt{n}).\end{aligned}$$

Combining this with (3.8.21), we get

$$\begin{aligned}\max_{a,b}|\tilde{U}_{ab}-T_{ab}| &= O_p(\frac{K}{\sqrt{n}}\max\{d^{3/2}\log(p),d^2(\log(p))^{3/2}/\sqrt{n_\dagger}\})\\ &\quad + O_p(K^2d\log(p)/\sqrt{n}) + O_p(Kd\log(p)/\sqrt{n})\\ &= O_p\bigg(\frac{K}{\sqrt{n}}\max\{d^{3/2}\log(p),d^2(\log(p))^{3/2}/\sqrt{n_\dagger},Kd\log(p)\}\bigg).\end{aligned}$$

By considering $K=O(n^{1/4}/(d\log(p)))$ and $d^{3/2}=o(\sqrt{n_\dagger}/(\log(p))^{3/2})$, the $o_p(1)$ result follows directly.

3.8.5 Proof of Theorem 3.4.5

The proof of this theorem deploys a similar technique as that of Theorem 3.4.2, and we omit it. However, when the given t_0' does not exist and we let $t_0'=\sqrt{2\log(p_1)}$, we need to condition on the event

$$\bigg\{\max_{a,b}|\tilde{U}_{ab}-T_{ab}|\leq\frac{C_1K}{\sqrt{n}}\max\{d^{3/2}\log(p),d^2(\log(p))^{3/2}/\sqrt{n_\dagger},Kd\log(p)\bigg\},$$

and its complement set, where C_1 is a positive constant. By following the same technique as the second part of the proof of Theorem 3.4.2 (when we consider $t_0=\sqrt{2\log(p_1)}$), we get

$$\mathbb{P}(\sum_{(a,b)\in S_{\leq 0}}\mathbb{I}(\hat{T}_{ab}\geq t_0)\geq 1)\leq p_1\max_{(a,b)\in S_{\leq 0}}\mathbb{P}(T_{ab}\leq t_0-A)+o(1), \qquad (3.8.26)$$

where $A := \frac{C_1 K}{\sqrt{n}} \max\{d^{3/2}\log(p), d^2(\log(p))^{3/2}/\sqrt{n_\dagger}, Kd\log(p)\}$.
If $A = \frac{C_1 K}{\sqrt{n}} d^{3/2}\log(p)$, then in (3.8.26) we have

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1\Big) \\
&\leq p_1 \max_{(a,b)\in S_{\leq 0}} \mathbb{P}(T_{ab} \leq t_0 - C_1 K d^{3/2}\log(p)/\sqrt{n}) + o(1) \\
&\leq \frac{p_1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - C_1 K d^{3/2}\log(p)/\sqrt{n})} \\
&\quad \times \exp\big(-(\sqrt{2\log(p_1)} - C_1 K d^{3/2}\log(p)/\sqrt{n})^2/2\big) + o(1) \\
&= \frac{p_1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - C_1 K d^{3/2}\log(p)/\sqrt{n})} \times \frac{1}{p_1} \\
&\quad \times \exp\Big(-\frac{C_1^2 K^2}{2n} d^3(\log(p))^2\Big) \times \exp\Big(\frac{C_1 K}{\sqrt{n}} d^{3/2}\log(p)\sqrt{2\log(p_1)}\Big) + o(1).
\end{aligned} \tag{3.8.27}$$

As $K = O(n^{1/4}/(d\log(p)))$, there exists a positive constant C_2, such that $K \leq C_2 n^{1/4}/(d\log(p))$. Therefore, $\frac{C_1 K}{\sqrt{n}} d^{3/2}\log(p)\sqrt{2\log(p_1)} \leq \frac{C_1 C_2}{n^{1/4}}\sqrt{2d\log(p_1)}$. Since $d^{3/2} = o(\sqrt{n_\dagger}/(\log(p))^{3/2})$, it can be easily shown that

$$\frac{C_1 K}{\sqrt{n}} d^{3/2}\log(p)\sqrt{2\log(p_1)} \leq \frac{C_1 C_2}{n^{1/4}}\sqrt{2d\log(p_1)} \to 0,$$

with increasing n. On the other hand,

$$\frac{1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - C_1 K d^{3/2}\log(p)/\sqrt{n})} \leq \frac{1}{\sqrt{2\pi}(\sqrt{2\log(p_1)} - C_1 C_2\sqrt{d}/n^{1/4})} \to 0,$$

with increasing n, under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/(\log(p))^{3/2})$. Considering these results in (3.8.27), one can conclude that $\mathbb{P}(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1)$ tends to zero with increasing n. By similar arguments, if $A = C_1 K d^2(\log(p))^{3/2}/\sqrt{nn_\dagger}$ or $A = C_1 K 2d\log(p)/\sqrt{n}$, one can show that $\mathbb{P}(\sum_{(a,b)\in S_{\leq 0}} \mathbb{I}(\hat{T}_{ab} \geq t_0) \geq 1)$ tends to zero with increasing n.

3.9 Appendix B: A technical lemma and its proof

Lemma 3.9.1. Let $p \leq n^r$ for some $r > 0$, and suppose that assumptions (A1)-(A2) in Chapter 2 hold. Consider the random variable T_{ab} defined in (3.4.7). Under the sparsity condition $d^{3/2} = o(\sqrt{n}/(\log(p))^{3/2})$, and considering cardinalities $\Upsilon = o(p^{\delta\rho})$, $\rho \in (0,1)$, $\delta \in (0, 1/2]$ and $\psi = O(p^{\delta\rho}/\sqrt{\log(p)})$, for the sets $\Gamma(\gamma, c_0)$ and

$\Psi(\rho)$, respectively, it holds that

$$\mathbb{E}\left\{\left(\frac{\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}}{p_1Q(t)}\right)^2\right\}=o(1). \tag{3.9.1}$$

Proof. To show (3.9.1), we have

$$\begin{aligned}&\mathbb{E}\left\{\left(\frac{\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}}{p_1Q(t)}\right)^2\right\}\\&=\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\left\{\frac{\mathbb{P}(T_{ab}\geq t,T_{cd}\geq t)-Q^2(t)}{Q^2(t)}\right\},\end{aligned}$$

where the equality is deduced first by squaring the terms inside the expectation and then using the Cramér result on the tail probability of T_{ab} from Lemma 3.4.1. To bound the above expectation, we consider two disjoint sets $\Gamma(\gamma,c_0)$ (highly correlated pair nodes) and $\Gamma^c(\gamma,c_0)$ (weakly correlated pair nodes). To simplify the notation, define

$$\Gamma_{\leq 0}(\gamma,c_0)=\Gamma(\gamma,c_0)\bigcap S_{\leq 0},\quad \Gamma^c_{\leq 0}(\gamma,c_0)=\Gamma^c(\gamma,c_0)\bigcap S_{\leq 0}.$$

Note that the set $S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)$ is equal to the union of two disjoint sets $\Gamma_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)$ and $\Gamma^c_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)$. As such,

$$\begin{aligned}&\mathbb{E}\left\{\left(\frac{\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}}{p_1Q(t)}\right)^2\right\}\\&\leq\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{\Gamma_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)\}}\left\{\frac{\mathbb{P}(T_{ab}\geq t,T_{cd}\geq t)}{Q^2(t)}\right\}\\&\quad+\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{\Gamma^c_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)\}}\left\{\frac{\mathbb{P}(T_{ab}\geq t,T_{cd}\geq t)}{Q^2(t)}-1\right\}.\end{aligned} \tag{3.9.2}$$

Recall that

$$\mathbf{Z}_{ab,l}=\boldsymbol{\Theta}_a^\top\dot{\mathbf{Y}}_l\dot{\mathbf{Y}}_l^\top\boldsymbol{\Theta}_b-\boldsymbol{\Theta}_{ab}=\mathbf{e}_a^\top\boldsymbol{\Theta}\dot{\mathbf{Y}}_l\dot{\mathbf{Y}}_l^\top\boldsymbol{\Theta}\mathbf{e}_b-\boldsymbol{\Theta}_{ab}=\mathbf{Z}_l^a\mathbf{Z}_l^b-\boldsymbol{\Theta}_{ab},$$

where $\mathbf{Z}_l^a$ is the a-th element of $\mathbf{Z}_l:=\boldsymbol{\Theta}\dot{\mathbf{Y}}_l\sim\mathcal{N}_p(\mathbf{0},\boldsymbol{\Theta})$ and $\mathbf{e}_a$ is a unit column vector with 1 at the a-th position and zero everywhere else. Thus, $\mathbb{E}(\mathbf{Z}_l^a\mathbf{Z}_l^b)=\mathbb{C}\mathrm{ov}(\mathbf{Z}_l^a,\mathbf{Z}_l^b)=\boldsymbol{\Theta}_{ab}$ and then $\mathbb{E}(\mathbf{Z}_{ab,l})=\mathbb{E}(\mathbf{Z}_l^a\mathbf{Z}_l^b)-\boldsymbol{\Theta}_{ab}=0$. Moreover,

$$\mathbb{E}(\mathbf{Z}_l^a\mathbf{Z}_l^b\mathbf{Z}_l^c\mathbf{Z}_l^d)=\boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{cd}+\boldsymbol{\Theta}_{ac}\boldsymbol{\Theta}_{bd}+\boldsymbol{\Theta}_{ad}\boldsymbol{\Theta}_{bc},$$

and as such, we get

$$\mathbb{C}\mathrm{ov}\left(\frac{\mathbf{Z}_{ab,l}}{\sigma_{ab}},\frac{\mathbf{Z}_{cd,l}}{\sigma_{cd}}\right)=\frac{\boldsymbol{\Theta}_{ac}\boldsymbol{\Theta}_{bd}+\boldsymbol{\Theta}_{ad}\boldsymbol{\Theta}_{bc}}{\sqrt{(\boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}+\boldsymbol{\Theta}_{ab}^2)(\boldsymbol{\Theta}_{cc}\boldsymbol{\Theta}_{dd}+\boldsymbol{\Theta}_{cd}^2)}}.$$

On $\Gamma_{\leq 0}(\gamma, c_0) \setminus \Psi_{\leq 0}(\rho)$, we have $|\boldsymbol{\Theta}_{ab}| \leq \frac{1}{\sqrt{2}}(\frac{1-\rho}{1+\rho})\sqrt{\boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}}$, and so,

$$\begin{aligned}\left|\mathbb{C}\text{ov}(\frac{\mathbf{Z}_{ab,l}}{\sigma_{ab}}, \frac{\mathbf{Z}_{cd,l}}{\sigma_{cd}})\right| &\leq (\frac{1-\rho}{1+\rho})^2 \\ &\times \left(\frac{\boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}\boldsymbol{\Theta}_{cc}\boldsymbol{\Theta}_{dd}}{\boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}\boldsymbol{\Theta}_{cc}\boldsymbol{\Theta}_{dd} + \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb}\boldsymbol{\Theta}^2_{cd} + \boldsymbol{\Theta}_{cc}\boldsymbol{\Theta}_{dd}\boldsymbol{\Theta}^2_{ab} + \boldsymbol{\Theta}^2_{ab}\boldsymbol{\Theta}^2_{cd}}\right)^{1/2} \\ &\leq (\frac{1-\rho}{1+\rho})^2.\end{aligned}$$

Considering the above bound on the covariance of $\mathbf{Z}_{ab,l}/\sigma_{ab}$ and $\mathbf{Z}_{cd,l}/\sigma_{cd}$ in Lemma 6.2 of Liu (2013) for which the assumptions hold, we can write on $\Gamma_{\leq 0}(\gamma, c_0) \setminus \Psi_{\leq 0}(\rho)$,

$$\begin{aligned}\mathbb{P}(T_{ab} \geq t, T_{cd} \geq t) &= \mathbb{P}\Big(\sum_{l=1}^{n} \frac{\mathbf{Z}_{ab,l}}{\sigma_{ab}} \geq t\sqrt{n}, \sum_{l=1}^{n} \frac{\mathbf{Z}_{cd,l}}{\sigma_{cd}} \geq t\sqrt{n}\Big) \\ &\leq C'(1+t)^{-2} \exp\Big(- t^2/(1 + (\frac{1-\rho}{1+\rho})^2)\Big) \\ &\leq C'(1+t)^{-2} \exp\big(- (1+\rho)^2 t^2/4\big),\end{aligned}$$

for some constant $C' > 0$. As such, on $\Gamma_{\leq 0}(\gamma, c_0) \setminus \Psi_{\leq 0}(\rho)$,

$$\begin{aligned}&\frac{1}{p_1^2} \sum_{(a,b),(c,d) \in \{\Gamma_{\leq 0}(\gamma,c_0) \setminus \Psi_{\leq 0}(\rho)\}} \frac{\mathbb{P}(T_{ab} \geq t, T_{cd} \geq t)}{Q^2(t)} \\ &\quad \leq \frac{C'\Upsilon^2}{p_1^2 Q^2(t)}(1+t)^{-2} \exp\big(- (1+\rho)^2 t^2/4\big) \qquad (3.9.3) \\ &\quad \leq \frac{C'\Upsilon^2}{p_1^2} \exp\big((1-\rho)(3+\rho)t^2/4\big), \qquad (3.9.4)\end{aligned}$$

where we recall that $\Upsilon = \#\Gamma(\gamma, c_0)$. The last inequality follows using the lower bound of Lemma 7.1 of Javanmard and Javadi (2019), that is $Q(t) > \frac{t}{t^2+1}\phi(t)$, for all $t \geq 0$, with $\phi(t) = \exp(-t^2/2)/\sqrt{2\pi}$. Using this lower bound on $Q(t)$, we get

$$\frac{1}{Q^2(t)(t+1)^2} < \frac{2\pi(t^2+1)^2}{t^2(t+1)^2} \exp(t^2) \leq 2\pi \exp(t^2).$$

The result in (3.9.4) is deduced by substituting this upper bound in (3.9.3).

Now, to bound the second term on the right hand side of (3.9.2), we have

$$\frac{\mathbb{P}(T_{ab} \geq t, T_{cd} \geq t)}{Q^2(t)} - 1 \leq \left|\frac{\mathbb{P}(|T_{ab}| \geq t, |T_{cd}| \geq t)}{Q^2(t)} - 1\right|.$$

As such,

$$\begin{aligned}
&\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{\Gamma^c_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)\}}\left\{\frac{\mathbb{P}(T_{ab}\geq t,T_{cd}\geq t)}{Q^2(t)}-1\right\}\\
&\quad\leq\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{\Gamma^c_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)\}}\left|\frac{\mathbb{P}(|T_{ab}|\geq t,|T_{cd}|\geq t)}{Q^2(t)}-1\right|\\
&\quad\leq\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\Gamma^c_{\leq 0}(\gamma,c_0)}\left|\frac{\mathbb{P}(|T_{ab}|\geq t,|T_{cd}|\geq t)}{Q^2(t)}-1\right|. \qquad (3.9.5)
\end{aligned}$$

Consider $\mathcal{Z}_l=(\frac{\mathbf{Z}_{ab,l}}{\sigma_{ab}},\frac{\mathbf{Z}_{cd,l}}{\sigma_{cd}})^\top$, $l=1,\ldots,n$, on $\Gamma^c_{\leq 0}(\gamma,c_0)$, then

$$||\mathbb{C}\mathrm{ov}(\mathcal{Z}_l)-\mathbf{I}_2||_F=\sqrt{2}\frac{|\mathbf{\Theta}_{ac}\mathbf{\Theta}_{bd}+\mathbf{\Theta}_{ad}\mathbf{\Theta}_{bc}|}{\sigma_{ab}\sigma_{cd}}\leq c_3(\log(p))^{-2-\gamma},$$

for every $\gamma>0$, and some constant $c_3>0$. The last inequality holds because on $\Gamma^c_{\leq 0}(\gamma,c_0)$, we have $|\mathbf{\Theta}^0_{ab}|\leq c_0(\log(p))^{-1-\gamma/2}$. By Lemma 6.1 of Liu (2013), for which the assumptions hold,

$$\sup_{0\leq t\leq 2\sqrt{\log(p)}}\left|\frac{\mathbb{P}(|T_{ab}|\geq t,|T_{cd}|\geq t)}{Q^2(t)}-1\right|\leq C(\log(p))^{-1-\gamma_1},$$

for some constant $C>0$, where $\gamma_1=\min\{\gamma,1/2\}$. Substituting this bound into (3.9.5), we get

$$\begin{aligned}
\frac{1}{p_1^2}\sum_{(a,b),(c,d)\in\{\Gamma^c_{\leq 0}(\gamma,c_0)\setminus\Psi_{\leq 0}(\rho)\}}\left\{\frac{\mathbb{P}(T_{ab}\geq t,T_{cd}\geq t)}{Q^2(t)}-1\right\}&\leq\frac{C(\#\Gamma^c_{\leq 0}(\gamma,c_0))^2}{p_1^2}\\
&\quad\times(\log(p))^{-1-\gamma_1}, \qquad (3.9.6)
\end{aligned}$$

where $\#\Gamma^c_{\leq 0}(\gamma,c_0)$ is the cardinality of the set $\Gamma^c_{\leq 0}(\gamma,c_0)$. By substituting (3.9.4) and (3.9.6) into (3.9.2), we have

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{\sum_{(a,b)\in\{S_{\leq 0}\setminus\Psi_{\leq 0}(\rho)\}}\{\mathbb{I}(T_{ab}\geq t)-Q(t)\}}{p_1Q(t)}\right)^2\right\}\\
&\leq\frac{C'o(p^{2\delta\rho})}{p_1^2}\exp\left((1-\rho)(3+\rho)t_p^2/4\right)+C(\log(p))^{-1-\gamma_1} \qquad (3.9.7)\\
&=\frac{C'o(p^{2\delta\rho})}{p_1^2}\times(\frac{p_1}{\log(p_1)})^{(1-\rho)(3+\rho)/2}+C(\log(p))^{-1-\gamma_1}=o(1),
\end{aligned}$$

where (3.9.7) is deduced using the fact that $\exp\left((1-\rho)(3+\rho)t^2/4\right)\leq\exp\left((1-\rho)(3+\rho)t_p^2/4\right)$ for all $t\leq t_p$, and the last equality follows by substituting $t_p=\sqrt{2\log(p_1)-2\log(\log(p_1))}$. □

CHAPTER 4

Estimation and inference for covariate adjusted Gaussian graphical models under an unbalanced distributed setting

This chapter proposes a distributed estimation and inferential framework for sparse multivariate regression and conditional Gaussian graphical models under the unbalanced splitting setting. This type of data splitting arises when the datasets from different sources cannot be aggregated on one single machine or when the available machines are of different powers. In this chapter, the number of covariates, responses and machines grow with the sample size, while sparsity is imposed. Debiased estimators of the coefficient matrix and of the precision matrix are proposed on every single machine and theoretical guarantees are provided. Moreover, new aggregated estimators that pool information across the machines using a pseudo log-likelihood function are proposed. It is shown that they enjoy efficiency and asymptotic normality as the number of machines grows with the sample size. The performance of these estimators is investigated via a simulation study and a real data example. It is shown empirically that the performances of these estimators are close to those of the non-distributed estimators which use the entire dataset. The content of this chapter is based on the paper

> Nezakati, E. and E. Pircalabelu (2023b). Estimation and inference in sparse multivariate regression and conditional Gaussian graphical models under an unbalanced distributed setting. pp. 1–50, *Submitted*.

4.1 Introduction

In multivariate linear regression, multiple response variables (say p) are regressed on multiple predictors (say q). This model has many applications in the real world,

especially in genomic data analysis, where one can model the dependence of RNA levels on DNA copy numbers through a multivariate regression model with RNA levels being responses and the DNA copy numbers being predictors as in Peng et al. (2010). Moreover, a natural way to characterize the regularization network between a set of genetic variants and a set of expressions is through multivariate regression models (see for example, Yin and Li, 2011, 2013; Wang et al., 2015). Estimation of the coefficient matrix in multivariate regression models requires one to estimate pq coefficients, which becomes challenging with high-dimensional predictors and responses. Moreover, by adjusting the effect of covariates on the mean of the random response, we are able to estimate the structure of a conditional graphical model constructed using the elements of the precision matrix related to the response vector. Most studies focusing on the estimation of the precision matrix for Gaussian graphical models assume that the random vector has zero or constant mean. For a treatment on the subject in the high-dimensional context for a mean zero Gaussian graphical model, we refer the reader to Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), Cai et al. (2011), Ravikumar et al. (2011), Cai et al. (2016), and Loh and Tan (2018) among many others. However, in many real applications, adjusting for the effect of covariates on the mean of the random vector is important for understanding the underlying graph structure.

Several studies used regularization-based approaches to estimate the coefficient and precision matrices with adjusted covariates. The works of Obozinski et al. (2011) and Yin and Li (2011) proposed a joint regularization penalty to estimate both the multivariate regression coefficients corresponding to the covariates and the precision matrix of a Gaussian graphical models iteratively. In Rothman et al. (2010) and Wang (2015), the coefficient and precision matrices were estimated simultaneously via a joint penalized likelihood function. The works of Cai et al. (2013) and Yin and Li (2013) proposed a two-stage strategy which first estimates the regression coefficients and then using the residuals from the first stage, estimates the precision matrix. In Chen et al. (2016), a tuning parameter free estimator was proposed that is asymptotically normal and efficient for the estimation of every finite sub-graph for covariate adjusted Gaussian graphical models. In these studies, due to the use of penalization, the estimators are biased. In this chapter, by introducing debiased estimators, we are able to perform not only the estimation, but also inference and hypothesis testing.

With the development of technology, the size of the datasets grows at a high rate, such that in certain situations it is not possible to store all needed datasets in the memory of one single machine. Moreover, in recent frameworks, like federated learning (McMahan et al., 2017), due to privacy concerns, it may be impossible to collect datasets from different resources on one single central machine. As such, the dataset is partitioned onto a cluster of machines. Distributed statistical approaches, also known as 'divide and conquer' approaches, have drawn a lot of attention in the last decade and have been developed for various statistical problems. The two

most popular techniques in distributed statistical inference and estimation problems are 'averaging' estimators from local machines and the 'one-step' approach, which combines the simple averaging estimator with a classical Newton's method to generate a one-step estimator. In Jordan et al. (2018), the authors presented a 'Communication-efficient Surrogate Likelihood' framework for solving distributed statistical inference problems in low-dimensional, high-dimensional and Bayesian frameworks. In Chen and Xie (2014), Lee et al. (2017) and Battey et al. (2018) the authors considered a high-dimensional sparse parameter vector estimation problem, where they adopted the penalized M-estimator setting. For a detailed review on aggregation methods for distributed estimators, the reader can refer to Huo and Cao (2019).

Nevertheless, most studies in the distributed setting have focused on the balanced sub-samples case, while in recent approaches like federated learning, some of the machines are more powerful than others and it is not efficient to distribute a dataset on different machines with equal sizes. In this situation, just taking a simple average is not an optimal approach for aggregating estimators. Recently, Nezakati and Pircalabelu (2023c) proposed a new weighted, aggregated estimator for the elements of the precision matrix in a zero-mean Gaussian graphical model, where the weights are a function of sub-sample sizes and the variances estimated based on the sub-samples. This paper was extensively presented in Chapter 2 of this thesis. In this chapter, we introduce new aggregated estimators for both the coefficient and precision matrices in multivariate regression models using a pseudo log-likelihood function which is constructed using the asymptotic distribution of the debiased estimators. It is shown empirically that these estimators perform better than the simple average in terms of accuracy and coverage probabilities. These estimators are constructed for the setting where the number of responses (p), covariates (q) and machines (K) grow with the sample size (n). As a consequence, sparsity assumptions are imposed on the true matrices as a function of p, q and n. Different upper bounds are derived on the number of machines to guarantee the consistency and asymptotic normality of the estimators.

The content of this chapter is organized as follows. Notation and preliminaries are presented in Section 4.2. The debiased distributed estimators and their statistical properties are provided for the coefficient matrix and the precision matrix separately in Sections 4.3 and 4.4, respectively. The final aggregated estimators for both target matrices are introduced in Section 4.5. Theoretical properties of these estimators are also investigated in this section. In Section 4.6, the performance of the estimators is evaluated by means of a controlled simulation study and in Section 4.7, the performance on a real data set is illustrated. We close with a discussion on the method in Section 4.8. Proofs of the supporting lemmas and theorems and more simulation results can be found in the Appendix.

4.2 Preliminaries

The multivariate regression model in this chapter is defined as

$$\dot{\mathbf{Y}} = \mathbf{\Gamma}^\top \dot{\mathbf{X}} + \dot{\boldsymbol{\varepsilon}}, \tag{4.2.1}$$

where $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top \in \mathbb{R}^p$ and $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top \in \mathbb{R}^q$ are the random response and covariate vectors, respectively. Moreover, the matrix $\mathbf{\Gamma}$ is a $q \times p$ regression coefficient matrix. Consider a random noise vector $\dot{\boldsymbol{\varepsilon}} = (\varepsilon^1, \ldots, \varepsilon^p)^\top \in \mathbb{R}^p$ independent of $\dot{\mathbf{X}}$, which follows a p-dimensional Gaussian distribution with mean zero, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. The components of $\dot{\mathbf{Y}}$ are mapped to the node set $\mathcal{V} = \{1, \ldots, p\}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ describes the set of edges between all pairs $(a, b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$. An undirected edge between nodes $a, b \in \mathcal{V}$ is drawn if $(a, b) \in \mathcal{E}$ and $(b, a) \in \mathcal{E}$. A pair (a, b) is included in the edge set $\mathcal{E}$ if and only if the variables Y^a and Y^b are conditionally dependent given $\dot{\mathbf{X}}$ and all remaining random variables in $\dot{\mathbf{Y}}$.

According to model (4.2.1), conditionally on $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, $\dot{\mathbf{Y}}$ follows a p-dimensional Gaussian distribution with mean vector $\mathbf{\Gamma}^\top \dot{\mathbf{x}}$, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta}$. Every off-diagonal entry (a, b) of $\mathbf{\Theta}$ is proportional to the partial correlation between Y^a and Y^b given $\dot{\mathbf{X}}$ and all other variables in $\dot{\mathbf{Y}}$. As such, a pair of variables in $\dot{\mathbf{Y}}$ is conditionally independent given all remaining variables of $\dot{\mathbf{Y}}$ and $\dot{\mathbf{X}}$, if and only if the corresponding entry in the precision matrix $\mathbf{\Theta}$ is zero. Denote $\mathbf{A}_{ab}$ by the (a, b)-th element of an arbitrary matrix $\mathbf{A}$. The support of the precision matrix $\mathbf{\Theta}$ is defined as the index set of its non-zero off-diagonal elements

$$S_1 := \{(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b : \quad \mathbf{\Theta}_{ab} \neq 0\},$$

with cardinality $s_1 = \#S_1$ and the maximum node degree or row sparsity of $\mathbf{\Theta}$ is defined as

$$d_1 := \max_{a \in \mathcal{V}} \#\{b \in \mathcal{V}, b \neq a : \quad \mathbf{\Theta}_{ab} \neq 0\}.$$

Analogously, the support of the regression coefficient matrix $\mathbf{\Gamma}$ is considered as

$$S_2 := \{(a, b) \in \{1, \ldots, q\} \times \{1, \ldots, p\} : \quad \mathbf{\Gamma}_{ab} \neq 0\},$$

which is the index set of its non-zero elements, with cardinality $s_2 = \#S_2$. Moreover, the support of the b-th column of the regression coefficient matrix $\mathbf{\Gamma}$ is defined as $S_2(b) = \{a \in \{1, \ldots, q\} : \quad \mathbf{\Gamma}_{ab} \neq 0\}$, $b = 1, \ldots, p$, with cardinality $s_2(b) = \#S_2(b)$.

Given n independent observations from the pair $(\dot{\mathbf{Y}}^\top, \dot{\mathbf{X}}^\top)$, our goal is to introduce distributed, debiased estimators for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ from model (4.2.1). Suppose that the n samples are randomly divided into K non-overlapping unbalanced sub-samples with size n_k for the k-th sub-sample, $k = 1, \ldots, K$, and denote by

$n_\dagger = \min_{1\leq k\leq K} n_k$, while $n = \sum_{k=1}^{K} n_k$. Suppose that $n_k/n \to c_k \in (0,1)$, as $n_k \to \infty$, such that $\lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1$. In this chapter, p and q can grow with n, such that they might be larger than n. However, we suppose that $\log(p) = o(n_\dagger)$ and $\log(q) = o(\sqrt{n_\dagger})$, where $o(\cdot)$ expresses the asymptotic behavior of a sequence and is defined below. The usual assumption in the high-dimensional context is that the underlying matrices (coefficient and precision matrices in this chapter) are sparse, which means that the number of non-zero elements in the matrices cannot grow too fast and a certain bound is imposed on it. This sparsity reflects that many predictors in the regression model are redundant and that the graph related to the precision matrix has a rather low number of edges. To impose this sparsity condition on the estimation, one common approach is to add an ℓ_1 penalty to the function which is going to be optimized. This kind of regularization effectively forces some of the elements to zero, thus resulting in sparse solutions. There exists a wide variety of methods making use of ℓ_1 regularization, see for example Friedman et al. (2008), Ravikumar et al. (2011) and van de Geer et al. (2014). For convenience of notation, denote by $\mathbf{Y}_k = [\dot{\mathbf{Y}}_{1,k}, \ldots, \dot{\mathbf{Y}}_{n_k,k}]^\top$ and $\boldsymbol{\xi}_k = [\dot{\boldsymbol{\varepsilon}}_{1,k}, \ldots, \dot{\boldsymbol{\varepsilon}}_{n_k,k}]^\top$ the k-th sub-sample of the response vector $\dot{\mathbf{Y}}$ and the random noise $\dot{\boldsymbol{\varepsilon}}$, respectively, both arranged as matrices of dimension $n_k \times p$, with $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$, $\dot{\boldsymbol{\varepsilon}}_{l,k} \in \mathbb{R}^p$ as the l-th row, $l = 1, \ldots, n_k$, and by $\mathbf{X}_k = [\dot{\mathbf{X}}_{1,k}, \ldots, \dot{\mathbf{X}}_{n_k,k}]^\top$ the k-th design matrix of dimension $n_k \times q$, with $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ as the l-th row, $l = 1, \ldots, n_k$. With this notation, the sample version of the regression model (4.2.1) corresponds to

$$\mathbf{Y}_k = \mathbf{X}_k \boldsymbol{\Gamma} + \boldsymbol{\xi}_k, \tag{4.2.2}$$

where the rows of $\boldsymbol{\xi}_k$ are i.i.d. p-dimensional Gaussian vectors with mean zero, covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Theta}$. In the next two sections, debiased estimators for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ based on the k-th sub-sample are derived.

4.3 Distributed estimator of the coefficient matrix using the k-th sub-sample

We make the following assumptions in this section.

(A1) The rows of the design matrix $\mathbf{X}_k$ are i.i.d. q-dimensional Gaussian vectors with mean zero and positive definite covariance matrix $\mathbf{Q}$, where $\max_a \mathbf{Q}_{aa} = O(1)$.

Assumption (A1) can be relaxed to a deterministic design matrix. However, in this case, one needs the mutual incoherence or irrepresentability condition on the design matrix to exhibit model selection consistency (Chapter 6 of Bühlmann and van de Geer, 2011). Another equivalent condition is the restricted eigenvalue condition (Bickel et al., 2009). Raskutti et al. (2010) showed that this condition holds for the random Gaussian design matrices. As such, one does not need any mutual incoherence assumption in this setting.

(A2) The eigenvalues of $\mathbf{Q}$ are bounded from above and below, i.e., there exists a constant $\Lambda_1 \geq 1$ such that $1/\Lambda_1 \leq \Lambda_{\min}(\mathbf{Q}) \leq \Lambda_{\max}(\mathbf{Q}) \leq \Lambda_1$, where $\Lambda_{\min}(\mathbf{Q})$ and $\Lambda_{\max}(\mathbf{Q})$ are the minimum and maximum eigenvalues of $\mathbf{Q}$, respectively.

Denote the maximum row sparsity of the inverse covariance matrix $\mathbf{Q}^{-1}$ with $d_2 := \max_{a\in\{1,\ldots,q\}} \#\{a' \in \{1,\ldots,q\}, a' \neq a : \quad \mathbf{Q}^{-1}_{aa'} \neq 0\}$. To find an initial estimator for the regression coefficient matrix $\mathbf{\Gamma}$ in the k-th sub-sample, following Yin and Li (2013), we minimize the following joint penalized residual sum of squares and denote by $\hat{\mathbf{\Gamma}}_k$,

$$\hat{\mathbf{\Gamma}}_k = \arg\min_{\mathbf{\Gamma}} \left\{ \frac{1}{2n_k} \text{trace}\{(\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})^\top (\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})\} + \rho_k \|\mathbf{\Gamma}\|_1 \right\}, \tag{4.3.1}$$

where $\rho_k > 0$ is a regularization parameter that forces $\hat{\mathbf{\Gamma}}_k$ to be sparse. The optimization problem (4.3.1) contains p decoupled Lasso regressions with q coefficients. This equation ignores the correlation among response variables when estimating the multiple regression coefficients. Rothman et al. (2010) showed empirically that only when the correlation of the errors is high, incorporation of such a dependency can lead to increased efficiency in estimating $\mathbf{\Gamma}$. Lemma 4.10.7 in Appendix 4.10, provides a bound on the ℓ_1 norm of the difference between $\hat{\mathbf{\Gamma}}_k$ and the true matrix $\mathbf{\Gamma}$ under additional regularity conditions.

Due to the ℓ_1 penalty which is added to this loss function, $\hat{\mathbf{\Gamma}}_k$ is a biased estimator. To obtain a debiased estimator of $\mathbf{\Gamma}$, we use the idea of van de Geer et al. (2014) by inverting the Karush-Kuhn-Tucker (KKT) conditions. They showcased this method in the linear regression and generalized linear models framework. Using the KKT conditions in (4.3.1), we have

$$-\mathbf{X}_k^\top \mathbf{Y}_k/n_k + \mathbf{X}_k^\top \mathbf{X}_k \hat{\mathbf{\Gamma}}_k/n_k + \rho_k \mathbf{B}_k = \mathbf{0}, \tag{4.3.2}$$

where $\mathbf{B}_k$ belongs to the sub-differential of the ℓ_1 norm evaluated at $\hat{\mathbf{\Gamma}}_k$. By adding and subtracting $\mathbf{X}_k^\top \boldsymbol{\xi}_k/n_k$ to (4.3.2), performing some algebra calculations and finally vectorizing, we get

$$\sqrt{n_k}\text{vec}(\hat{\mathbf{\Gamma}}_k^d - \mathbf{\Gamma}) = \mathbf{T}_k + \mathbf{R}_{k,\mathbf{\Gamma}}, \tag{4.3.3}$$

where $\text{vec}(\cdot)$ is an operator which converts the matrix into a column vector by stacking the columns of the matrix on top of one another. Moreover

$$\begin{aligned}
\hat{\mathbf{\Gamma}}_k^d &= \hat{\mathbf{\Gamma}}_k + \mathbf{M}_k \mathbf{X}_k^\top (\mathbf{Y}_k - \mathbf{X}_k \hat{\mathbf{\Gamma}}_k)/n_k, \\
\mathbf{T}_k &= \text{vec}(\mathbf{M}_k \mathbf{X}_k^\top \boldsymbol{\xi}_k)/\sqrt{n_k}, \\
\mathbf{R}_{k,\mathbf{\Gamma}} &= \sqrt{n_k}\text{vec}((\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k)(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})),
\end{aligned}$$

where $\mathbf{I}_q$ is the identity matrix of dimension $q \times q$ and $\mathbf{M}_k$ is a reasonable approximation for the inverse of the sample covariance matrix $\mathbf{C}_k = \mathbf{X}_k^\top \mathbf{X}_k/n_k$. As q can be larger than n_k, the sample covariance matrix $\mathbf{C}_k$ is not always invertible. As

such, we find $\mathbf{M}_k$ such that $\mathbf{M}_k\mathbf{C}_k \approx \mathbf{I}_q$. The estimator $\hat{\mathbf{\Gamma}}_k^d$, $k = 1, \ldots, K$, is the multivariate version of the one introduced in van de Geer et al. (2014). Note that the quantities in (4.3.3) are indexed by k, and as such one obtains a collection of debiased estimators $\hat{\mathbf{\Gamma}}_1^d, \ldots, \hat{\mathbf{\Gamma}}_K^d$, each using a particular sub-sample. In Section 4.5, we propose a novel, aggregated estimator based on the collection $\hat{\mathbf{\Gamma}}_1^d, \ldots, \hat{\mathbf{\Gamma}}_K^d$. In order to construct the aggregated estimator one needs the asymptotic distribution of $\hat{\mathbf{\Gamma}}_k^d$, $k = 1, \ldots, K$, which is investigated in the sequel.

Similarly to the case of a univariate response, by conditioning on $\mathbf{X}_k$, one can show the normality of $\mathbf{T}_k$ from (4.3.3). One just needs next to show that the remainder term $\mathbf{R}_{k,\mathbf{\Gamma}}$ vanishes with increasing n_k. To this end, we consider a suitable approximation for $\mathbf{M}_k$, such that with increasing n_k, the entries in $(\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k)$ get closer to zero. Several works (for instance, van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2018) assumed that $\mathbf{Q}^{-1}$ is sparse and then using the method of nodewise Lasso, estimated $\mathbf{Q}^{-1}$ and have set $\mathbf{M}_k = \hat{\mathbf{Q}}^{-1}$, where $\hat{\mathbf{Q}}^{-1}$ is the nodewise Lasso estimator of $\mathbf{Q}^{-1}$. We follow the same procedure, and for the reader's convenience, a brief description of the method is provided in Appendix 4.9.1. Furthermore, we need to control the randomness of the product between the noise matrix $\boldsymbol{\xi}_k$ and the design matrix $\mathbf{X}_k$ in the multivariate linear regression. Many studies, focusing on Lasso regression models, control the randomness of the noise by conditioning on an event of interest (see for instance, Bühlmann and van de Geer, 2011; Javanmard and Montanari, 2018). Similarly, considering the threshold ρ_0, such that $2\rho_0 \leq \rho_k$, recall that ρ_k is the regularization parameter in (4.3.1), we consider the event

$$\mathcal{F}_k(n_k, p, q) = \left\{ \|\mathrm{vec}(\mathbf{X}_k^\top \boldsymbol{\xi}_k)\|_\infty / n_k \leq \rho_0 \right\}.$$

Lemmas 4.10.1 and 4.10.2 in Appendix 4.10 show that for a suitable value of ρ_0, the event $\mathcal{F}_k(n_k, p, q)$ happens with a large probability for every fixed k and jointly in $k = 1, \ldots, K$, respectively. By controlling the randomness of the product between $\mathbf{X}_k$ and $\boldsymbol{\xi}_k$, one can control the ℓ_1 error bound on the estimation of the coefficient matrix. The reader can refer to Theorem 4.3.2 and its proof for more details. To show the asymptotic normality of $\mathbf{T}_k$ from (4.3.3), we need to impose as well a sparsity condition on the inverse covariance matrix $\mathbf{Q}^{-1}$. The sparsity condition in van de Geer et al. (2014) is considered as $d_2 = o(n/\log(q))$. In this chapter, we need to restrict the elements of $\mathbf{Q}^{-1}$ to a sparser regime such that the maximum row sparsity grows at the rate $d_2 = o(\sqrt{n_\dagger}/\log(q))$. With this sparsity condition, we show the asymptotic normality of the final aggregated estimator in Section 4.5 as this condition is needed in Lemma 4.10.10, where we show that under such an assumption two sequences are equivalent in probability. Moreover, due to the fact that both K and n_k, $k = 1, \ldots, K$, grow in the distributed case, we impose the sparsity condition $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$ on $\mathbf{\Gamma}$ to guarantee the theoretical properties of the aggregated estimator.

Theorem 4.3.1. *Consider the regression model* (4.2.2) *for the k-th sub-sample with zero-mean Gaussian noise matrix $\boldsymbol{\xi}_k$ having covariance matrix $\boldsymbol{\Sigma}$ and random design matrix $\mathbf{X}_k$, which satisfies assumptions* (A1) *and* (A2) *from Section 4.3 with sparsity condition $d_2 = o(\sqrt{n_\dagger}/\log(q))$. On the event $\mathcal{F}_k(n_k, p, q)$, with regularization parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$ in* (4.3.1) *and $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$, $j = 1, \ldots, q$, from the nodewise Lasso procedure, defined in Appendix 4.9.1, we have*

$$\mathbf{T}_k|\mathbf{X}_k \sim \mathcal{N}_{pq}(\mathbf{0}, \boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)), \quad \|\mathbf{R}_{k,\boldsymbol{\Gamma}}\|_\infty = O_p(s_2\sqrt{\log(q)\log(pq)/n_k}), \tag{4.3.4}$$

and under the additional sparsity condition $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$, we have $\|\mathbf{R}_{k,\boldsymbol{\Gamma}}\|_\infty = o_p(1)$.

The proof is given in Appendix 4.9.1.

One can use Theorem 4.3.1 to construct asymptotic inferential tools for $\boldsymbol{\Gamma}$ based on the k-th sub-sample. However, the covariance matrix $\boldsymbol{\Sigma}$ is in general unknown and it can be replaced in practice by a consistent estimator. Lemma 4.10.8 in Appendix 4.10 shows that if the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded, then under the random design setting, with sparsity condition $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$, the estimator $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)/n_k$ is a consistent estimator for the covariance matrix $\boldsymbol{\Sigma}$ with convergence rate in ℓ_∞ norm of order $O_p(\max\{\sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\})$.

Using (4.3.3), one can also obtain the convergence rate of the debiased estimator $\hat{\boldsymbol{\Gamma}}_k^d$. This bound is investigated in Theorem 4.3.2.

Theorem 4.3.2. *Consider the regression model* (4.2.2) *for the k-th sub-sample with design matrix $\mathbf{X}_k$ which satisfies assumptions* (A1) *and* (A2) *from Section 4.3. On the event $\mathcal{F}_k(n_k, p, q)$, with regularization parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$, we have*

$$\|\hat{\boldsymbol{\Gamma}}_k^d - \boldsymbol{\Gamma}\|_\infty = O_p(\max\{\sqrt{d_2\log(pq)/n_k}, s_2\sqrt{\log(q)\log(pq)}/n_k\}), \tag{4.3.5}$$

where under the assumption $\log(p)/\log(q) = o(n_\dagger^{\pi_2})$, $0 < \pi_2 < 1/2$, and the sparsity conditions $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$, we obtain $\|\hat{\boldsymbol{\Gamma}}_k^d - \boldsymbol{\Gamma}\|_\infty = o_p(1)$.

The proof is given in Appendix 4.9.2. In the next section, the distributed estimation of $\boldsymbol{\Theta}$ based on the k-th sub-sample is provided.

4.4 Distributed estimator of the precision matrix using the k-th sub-sample

Given the design matrix $\mathbf{X}_k$, the assumptions of Chapter 2 are adapted to our context and are reproduced here for the reader's convenience.

(B1) The eigenvalues of the precision matrix $\boldsymbol{\Theta}$ are bounded, i.e., there exists a constant $\Lambda_2 \geq 1$ such that $1/\Lambda_2 \leq \Lambda_{\min}(\boldsymbol{\Theta}) \leq \Lambda_{\max}(\boldsymbol{\Theta}) \leq \Lambda_2$, where $\Lambda_{\min}(\boldsymbol{\Theta})$ and $\Lambda_{\max}(\boldsymbol{\Theta})$ are the minimum and maximum eigenvalues of $\boldsymbol{\Theta}$, respectively. Moreover, $\max_a \boldsymbol{\Theta}_{aa} = O(1)$, where $\boldsymbol{\Theta}_{aa}$ is the a-th diagonal element of $\boldsymbol{\Theta}$.

Recall that $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)/n_k$ and consider $\mathbf{H}$ as the Hessian of the negative log-likelihood function proportional to $l(\boldsymbol{\Theta}) = \text{trace}(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\boldsymbol{\Theta})$ $- \log\det(\boldsymbol{\Theta})$. By definition, $\mathbf{H}$ is a $p^2 \times p^2$ matrix indexed by the pair of elements from the node set, such that $\mathbf{H} = [\mathbf{H}_{(a,b),(c,d)}]$, where $(a,b),(c,d) \in \mathcal{V} \times \mathcal{V}$. Let $\mathbb{S}_1 = \{S_1 \cup \{(1,1),(2,2),\ldots,(p,p)\}\}$ with cardinality ϖ, which is equal to $\varpi = s_1 + p$, and denote its complement set with $\mathbb{S}_1^c$.

(B2) The irrepresentability condition holds for the true precision matrix $\boldsymbol{\Theta}$, i.e., there exists $\alpha_1 \in (0,1]$ such that $\max_{e\in\mathbb{S}_1^c}\|\mathbf{H}_{e\mathbb{S}_1}(\mathbf{H}_{\mathbb{S}_1\mathbb{S}_1})^{-1}\|_1 \leq 1-\alpha_1$, where $\mathbf{H}_{\mathbb{S}_1\mathbb{S}_1} \in \mathbb{R}^{\varpi\times\varpi}$ is a sub-matrix of $\mathbf{H}$ whose rows and columns are indexed by the elements of $\mathbb{S}_1$. Moreover, e is a pair $(a,b) \in \mathbb{S}_1^c$ such that $\mathbf{H}_{e\mathbb{S}_1}$ is an ϖ-dimensional column vector with elements $\mathbf{H}_{e,(c,d)}$, where $(c,d) \in \mathbb{S}_1$.

Given $\mathbf{X}_k$ and using $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)/n_k$, one can construct an estimator for $\boldsymbol{\Theta}$ using the following graphical Lasso optimization problem

$$\hat{\boldsymbol{\Theta}}_k = \arg\min_{\boldsymbol{\Theta}\in\mathcal{S}_{++}^p}\Big\{\text{trace}(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \lambda_k\|\boldsymbol{\Theta}\|_{1,\text{off}}\Big\}, \tag{4.4.1}$$

where $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal norm of the matrix defined as $\|\boldsymbol{\Theta}\|_{1,\text{off}} = \sum_{a\neq b}|\boldsymbol{\Theta}_{ab}|$ and $\mathcal{S}_{++}^p$ is the space of positive definite matrices of dimension $p \times p$. To obtain a debiased estimator of $\boldsymbol{\Theta}$, we invert the KKT conditions from the optimization problem (4.4.1), which is of the form

$$\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \hat{\boldsymbol{\Theta}}_k^{-1} + \lambda_k\hat{\mathbf{D}}_k = \mathbf{0}, \tag{4.4.2}$$

where the matrix $\hat{\mathbf{D}}_k$ belongs to the sub-differential of the ℓ_1 off-diagonal norm evaluated at $\hat{\boldsymbol{\Theta}}_k$. The difference between (4.4.2) and the problem in Chapter 2 is that the previous work used the sample covariance matrix $\mathbf{Y}_k^\top\mathbf{Y}_k/n_k$, but here due to the non-zero mean of the model, we use $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ which contains the plug-in estimation of the regression coefficient matrix $\boldsymbol{\Gamma}$ therein, thus making the analysis more complicated. To simplify notation, define

$$\mathbf{W}_k = \hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{\Gamma}} - \boldsymbol{\Sigma}, \quad \mathbf{W}_k' = \hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}, \quad \mathbf{W}_k'' = \hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{\Gamma}},$$

where

$$\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{\Gamma}} = (\mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\Gamma})^\top(\mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\Gamma})/n_k.$$

Using the KKT condition (4.4.2) and performing some algebra calculations, leads to

$$\sqrt{n_k}(\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}) = -\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta} + \mathbf{R}_{k,\mathbf{\Theta}}, \tag{4.4.3}$$

where $\hat{\mathbf{\Theta}}_k^d = 2\hat{\mathbf{\Theta}}_k - \hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\hat{\mathbf{\Theta}}_k$ is the k-th debiased estimator of $\mathbf{\Theta}$ and the term $\mathbf{R}_{k,\mathbf{\Theta}}$ is defined as

$$\mathbf{R}_{k,\mathbf{\Theta}} := -\sqrt{n_k}(\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p)(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}) - \sqrt{n_k}(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{W}_k'\mathbf{\Theta} - \sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k''\mathbf{\Theta}. \tag{4.4.4}$$

In the sequel, it is shown that under suitable conditions, $\|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty = o_p(1)$ and that the term $\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}$ is elementwise asymptotically normal.

Relative to the work of Jankova and van de Geer (2015), in addition to different covariance matrices used in (4.4.4), our remainder contains the extra term $\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k''\mathbf{\Theta}$. This extra term is bounded in elementwise ℓ_∞ norm at the rate of $O_p(d_1 s_2 \log(pq)/\sqrt{n_k})$, as it is shown in Lemma 4.4.1. In the estimation procedure, if one considers $\mathbf{\Gamma} = \mathbf{0}$ as a special case, then the KKT condition (4.4.2) will simplify to the one in Jankova and van de Geer (2015), and as such the estimator we propose, $\hat{\mathbf{\Theta}}_k^d$, will simplify to the same debiased estimator with the same convergence rate from that work. We recall $\kappa_{\mathbf{\Sigma}} := |||\mathbf{\Sigma}|||_\infty$ as the matrix ℓ_∞ norm of $\mathbf{\Sigma}$ and $\kappa_{\mathbf{H}} = |||(\mathbf{H}_{\mathbb{S}_1,\mathbb{S}_1})^{-1}|||_\infty$ as the matrix ℓ_∞ norm of the inverse of $\mathbf{H}_{\mathbb{S}_1,\mathbb{S}_1}$ from the irrepresentability condition (B2). In Lemma 4.4.1, negligibility of $\mathbf{R}_{k,\mathbf{\Theta}}$ is shown.

Lemma 4.4.1. Consider the multivariate regression model (4.2.2) with random Gaussian noise matrix $\boldsymbol{\xi}_k$ having zero-mean rows, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta}$. Let assumptions (B1)-(B2), and (C1)-(C3) from Appendix 4.9.3 hold. Consider the debiased estimator in (4.4.3) with regularization parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$. On the event $\mathcal{F}_k(n_k, p, q)$ with $2\rho_0 \leq \rho_k$, where $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and under the assumptions $1/\alpha_1 = O(1)$, $\kappa_{\mathbf{\Sigma}} = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$, we have

$$\|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty = O_p\bigg(\max\{d_1^{3/2}\log(p)/\sqrt{n_k}, d_1^2(\log(p))^{3/2}/n_k, d_1 s_2 \log(pq)/\sqrt{n_k}\}\bigg), \tag{4.4.5}$$

and under the sparsity conditions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, where $0 < \pi_3 \leq 1/6$, we get $\|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty = o_p(1)$.

The proof is given in Appendix 4.9.3.

Remark 4.1. A similar convergence rate for the remainder term is obtained in Jankova and van de Geer (2015) but for the zero-mean model, which is of order $O_p(\max\{d_1^{3/2}\log(p)/\sqrt{n_k}, d_1^2(\log(p))^{3/2}/n_k\})$. Comparing it with (4.4.5), it is observed that the convergence rate of the extra term $\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k''\mathbf{\Theta}$ is of order $O_p(d_1 s_2 \log(pq)/\sqrt{n_k})$ which also adds more constraints on the negligibility of $\mathbf{R}_{k,\mathbf{\Theta}}$.

Theorem 4.4.2 investigates the elementwise asymptotic normality of the debiased estimator $\hat{\mathbf{\Theta}}_k^d$, which will be leveraged further in Section 4.5 to construct an aggregated estimator using all K estimators $\hat{\mathbf{\Theta}}_1^d, \ldots, \hat{\mathbf{\Theta}}_K^d$.

Theorem 4.4.2. *Under the assumptions of* Lemma 4.4.1 *and the sparsity conditions* $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ *and* $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, *for all* $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, *it holds that*

$$\sqrt{n_k}(\hat{\mathbf{\Theta}}_{ab,k}^d - \mathbf{\Theta}_{ab})/\sigma_{ab} = \mathbf{Z}_{ab,k} + o_p(1), \tag{4.4.6}$$

where $\mathbf{\Theta}_{ab}$ *and* $\hat{\mathbf{\Theta}}_{ab,k}^d$ *are the* (a,b)*-th element of* $\mathbf{\Theta}$ *and* $\hat{\mathbf{\Theta}}_k^d$, *respectively, and* $\sigma_{ab}^2 = \mathbf{\Theta}_{aa}\mathbf{\Theta}_{bb} + \mathbf{\Theta}_{ab}^2$. *Moreover,* $1/\sigma_{ab} = O(1)$ *and* $\mathbf{Z}_{ab,k}$ *converges weakly to* $\mathcal{N}(0,1)$ *as* n_k *grows.*

The proof is given in Appendix 4.9.4.

Remark 4.2. One can use Theorem 4.4.2 to construct asymptotic confidence intervals and hypothesis testing procedures. However, to perform inference, one needs a consistent estimator for σ_{ab} as it is unknown. By similar arguments as in Lemma 2 of Jankova and van de Geer (2015) and considering $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, the estimator $\hat{\sigma}_{ab,k}^2 = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}_{ab,k}^2$ is a consistent estimator for σ_{ab}^2 with convergence rate $O_p(\max\{\log(p)/n_k, \sqrt{d_1 \log(p)/n_k}\})$.

Remark 4.3. To quantify the convergence rate of the debiased estimator $\hat{\mathbf{\Theta}}_k^d$, from (4.4.3), we have that

$$\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty \leq |||\mathbf{\Theta}|||_\infty^2 \|\mathbf{W}_k\|_\infty + \|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty/\sqrt{n_k}.$$

Given $\mathbf{X}_k$ and under assumption (B1), by considering $\boldsymbol{\xi}_k = \mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma}$ and setting $\mathbf{A} = \mathbf{B} = \mathbf{I}_p$ in Lemma 2 of Lam and Fan (2009), for which the conditions are fulfilled, one can write $\|\mathbf{W}_k\|_\infty = O_p(\sqrt{\log(p)/n_k})$. Combining this bound with (4.4.5) and (4.9.9) from Appendix 4.9.3, we get

$$\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty = O_p\Big(\max\{d_1\sqrt{\log(p)/n_k}, d_1^{3/2}\log(p)/n_k, \\ d_1^2(\log(p)/n_k)^{3/2}, d_1 s_2 \log(pq)/n_k\}\Big),$$

and under the sparsity conditions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, the consistency of $\hat{\mathbf{\Theta}}_k^d$ follows.

4.5 The final aggregated estimators across the sub-samples

The results in the previous sections are provided at the level of the k-th sub-sample, $k = 1, \ldots, K$. To construct more reliable estimators, we aggregate estimators

drawn from different distributed locations. There are multiple ways to aggregate these K estimators into a combined estimator. The aggregation procedure in this chapter is the same as the one in Chapter 2, and for the reader's convenience, it is again provided here.

Due to the asymptotic normal distribution of the local estimators, we derive the final aggregated estimator by maximizing a pseudo log-likelihood function, which is constructed using the asymptotic normal density of local estimators constructed based on the sub-samples. Consider Δ as a general parameter of interest, for example $\mathbf{\Theta}_{ab}$ or $\text{vec}(\mathbf{\Gamma})_a$ in this chapter, where $\text{vec}(\mathbf{\Gamma})_a$ is the a-th element of the vectorized form of $\mathbf{\Gamma}$, and $\hat{\Delta}_k$ as the k-th estimator based on the k-th sub-sample with variance σ^2 and denote its consistent estimator by $\hat{\sigma}_k^2$ which is also derived based on the k-th sub-sample. Consider the asymptotic normal density of $\hat{\Delta}_k$ at point ι_k as $\hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)$, where the variance σ^2 is replaced by its consistent estimator $\hat{\sigma}_k^2$. By maximizing the pseudo log-likelihood function constructed as

$$l(\Delta) = \log\Big(\prod_{k=1}^{K} \hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)\Big) \propto \sum_{k=1}^{K}(-n_k/2)(\iota_k - \Delta)^2/\hat{\sigma}_k^2,$$

with respect to Δ, the final aggregated estimator is of the form

$$\tilde{\Delta}_{\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2}\hat{\Delta}_k, \tag{4.5.1}$$

where the subscript "owAvg" stands for the optimally weighted average. We call this estimator an "optimally weighted average", as it is obtained using a maximization problem. Substituting debiased estimators $\hat{\mathbf{\Gamma}}_k^d$ and $\hat{\mathbf{\Theta}}_k^d$ in (4.5.1), one can aggregate distributed estimators of the coefficient and precision matrices from the sub-samples into the final combined estimators $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$, respectively. Note that if the variance σ^2 is known, then replacing $\hat{\sigma}_k^2$ by σ^2 simplifies the aggregated estimator $\tilde{\Delta}_{\text{owAvg}}$ to two special cases. In the case of unbalanced sub-samples, $\tilde{\Delta}_{\text{owAvg}}$ is simplified to the sample size weighted average estimator

$$\tilde{\Delta}_{\text{wAvg}} = \sum_{k=1}^{K} \frac{n_k}{n}\hat{\Delta}_k, \tag{4.5.2}$$

where "wAvg" stands for the sample size weighted average proportional to the sub-sample sizes. Similarly to the optimally weighted average estimator, by substituting debiased estimators $\hat{\mathbf{\Gamma}}_k^d$ and $\hat{\mathbf{\Theta}}_k^d$ in (4.5.2), the weighted average aggregated estimators of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ can be constructed and denoted by $\tilde{\mathbf{\Gamma}}_{\text{wAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$, respectively. Moreover, if the sub-samples are also balanced, $\tilde{\Delta}_{\text{owAvg}}$ further simplifies to the simple average estimator

$$\tilde{\Delta}_{\text{sAvg}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\Delta}_k, \tag{4.5.3}$$

where "sAvg" stands for the simple average. By the same argument as for the optimally weighted and sample size weighted averages, the simple average estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ can be constructed and denoted by $\tilde{\boldsymbol{\Gamma}}_{\text{sAvg}}$ and $\tilde{\boldsymbol{\Theta}}_{\text{sAvg}}$, respectively.

Using (4.3.3) and (4.5.1), by considering the consistent estimator $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ for $\boldsymbol{\Sigma}$, the optimally weighted average estimator for the a-th element of $\text{vec}(\boldsymbol{\Gamma})$, $a = 1, \ldots, qp$, is of the form

$$\begin{aligned}\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{owAvg}})_a = &\left(\sum_{k=1}^{K} \frac{n_k}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\right)^{-1} \\ &\times \sum_{k=1}^{K} \frac{n_k}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \text{vec}(\hat{\boldsymbol{\Gamma}}_k^d)_a. \end{aligned} \tag{4.5.4}$$

It can be shown that

$$\sqrt{\sum_{k=1}^{K} n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}(\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{owAvg}})_a - \text{vec}(\boldsymbol{\Gamma})_a) = \mathbf{W}_{a,\boldsymbol{\Gamma}} + \mathbf{R}_{a,\boldsymbol{\Gamma}}, \tag{4.5.5}$$

where

$$\begin{aligned}\mathbf{R}_{a,\boldsymbol{\Gamma}} &= \sqrt{\frac{\sum_{k=1}^{K} n_k/[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \mathbf{R}_{a,k,\boldsymbol{\Gamma}}, \\ \mathbf{W}_{a,\boldsymbol{\Gamma}} &= \sqrt{\frac{\sum_{k=1}^{K} n_k/[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \mathbf{T}_{a,k},\end{aligned}$$

and $\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$ and $\mathbf{T}_{a,k}$ are the a-th element of $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ and $\mathbf{T}_k$, respectively, defined in (4.3.3).

Similarly, using (4.4.3) and (4.5.1), by considering the consistent estimator $\hat{\sigma}^2_{ab,k}$ for σ^2_{ab} in (4.4.6), the aggregated estimator for the (a,b)-th element of $\boldsymbol{\Theta}$, $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, is of the form

$$\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}} = \left(\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}\right)^{-1} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \hat{\boldsymbol{\Theta}}^d_{ab,k}. \tag{4.5.6}$$

Moreover,

$$\sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}}\left(\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}} - \boldsymbol{\Theta}_{ab}\right) = \mathbf{W}_{ab,\boldsymbol{\Theta}} + \mathbf{R}_{ab,\boldsymbol{\Theta}}, \tag{4.5.7}$$

where

$$\mathbf{R}_{ab,\boldsymbol{\Theta}} = \sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}} \sum_{k=1}^K \frac{\sqrt{n_k}\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2} \mathbf{R}_{ab,k,\boldsymbol{\Theta}},$$

$$\mathbf{W}_{ab,\boldsymbol{\Theta}} = \sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}} \sum_{k=1}^K \frac{\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k}) \times (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

and $\mathbf{R}_{ab,k,\boldsymbol{\Theta}}$ is the (a,b)-th element of $\mathbf{R}_{k,\boldsymbol{\Theta}}$ from (4.4.3), $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$ and $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ are the l-th row of $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively, while $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ are p-dimensional vectors coming from the a-th row and the b-th row of $\boldsymbol{\Theta}$, respectively.

Theorem 4.5.1. *Consider the regression model* (4.2.2)*, where* $\mathbf{X}_k$ *satisfies assumptions* (A1) *and* (A2) *from Section 4.3, and consider the maximum row sparsity of* $\mathbf{Q}^{-1}$ *as* $d_2 = o(\sqrt{n_\dagger}/\log(q))$*. Moreover, consider the coefficient matrix* $\boldsymbol{\Gamma}$ *with sparsity condition* $s_2 = o(n_\dagger^{\pi_1}/(\log(pq)\log(q)))$*,* $0 < \pi_1 \leq 1/2$*. Suppose that the event* $\mathcal{F}_k(n_k, p, q)$ *holds jointly in* $k = 1, \ldots, K$*. Moreover, suppose that* $\hat{\boldsymbol{\Gamma}}_k^d$*,* $k = 1, \ldots, K$*, is the* k*-th debiased estimator in* (4.3.3) *with tuning parameter* $\rho_k \asymp \sqrt{\log(pq)/n_k}$ *and let* $n_k/n \to c_k \in (0,1)$ *as* n_k *grows such that* $\lim_{K\to\infty} \sum_{k=1}^K c_k = 1$*.*

a) If K *grows at the rate of* $K = O(n^{1/4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\}))$*, and* $\log(p)/\log(q) = o(n_\dagger^{\pi_2})$*,* $0 < \pi_2 < 1/2$*, for the* a*-th element of the vectorized form of* $\tilde{\boldsymbol{\Gamma}}_{\text{owAvg}}$*,* $a = 1, \ldots, qp$*, in* (4.5.5)*, we have*

$$\mathbf{W}_{a,\boldsymbol{\Gamma}} \xrightarrow{d} \mathcal{N}(0,1), \quad \textit{and} \quad |\mathbf{R}_{a,\boldsymbol{\Gamma}}| = o_p(1), \tag{4.5.8}$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

b) If K *grows at the rate of* $K = O(n^{1/3}/(\sqrt{\log(pq)\log(q)}\max\{s_2, d_2\}))$*, in the spirit of the special case* (4.5.2)*, for the* a*-th element of the vectorized form of* $\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}}$*,* $a = 1, \ldots, qp$*, which is denoted by* $\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}})_a$*, we have*

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^K [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} (\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}})_a - \text{vec}(\boldsymbol{\Gamma})_a) = \mathbf{W}'_{a,\boldsymbol{\Gamma}} + \mathbf{R}'_{a,\boldsymbol{\Gamma}},$$

where $\mathbf{W}'_{a,\boldsymbol{\Gamma}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}'_{a,\boldsymbol{\Gamma}}| = o_p(1)$*.*

c) If K *grows at the rate of* $\sqrt{K} = O(n_\dagger^{1/3}/(\sqrt{\log(pq)\log(q)}\max\{s_2, d_2\}))$*, in the spirit of the special case* (4.5.3)*, for the* a*-th element of the vectorized*

form of $\tilde{\mathbf{\Gamma}}_{\text{sAvg}}$, $a = 1, \ldots, qp$, *which is denoted by* $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a$, *we have*

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^{K} 1/n_k)}}(\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a - \text{vec}(\mathbf{\Gamma})_a)$$
$$= \mathbf{W}''_{a,\mathbf{\Gamma}} + \mathbf{R}''_{a,\mathbf{\Gamma}},$$

where $\mathbf{W}''_{a,\mathbf{\Gamma}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}''_{a,\mathbf{\Gamma}}| = o_p(1)$.

The proof is given in Appendix 4.9.5.

Remark 4.4. Note that Theorem 4.3.1 in Section 4.3 provides the asymptotic normality, conditionally on the design matrix $\mathbf{X}_k$, while the asymptotic normality result in Theorem 4.5.1 is more general as it is an unconditional result. By using a similar argument to the proof of Theorem 4.5.1, it can be shown that both $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ have an asymptotic variance equal to $[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}$ which implies that their asymptotic relative efficiency is equal to 1, and they are as efficient as the full estimator using the full sample data. This result is expected because (i) the definitions of $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ are closely related, and (ii) the estimated variances based on the sub-samples are consistent estimators for the true variance, i.e., $[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}/[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa} \xrightarrow{p} 1$. However, it has been brought to our attention that the same final result can be obtained via the asymptotic normality result characterizing M estimators (see for instance, Theorem 5.21 of van der Vaart, 2000) since the proposed estimators are asymptotically linear and hence asymptotically equivalent to the estimator obtained using the full sample data. A comparison of the owAvg and wAvg estimators from a finite sample perspective is also provided through simulation and real data examples in Sections 4.6 and 4.7, respectively.

In Theorem 4.5.2, the asymptotic normality of the estimator in (4.5.6) and its two special cases is investigated as $n_\dagger$ and K both grow.

Theorem 4.5.2. *Consider the regression model* (4.2.2) *and suppose that assumptions* (B1)-(B2), *and* (C1)-(C3) *from Appendix 4.9.3 hold. Moreover, consider the coefficient matrix* $\mathbf{\Gamma}$ *with sparsity condition* $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$. *Suppose that the event* $\mathcal{F}_k(n_k, p, q)$ *holds jointly in* $k = 1, \ldots, K$. *Moreover, suppose that* $\hat{\mathbf{\Theta}}_k^d$, $k = 1, \ldots, K$, *is the* k*-th debiased estimator in* (4.4.3) *with tuning parameter* $\lambda_k \asymp \sqrt{\log(p)/n_k}$ *and let* $n_k/n \to c_k \in (0,1)$ *as* n_k *grows, such that* $\lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1$.

a) If K *and the maximum node degree of* $\mathbf{\Theta}$ *grow at the rates of* $K = O(n^{1/3}/(d_1 \log(p)))$, *and* $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, *respectively, then for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, *of the pooled estimator* $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ *in* (4.5.7), *we have*

$$\mathbf{W}_{ab,\mathbf{\Theta}} \xrightarrow{d} \mathcal{N}(0,1), \quad \textit{and} \quad |\mathbf{R}_{ab,\mathbf{\Theta}}| = o_p(1). \tag{4.5.9}$$

b) Under the same conditions as in part a), in the spirit of the special case (4.5.2)*, for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}$*,* $a \neq b$*, of* $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$*, which is denoted by* $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$*, we have*

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k}}}(\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} - \mathbf{\Theta}_{ab}) = \mathbf{W}'_{ab,\mathbf{\Theta}} + \mathbf{R}'_{ab,\mathbf{\Theta}},$$

where $\mathbf{W}'_{ab,\mathbf{\Theta}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}'_{ab,\mathbf{\Theta}}| = o_p(1)$*.*

c) If K *and the maximum node degree of* $\mathbf{\Theta}$ *grow at the rates of* $K = O(n_\dagger^{4/3}/n)$ *and* $d_1^{3/2} = o(n_\dagger^{1/3}/\log(p))$*, in the spirit of the special case* (4.5.3)*, for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}$*,* $a \neq b$*, of* $\tilde{\mathbf{\Theta}}_{\text{sAvg}}$*, which is denoted by* $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$*, we have*

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K} 1/n_k)}}(\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} - \mathbf{\Theta}_{ab}) = \mathbf{W}''_{ab,\mathbf{\Theta}} + \mathbf{R}''_{ab,\mathbf{\Theta}},$$

where $\mathbf{W}''_{ab,\mathbf{\Theta}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}''_{ab,\mathbf{\Theta}}| = o_p(1)$*.*

The proof is given in Appendix 4.9.6.

Remark 4.5. By a similar argument to that of Remark 4.4, it can be shown that the asymptotic variances of $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ and $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ are both equal to σ^2_{ab}, where we recall that $\sigma^2_{ab} = \mathbf{\Theta}_{aa}\mathbf{\Theta}_{bb} + \mathbf{\Theta}^2_{ab}$. Additionally, by rewriting Theorem 4.4.2 for the full sample estimator, which uses the entire dataset of size n, it can be shown that the asymptotic variance of the debiased full estimator is also equal to σ^2_{ab}. Therefore, we can deduce that both the optimally weighted and the weighted distributed estimators are asymptotically as efficient as the debiased full sample estimator, which implies that the efficiency loss from the distributed setting is asymptotically zero.

According to the asymptotic normal distribution of the proposed estimators, one can construct confidence intervals and perform hypothesis testing for the elements of the coefficient matrix $\mathbf{\Gamma}$ and of the precision matrix $\mathbf{\Theta}$. The $(1-\alpha)100\%$ asymptotic confidence intervals for a general quantity of interest Δ, namely $\mathbf{\Theta}_{ab}$ or $\text{vec}(\mathbf{\Gamma})_a$ in this chapter, using the optimally weighted, weighted and simple average

estimators can be constructed as

$$\tilde{\Delta}_{\text{owAvg}} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}_k^2}, \tag{4.5.10}$$

$$\tilde{\Delta}_{\text{wAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)/(nK)}, \tag{4.5.11}$$

$$\tilde{\Delta}_{\text{sAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)(\sum_{k=1}^{K} 1/n_k)/K^3}, \tag{4.5.12}$$

where $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-th quantile of standard normal distribution and $\hat{\sigma}_k^2$ is the k-th consistent estimator of σ^2. Substituting $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$, $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a$, respectively in (4.5.10), (4.5.11) and (4.5.12) and then replacing the estimated variance $\hat{\sigma}_k^2$ by $[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$, one can construct $(1-\alpha)100\%$ asymptotic confidence intervals for the a-th element, $a = 1, \ldots, qp$, of the vectorized from of the coefficient matrix $\mathbf{\Gamma}$. Similarly, substituting $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$, $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ and $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$, respectively in (4.5.10), (4.5.11) and (4.5.12) and then replacing the estimated variance $\hat{\sigma}_k^2$ by $\hat{\sigma}_{ab,k}^2 = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}_{ab,k}^2$, the $(1-\alpha)100\%$ asymptotic confidence intervals for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$, of the precision matrix $\mathbf{\Theta}$ can be constructed. Moreover, by substituting the same quantities in

$$|\tilde{\Delta}_{\text{owAvg}}| > \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}_k^2},$$

$$|\tilde{\Delta}_{\text{wAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)/(nK)},$$

$$|\tilde{\Delta}_{\text{sAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)(\sum_{k=1}^{K} 1/n_k)/K^3},$$

the rejection regions at level α for the hypothesis tests $H_{0,a} : \text{vec}(\mathbf{\Gamma})_a = 0$ vs $H_{1,a} : \text{vec}(\mathbf{\Gamma})_a \neq 0$, where $a = 1, \ldots, pq$ and $H_{0,ab} : \mathbf{\Theta}_{ab} = 0$ vs $H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0$, where $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$ can be constructed.

As it was mentioned, the distributed estimator has an efficiency loss approaching zero as the sample size grows. This allows us to compare its incurred loss with that of the practically unavailable, full-sample debiased estimator. Considering the

definition of $\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}})_a$, it can be written

$$\begin{aligned}
&\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}})_a - \mathrm{vec}(\mathbf{\Gamma})_a \\
&= \frac{\sum_{k=1}^K n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^K n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \times \sum_{k=1}^K \frac{\sqrt{n_k}[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{T}_{a,k} \\
&+ \frac{\sum_{k=1}^K n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^K n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \times \sum_{k=1}^K \frac{\sqrt{n_k}[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{R}_{a,k,\mathbf{\Gamma}}.
\end{aligned}$$

Since $[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} \xrightarrow{p} 1$ and $\sum_{k=1}^K(n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})/\sum_{k=1}^K(n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}) \xrightarrow{p} 1$ as $K\to\infty$ and $n_k\to\infty$, $k=1,\ldots,K$, using similar arguments to the proof of Theorem 4.3.2, we can write

$$\begin{aligned}
\|\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}} - \mathbf{\Gamma}\|_\infty &= \max_{a\in\{1,\ldots,qp\}} |\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}})_a - \mathrm{vec}(\mathbf{\Gamma})_a| \\
&= O_p(K\max\{\sqrt{d_2\log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}), \quad (4.5.13)
\end{aligned}$$

where the last equality is deduced using (4.3.4), (4.9.5) and (4.10.20). By using the full sample data in Theorem 4.3.2, the convergence rate of the debiased full estimator, call it $\hat{\mathbf{\Gamma}}_F^d$, is of order

$$\|\hat{\mathbf{\Gamma}}_F^d - \mathbf{\Gamma}\|_\infty = O_p(\max\{\sqrt{d_2\log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}). \quad (4.5.14)$$

Combining the triangle inequality with (4.5.13) and (4.5.14), it is deduced that

$$\|\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}} - \hat{\mathbf{\Gamma}}_F^d\|_\infty = O_p(K\max\{\sqrt{d_2\log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}),$$

which implies that the convergence rate of the distance between the distributed estimator and the full one is equal to the rate of the distance between the distributed estimator and the true matrix $\mathbf{\Gamma}$. As such, it is noteworthy that $\tilde{\mathbf{\Gamma}}_{\mathrm{owAvg}}$ not only follows an asymptotic normal distribution, but also approximates the debiased full estimator $\hat{\mathbf{\Gamma}}_F^d$ well, and it exhibits a similar statistical error as $\hat{\mathbf{\Gamma}}_F^d$ as long as the number of machines K is not too large.

By a similar argument, one can derive the convergence rate of $\tilde{\mathbf{\Theta}}_{\mathrm{owAvg}}$ to be of the order

$$\begin{aligned}
\|\tilde{\mathbf{\Theta}}_{\mathrm{owAvg}} - \mathbf{\Theta}\|_\infty = O_p(K\max\{&d_1\sqrt{\log(p)/n}, d_1^{3/2}\log(p)/n, \\
&d_1^2(\log(p))^{3/2}/(n\sqrt{n_\dagger}), d_1 s_2\log(pq)/n\}), \quad (4.5.15)
\end{aligned}$$

which is obtained by combining the definition of $\tilde{\mathbf{\Theta}}_{\mathrm{owAvg}}$ with the rate in (4.4.5), the upper bound on $\|\mathbf{W}_k\|_\infty$ from Remark 4.3 and (4.9.9) from Appendix 4.9.3. Comparing this convergence rate with the one from Remark 4.3 for the full sample data estimator, call it $\hat{\mathbf{\Theta}}_F^d$, which is of the order

$$\begin{aligned}
\|\hat{\mathbf{\Theta}}_F^d - \mathbf{\Theta}\|_\infty = O_p(\max\{&d_1\sqrt{\log(p)/n}, d_1^{3/2}\log(p)/n, \\
&d_1^2(\log(p)/n)^{3/2}, d_1 s_2\log(pq)/n\}),
\end{aligned}$$

one can achieve the same conclusion as for the estimation of $\mathbf{\Gamma}$, which implies that if K is not too large, $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ attains a similar statistical error as the debiased full estimator $\hat{\mathbf{\Theta}}_F^d$.

Remark 4.6. Comparing the bound given by (4.5.15) with the one presented in (2.4.10) from Chapter 2, it is observed that the difference in the convergence rates between a covariate adjusted Gaussian graphical model with a zero-mean Gaussian graphical model is within the term $d_1 s_2 \log(pq)/n$. However, when considering the condition $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, the cardinality of the non-zero entries of $\mathbf{\Gamma}$ does not grow fast, such that $\mathbf{\Gamma}$ will be much sparser than $\mathbf{\Theta}$, and as a result, the term $d_1 s_2 \log(pq)/n$ will be dominated by the other terms in the bound (4.5.15). This result can be also investigated by a comparison of the bounds on K between Theorem 4.5.2 and Theorem 2.4.2 from Chapter 2.

In the next section, the statistical error and coverage probability of estimators are compared from a finite sample perspective.

4.6 Simulation study

In this section, we examine empirically the performance of our proposed estimators. To this end, we followed the simulation setup of Yin and Li (2013) for generating sparse matrices $\mathbf{\Gamma}$ and $\mathbf{\Theta}$.

First, to generate the precision matrix $\mathbf{\Theta}$, we randomly generated a link between all pairs $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, with probability of connection of 0.01. Then, the corresponding entry in the precision matrix is generated uniformly from $[-1, -0.5] \bigcup [0.5, 1]$, for each link. After that, for each row, each entry except the diagonal one is divided by the sum of the absolute values of the off-diagonal entries multiplied by 1.5. Finally, the matrix is symmetrized and the diagonal entries are fixed to 1. To generate the regression coefficient matrix, we first generated a sparse indicator matrix with non-zero elements having a probability of 0.01 of occurring for every pair $(a, b); a = 1, \ldots, q, b = 1, \ldots, p$. Then, corresponding to the non-zero entries of this indicator matrix, we generated uniformly the entries of $\mathbf{\Gamma}$ from $[-1, -\nu_m] \bigcup [\nu_m, 1]$, where ν_m is the minimum absolute non-zero value of the generated precision matrix. To generate a dataset, we first generated $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top$ from a q-dimensional Gaussian distribution with mean vector zero and covariance matrix $\mathbf{Q}$. We used the Toeplitz structure $\varrho^{|a-b|}$, $a, b \in \{1, \ldots, q\}$ with $\varrho = 0.9$ to generate the covariance matrix $\mathbf{Q}$. Finally, given $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, we generated $\dot{\mathbf{Y}}$ from a p-dimensional Gaussian distribution with mean $\mathbf{\Gamma}^\top \dot{\mathbf{x}}$ and covariance matrix $\mathbf{\Theta}^{-1}$.

To conduct the simulation, we set $n = 25000$ and 50000 and the number of machines to $K = 5, 10$ and 20. To show the performance of the distributed estimator in the unbalanced setting, we considered the same splitting procedure as in Chapters 2 and 3. Suppose that among all available machines, two of them are powerful. The first one is the most powerful one and $(55 - K)\%$ of the dataset

is distributed on this machine. The second one is less powerful than the first one and $(60-K)\%$ of the remaining dataset is distributed on this one. The remaining dataset is distributed roughly equally on the remaining machines. By considering this splitting procedure and setting the number of responses to $p = 1100$ and the number of predictors to $q = 550$, for some sub-samples we have high-dimensionality. For example, when $n = 25000$ and $K = 10$, we have $p > n_k$, $\forall k = 3, \ldots, 10$ and when $K = 20$, we have $p, q > n_k$, $\forall k = 3, \ldots, 20$. Moreover, when $n = 50000$ and $K = 20$, we have $p > n_k$, $\forall k = 3, \ldots, 20$. To compare the performance of the distributed estimators, we considered two types of estimators: debiased (which are non-sparse) and sparse. The debiased estimators consist of:

1) (Full) A debiased estimator based on the full non-distributed data.

2) (sAvg) An estimator based on splitting the data and averaging directly the debiased estimators from each machine.

3) (wAvg) An estimator based on splitting the data and taking the weighted average of the debiased estimators from each machine, where the weight for the k-th sub-sample is set to (n_k/n).

4) (Top1) The estimator produced by the most powerful machine which takes $(55-K)\%$ of dataset. Since estimation on each machine is consistent and asymptotically normal, investigating the performance on the first machine which takes most of the dataset, is relevant.

The sparse competitors, which are shown respectively by SFull, SsAvg, SwAvg, STop1, are obtained in the same way as estimators in 1)-4) but without the debiasing step. A comparison with the full estimator reveals how much the performance deteriorates due to splitting the data, while a comparison with the simple and sample size weighted average estimators has the purpose of evaluating if indeed the proposed owAvg estimator is better equipped to tackle unbalanced settings due to a more appropriate weighting. A comparison with the Top1 estimator has the purpose to evaluate if the remaining $(K-1)$ machines which account for $(45+K)\%$ of the original data are still able to produce informative estimates even though they receive low amounts of data. In this study, the tuning parameters in (4.3.1), (4.4.1) and (4.9.1), which are needed to obtain $\mathbf{M}_k$, see Appendix 4.9.1, are set to $\rho_k = \rho = \sqrt{\log(pq)/n}$, $\lambda_k = \lambda = \sqrt{\log(p)/n}$ and $\tilde{\rho}_{j,k} = \tilde{\rho} = \sqrt{\log(q)/n}$ for all simulation runs. All simulation results are calculated as averages over $R = 500$ different repetitions.

To compare the performance of the estimators, we used the Frobenius norm between the estimated matrix for each competitor and the true matrix from the data generating process. The results for the Frobenius norm and their standard deviation (between parentheses) for $n = 50000$ are presented in Table 4.1 for each competitor, separately on the active and the non-active sets. Note that the active sets on $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ are indexed by S_1 and S_2, respectively. The results for $n = 25000$ are similar and are presented in Table 4.4 in Appendix 4.11.

From Table 4.1, it is observed that the performance of the proposed debiased owAvg estimator is similar to that of the non-distributed, full estimator in terms of Frobenius norm. With increasing K from 5 to 20, the norm of the distributed estimator stays relatively constant. This suggests that by splitting observations in combination with the proposed aggregation, one might not lose much information. On the other hand, wAvg is quite sensitive to the number of machines and with increasing K, its norm performance deteriorates. Especially when $K = 20$, the difference relative to the full one is large on the non-active set. The worst estimator between debiased ones is sAvg which imposes the same weights on all sub-samples and it is observed that its norm increases substantially, which suggests that it is highly sensitive to K, as opposed to the distributed owAvg estimator. Furthermore, the Frobenius norm of Top1 is much larger than that of the centralized full estimator, which implies that by considering just the first machine with the largest amount of data and disregarding the remaining machines one loses information as this strategy does not provide an accurate estimate. Moreover, as it is expected, the performance of the sparse estimators is much better than the performance of the debiased estimators on the non-active set only as they shrink most of the elements to zero. However, their errors are much larger than the errors of the debiased estimators on the active set as they do not correct for the bias.

Due to the asymptotic distribution of the proposed estimators, investigating their inferential properties is also of interest. However, since there is no distributional result available for the sparse estimators, it is not possible to perform inference using these estimators. Figure 4.1 shows the normalized distribution of the proposed debiased optimally weighted estimator and the full non-distributed estimator for both $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively from top to bottom. As an illustrative example, these figures are reported for $(a, b) = (1, 3)$, others are available from the authors, but are similar. The light gray histograms correspond to the normalized distribution of the full estimators which are obtained from (4.3.4) and (4.4.6), by setting $K = 1$, for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively. The dark gray histograms correspond to the asymptotic distributions of (4.5.8) and (4.5.9), respectively. The presented histograms confirm the asymptotic normal distribution of the proposed estimators and its similarity to the full one.

Using the asymptotic distribution of the estimators, the coverage probabilities and the length of the confidence intervals are presented in Table 4.2 at significance level $\alpha = .05$. Here, we explain the procedure for computing the empirical coverage probability for the elements of $\mathbf{\Theta}$. The same procedure is applied to compute the empirical coverage probability and the length of confidence intervals for the elements of $\mathbf{\Gamma}$. Similarly to Jankova and van de Geer (2015) and Chapter 2, the empirical probability that the true parameter $\mathbf{\Theta}_{ab}$ is included in the confidence interval is defined as $\hat{\mathbb{P}}_{ab} = \#\{\mathbf{\Theta}_{ab} \in \mathrm{CI}_{ab,r}\}/R$, where R is the number of repetitions in the simulation, $\mathrm{CI}_{ab,r}$ is the estimated confidence interval for $\mathbf{\Theta}_{ab}$ at the r-th repetition, and $\#$ denotes the number of times for which the true parameter $\mathbf{\Theta}_{ab}$ belongs to the confidence interval. After obtaining $\hat{\mathbb{P}}_{ab}$ for

Table 4.1: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets, for the proposed estimators and different competitors, when $n = 50000$.

			Active set				Non-active set			
					K				K	
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Debiased	Full	0.73				4.75			
			(.01)				(.01)			
		owAvg		0.74	0.78	0.89		4.86	5.13	5.59
				(.00)	(.01)	(.01)		(.00)	(.01)	(.01)
		Top1		1.00	1.05	1.19		6.60	6.95	7.85
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		sAvg		1.07	1.81	2.94		7.00	10.44	12.96
				(.01)	(.02)	(.02)		(.01)	(.05)	(.03)
		wAvg		0.75	0.84	1.32		4.89	5.42	6.74
				(.00)	(.01)	(.01)		(.00)	(.01)	(.01)
	Sparse	SFull	2.03				.38			
			(.01)				(.00)			
		STop1		2.13	2.15	2.20		1.02	1.19	1.69
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SsAvg		2.02	1.82	1.49		3.12	5.06	6.06
				(.01)	(.01)	(.01)		(.00)	(.02)	(.01)
		SwAvg		1.99	1.90	1.68		1.44	1.95	2.78
				(.01)	(.01)	(.01)		(.00)	(.00)	(.00)
$\boldsymbol{\Gamma}$	Debiased	Full	1.09				10.85			
			(.01)				(.01)			
		owAvg		1.11	1.15	1.23		11.10	11.44	12.24
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		Top1		1.55	1.63	1.86		15.43	16.29	18.54
				(.01)	(.01)	(.02)		(.02)	(.02)	(.02)
		sAvg		1.57	2.02	2.23		15.67	20.15	22.23
				(.01)	(.02)	(.02)		(.02)	(.03)	(.03)
		wAvg		1.11	1.16	1.29		11.12	11.59	12.89
				(.01)	(.01)	(.01)		(.01)	(.01)	(.02)
	Sparse	SFull	4.74				1.93			
			(.00)				(.00)			
		STop1		4.74	4.74	4.75		1.99	2.00	2.04
				(.00)	(.00)	(.00)		(.00)	(.00)	(.00)
		SsAvg		4.75	4.36	3.97		2.04	2.27	2.65
				(.00)	(.13)	(.03)		(.00)	(.05)	(.02)
		SwAvg		4.74	4.61	4.40		1.93	1.90	1.93
				(.00)	(.04)	(.01)		(.00)	(.01)	(.00)

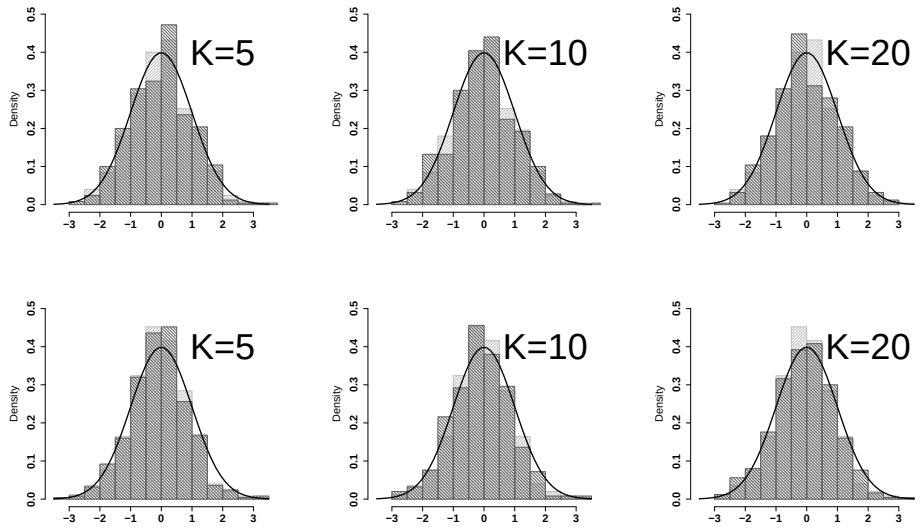


Figure 4.1: Histogram of the normalized debiased full (light gray bars) and distributed (dark gray bars), respectively, when $n = 50000$, $(a, b) = (1, 3)$ and $K = 5$, 10 or 20. From top to bottom, the figures present the asymptotic distributions for the estimation of $\mathbf{\Gamma}$ (top) and $\mathbf{\Theta}$ (bottom).

all $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, the average coverage probability on the active set S_1 is obtained as $\text{Avg.Cov}_{S_1} = (1/s_1) \sum_{(a,b)\in S_1} \hat{\mathbb{P}}_{ab}$, where s_1 is the number of active components. Similar computations have been implemented for obtaining the estimated coverage probability over the non-active set S_1^c.

From Table 4.2, it is observed that the performance of the owAvg estimator is close to the full one on both the active and non-active sets, as the coverage probabilities are close to the nominal level of 95%. Moreover, the average lengths are relatively low and are stable with increasing K. The results for Top1 are also close to the nominal level, but its length is slightly larger than that of the full and of the distributed estimator. However, in Table 4.1, we observed that its Frobenius norm performance is far away from the norm of the full estimator making it a less interesting alternative. The coverage probability of sAvg is low in some cases, especially in estimating $\mathbf{\Theta}$ on the active set. The length of its confidence interval is also relatively large, especially when estimating $\mathbf{\Gamma}$. The performance of wAvg is generally better than that of sAvg but over-coverage is observed. More than that, the length of its confidence interval is not stable and it increases with increasing K. The results for $n = 25000$ are similar and they are presented in Table 4.5 in

Table 4.2: Average coverage probability and average length of the confidence intervals over 500 repetitions for the proposed estimators and different competitors, when $n = 50000$.

			Avg.Cov				Avg.Len			
			K				K			
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Active set	Full	.93				.02			
		owAvg		.93	.93	.92		.02	.02	.02
		Top1		.94	.94	.94		.02	.03	.02
		sAvg		.92	.85	.71		.02	.03	.04
		wAvg		.94	.96	.91		.02	.02	.03
	Non-active set	Full	.93				.02			
		owAvg		.95	.95	.95		.02	.02	.02
		Top1		.95	.95	.95		.02	.03	.05
		sAvg		.94	.92	.92		.02	.03	.04
		wAvg		.95	.97	.98		.02	.02	.03
$\boldsymbol{\Gamma}$	Active set	Full	.95				.06			
		owAvg		.95	.95	.95		.06	.06	.06
		Top1		.95	.95	.95		.09	.08	.09
		sAvg		.95	.94	.93		.09	.10	.10
		wAvg		.96	.96	.97		.06	.06	.07
	Non-active set	Full	.96				.08			
		owAvg		.95	.95	.95		.06	.06	.06
		Top1		.95	.95	.95		.09	.08	.09
		sAvg		.95	.94	.93		.09	.10	.10
		wAvg		.96	.96	.97		.06	.06	.07

Appendix 4.11.

Another quantity which is important to keep track of, is the running time. In this chapter, it is considered as the maximum running time among all parallel jobs plus the time to combine the results. These results are shown in Figure 4.2 for the debiased estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ for different sample sizes from $n = 50000$ to 200000 and $K = 5, 10, 20$. Not surprisingly, the running times of the sparse estimators were less than the running times of the debiased ones as they do not have the debiasing step in the estimation procedure, and they are not reported in this figure. It is observed from Figure 4.2 that for any fixed sample size, the computation time of the distributed estimators is less than that of the full one and as expected it decreases with increasing K. This running time is quite close for all owAvg, wAvg, sAvg and Top1 estimators. This behavior is the same for both estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ and shows, as expected, the efficiency of the proposed

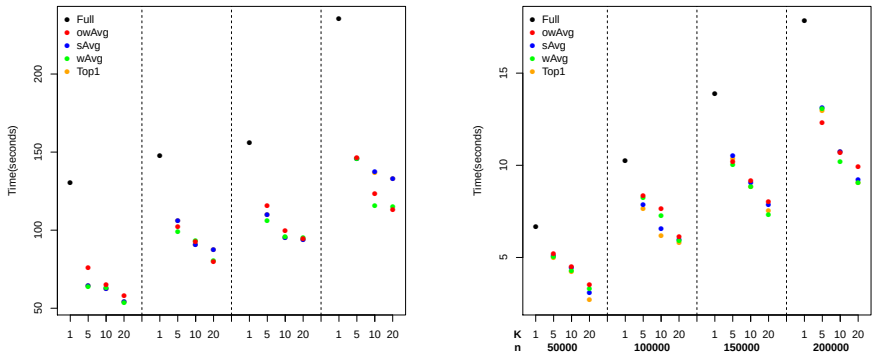


Figure 4.2: Running time in seconds for the full and proposed estimators in estimating $\mathbf{\Gamma}$ (left) and $\mathbf{\Theta}$ (right), when $p = 1100$ and $q = 550$. The regularization parameters for the distributed estimators are considered as $\lambda_k = \sqrt{\log(p)/n_k}$, $\rho_k = \sqrt{\log(pq)/n_k}$ and $\tilde{\rho}_k = \sqrt{\log(q)/n_k}$.

estimators in terms of computation time.

4.7 Real data example

To explore the performance of the proposed methodology, we used the Pan-Cancer dataset from The Cancer Genome Atlas (TCGA) project (available at https://xenabrowser.net/datapages/) that has been used as well in Chapter 3. Although the estimation of the graph structure of genes can be effective in identifying the associated genetic variants, external covariates such as single nucleotide polymorphisms may affect their structure. As such, we applied the proposed conditional multivariate regression to regress 743 microRNA mature strand expressions on 27147 tumor gene-level copy numbers and to model how the gene expressions regulate the microRNA expressions. Estimation of the underlying graph among the microRNAs is also of interest. We further selected 1164 tumor gene expressions with variance greater than 0.3, and the final dataset consists of $n = 9986$ subjects having both $q = 1164$ covariates and $p = 743$ responses.

To compare the performance of the proposed method, we split the dataset on $K = 3, 5$ and 10 different machines with the same splitting setting as in the simulation study in Section 4.6. Tuning parameters are also fixed as in the simulation study with $\alpha = 5\%$ and a Bonferroni correction is applied for multiple testing. The results are presented in Table 4.3 and it is observed that when estimating $\mathbf{\Gamma}$, the

Table 4.3: Significant and common pairs between four competitors and the estimator using the full dataset. For all methods a Bonferroni correction is applied.

		Significant pairs				Common pairs with Full		
		K				K		
		1	3	5	10	3	5	10
$\mathbf{\Gamma}$	Full	964						
	owAvg		897	832	828	797	756	685
	Top1		300	290	295	255	241	186
	sAvg		817	466	234	735	416	217
	wAvg		888	948	1334	812	761	778
$\mathbf{\Theta}$	Full	6564						
	owAvg		6168	5917	5535	5888	5675	5299
	Top1		3095	2966	2655	3060	2933	2632
	sAvg		5234	2923	1914	5099	2896	1902
	wAvg		6120	5463	4169	5854	5331	4130

owAvg and wAvg estimators identify more common non-zero coefficients with the full estimator, while sAvg and Top1 are further away from the full estimator. When $K = 3$, sAvg is close to the full one, but with increasing K, it grows further apart, such that when $K = 10$, it identified only 234 coefficients of which 217 of them are common with the full estimator. The same holds for Top1 which identified much less coefficients compared to the full, owAvg and wAvg estimators. We conclude that there are more common edges between the graphs estimated by the distributed owAvg method and the full one, while the least similar competitors are Top1 and sAvg. With increasing K, the number of common edges tends to decrease for all competitors due to the loss of information by splitting data on more machines.

As it is mentioned in Wang (2015), the tuple hsa-miR-136, hsa-miR-376a and hsa-miR-377 are the most important genes in identifying GBM cancer. Top1 could not identify any edges between these genes, while sAvg could identify an edge between hsa-miR-136 and hsa-miR-377, when $K = 3$. With the full, distributed owAvg and wAvg estimators, we could successfully identify an edge between hsa-miR-136 and hsa-miR-377 and another one between hsa-miR-376a and hsa-miR-377. However, wAvg missed the couple hsa-miR-376a and hsa-miR-377 when $K = 10$. The sparse graphical Lasso estimator is also considered as a competitor and with this method we identified 37376 edges suggesting that many estimated edges might in fact be false positive edges.

Afterwards, to investigate the performance of the estimators for completely independent variables, we permuted randomly sample data for each variable, thus breaking up the correlation structure in order to construct a dataset with mutually independent variables. As such, we expect that there are no selected variables in the estimated coefficient matrix and zero off-diagonal elements in the estimated precision matrix. By fitting separately the proposed owAvg estimator and all competitors on the dataset with $K = 3, 5$ and 10, we identified no edges in the

estimated precision matrix which confirms this assertion. However, a couple of non-zero coefficients were wrongly identified in the estimated coefficient matrix. The owAvg and sAvg estimators, both identified around 10 non-zero coefficients, while wAvg identified 84, which indicates more false positive discoveries produced by this estimator.

4.8 Discussion

Splitting a dataset on multiple locations with different sizes is an inevitable method to tackle the problem of large scale datasets which cannot be read nor stored in one single location. Separate analysis at multiple locations are also of contemporary interest due to today's security and privacy concerns. The most essential step in distributed problems is choosing how to aggregate different estimators to a final one.

In this chapter, aggregated estimators are introduced for the coefficient matrix in the multivariate regression models and the precision matrix corresponding to the graph structure of the response vector. To build these estimators, first debiased estimators were provided on each machine and then, they were pooled together to create the final estimators by maximizing a pseudo log-likelihood function which is constructed using the asymptotic distribution of the debiased estimators. Two special cases of the aggregated estimators, including simple and weighted averages, were provided under the known variance assumption. Statistical guarantees and the asymptotic distribution of the aggregated estimators were investigated under sparsity conditions and a growing number of machines as a function of the sample size.

It is shown that the aggregated estimators are asymptotically as efficient as the debiased, full non-distributed estimators and confirm the asymptotic negligibility of the efficiency loss in the distributed procedure. Moreover, based on the provided convergence rates, it is deduced that the aggregated estimators exhibit a similar statistical error as the debiased full estimator as long as the number of machine is not too large. In the estimation of the coefficient matrix, it is shown that as the number of machines grows at the rate of $K = O(n^{1/4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\}))$, the final estimator is consistent and asymptotically normal. This growth rate adjusts to $K = O(n^{1/3}/(d_1 \log(p)))$, in estimating the precision matrix. Statistical inference was also proposed based on the asymptotic normal distribution of the aggregated estimators. These distributions are valid as long as the number of machines grows at the mentioned rates.

The finite sample performance of the proposed estimators was evaluated with a simulation study where it was observed that the estimators produced competitive results relative to the non-distributed estimators that use the entire data. Since we perform estimation by distributing the computational load across multiple machines, not surprisingly the computational time comparison favors our novel estimator. Moreover, this estimator performed substantially better than the simple

average-based estimator in terms of accuracy. It was also observed that the coverage probabilities of the distributed estimators are close to those of the non-distributed estimators. This points to the fact that in practice, performing distributed estimation across multiple machines in our unbalanced framework induces a minimal loss in performance relative to models using all the data in a centralized location.

4.9 Appendix A: Proofs of the main theorems and lemmas

4.9.1 Proof of Theorem 4.3.1

Before starting the proof of Theorem 4.3.1, we provide a short description of the nodewise Lasso method from van de Geer et al. (2014) using the k-th sub-sample, which is needed later in the proof.

For every $j \in \{1, \ldots, q\}$, use Lasso for the regression problem $\mathbf{X}_{j,k}$ against $\mathbf{X}_{-j,k}$, where $\mathbf{X}_{j,k}$ and $\mathbf{X}_{-j,k}$ are, the j-th column and the $n_k \times (q-1)$ dimensional sub-matrix of $\mathbf{X}_k$ obtained by removing the j-th column, respectively. Formally,

$$\hat{\boldsymbol{\eta}}_{j,k} = \arg \min_{\boldsymbol{\eta} \in \mathbb{R}^{q-1}} \left\{ \frac{1}{2n_k} \|\mathbf{X}_{j,k} - \mathbf{X}_{-j,k}\boldsymbol{\eta}\|_2^2 + \tilde{\rho}_{j,k}\|\boldsymbol{\eta}\|_1 \right\}, \tag{4.9.1}$$

with components $\hat{\boldsymbol{\eta}}_{j,k} = \{\hat{\eta}_{(j,j'),k}; j' = 1, \ldots, q, j' \neq j\}$, where $\hat{\eta}_{(j,j'),k}$ is the estimated regression coefficient associated to the j'-th column of $\mathbf{X}_k$ when column j is the response vector. Define

$$\hat{\boldsymbol{\Psi}}_k = \begin{pmatrix} 1 & -\hat{\eta}_{(1,2),k} & \cdots & -\hat{\eta}_{(1,q),k} \\ -\hat{\eta}_{(2,1),k} & 1 & \cdots & -\hat{\eta}_{(2,q),k} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\eta}_{(q,1),k} & -\hat{\eta}_{(q,2),k} & \cdots & 1 \end{pmatrix}$$

and write

$$\hat{\boldsymbol{\Upsilon}}_k^2 = \text{diag}(\hat{\tau}_{1,k}^2, \ldots, \hat{\tau}_{q,k}^2), \quad \hat{\tau}_{j,k}^2 = \|\mathbf{X}_{j,k} - \mathbf{X}_{-j,k}\hat{\boldsymbol{\eta}}_{j,k}\|_2^2/n_k + \tilde{\rho}_{j,k}\|\hat{\boldsymbol{\eta}}_{j,k}\|_1,$$

and finally, set

$$\mathbf{M}_k = [\hat{\boldsymbol{\Upsilon}}_k^2]^{-1}\hat{\boldsymbol{\Psi}}_k.$$

Now to prove Gaussianity in Theorem 4.3.1, we mention that $\boldsymbol{\xi}_k$ is an $n_k \times p$ matrix with independent rows and distributed as $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. As such, the random vector $\text{vec}(\boldsymbol{\xi}_k)$ is distributed as a pn_k-dimensional Gaussian vector with mean zero and covariance matrix $\boldsymbol{\Sigma} \otimes \mathbf{I}_{n_k}$, which is a $pn_k \times pn_k$ matrix. Due to the properties of the $\text{vec}(\cdot)$ function,

$$\mathbf{T}_k := \text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\boldsymbol{\xi}_k/\sqrt{n_k}) = (1/\sqrt{n_k})(\mathbf{I}_p \otimes (\mathbf{M}_k\mathbf{X}_k^\top))\text{vec}(\boldsymbol{\xi}_k).$$

As such, given $\mathbf{X}_k$, the random vector $\mathbf{T}_k$ is a linear transformation of $\text{vec}(\boldsymbol{\xi}_k)$ and it is thus a pq-dimensional Gaussian vector with mean vector zero and covariance matrix

$$\begin{aligned}\mathbb{V}\text{ar}(\mathbf{T}_k|\mathbf{X}_k) &= (1/n_k)(\mathbf{I}_p \otimes (\mathbf{M}_k\mathbf{X}_k^\top))(\boldsymbol{\Sigma} \otimes \mathbf{I}_{n_k})(\mathbf{I}_p \otimes (\mathbf{X}_k\mathbf{M}_k^\top)) \\ &= \boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top).\end{aligned}$$

To show negligibility of the remainder term $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ in (4.3.3), we have

$$\begin{aligned}\|\mathbf{R}_{k,\boldsymbol{\Gamma}}\|_\infty &\leq \sqrt{n_k}\|\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k\|_\infty \|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_\infty \\ &\leq \sqrt{n_k}\|\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k\|_\infty \|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_1. \end{aligned} \tag{4.9.2}$$

Using the KKT conditions, van de Geer et al. (2014) showed that $\|\mathbf{C}_k\mathbf{M}_{j,k}^\top - \mathbf{e}_j\|_\infty \leq \tilde{\rho}_{j,k}/\hat{\tau}_{j,k}^2$, where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$ and $\mathbf{e}_j$ is the j-th unit column vector with 1 at the j-th position and zero everywhere else. Under the assumptions (A1) and (A2) and considering the maximum row sparsity $d_2 = o(\sqrt{n_\dagger}/\log(q))$ in Lemma 5.3 of van de Geer et al. (2014), we have

$$\max_{j\in\{1,\ldots,q\}} 1/\hat{\tau}_{j,k}^2 = O_p(1).$$

Therefore, by choosing uniformly $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$ for each $j = 1, \ldots, q$, we get

$$\|\mathbf{C}_k\mathbf{M}_k^\top - \mathbf{I}_q\|_\infty = \max_j \|\mathbf{C}_k\mathbf{M}_{j,k} - \mathbf{e}_j\|_\infty = O_p(\sqrt{\log(q)/n_k}). \tag{4.9.3}$$

Substituting (4.9.3) and (4.10.14) from Lemma 4.10.7 (see Appendix 4.10) in (4.9.2), we get

$$\|\mathbf{R}_{k,\boldsymbol{\Gamma}}\|_\infty = O_p(s_2\sqrt{\log(q)\log(pq)/n_k}).$$

Finally, by considering $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$, the required result follows. □

4.9.2 Proof of Theorem 4.3.2

Using (4.3.3), we have that,

$$\|\hat{\boldsymbol{\Gamma}}_k^d - \boldsymbol{\Gamma}\|_\infty \leq \|\text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\boldsymbol{\xi}_k)\|_\infty/n_k + \|\mathbf{R}_{k,\boldsymbol{\Gamma}}\|_\infty/\sqrt{n_k}. \tag{4.9.4}$$

Due to the properties of the $\text{vec}(\cdot)$ function,

$$\|\text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\boldsymbol{\xi}_k)\|_\infty/n_k \leq \|\mathbf{I}_p \otimes \mathbf{M}_k\|_\infty \|\text{vec}(\mathbf{X}_k^\top\boldsymbol{\xi}_k)\|_\infty/n_k = O_p(\sqrt{d_2\log(pq)/n_k}), \tag{4.9.5}$$

where the last equality is obtained by (i) working on the event $\mathcal{F}_k(n_k, p, q)$ with $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and (ii) by the same argument as in the proof of Lemma 5.4 from van de Geer et al. (2014), as

$$|||\mathbf{I}_p \otimes \mathbf{M}_k|||_\infty = |||\mathbf{M}_k|||_\infty = \max_{j\in\{1,\ldots,q\}} \|\mathbf{M}_{j,k}\|_1 = O_p(\sqrt{d_2}),$$

where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$. Under assumptions (A1) and (A2) and using Theorem 4.3.1, by substituting (4.3.4) and (4.9.5) in (4.9.4), the result in (4.3.5) follows directly. □

4.9.3 Proof of Lemma 4.4.1

Before starting the proof of this Lemma, we provide some technical assumptions on the sample covariance matrix $\mathbf{C}_k$, which are needed later in the proof. For more information on these assumptions, the reader can refer to Yin and Li (2013).

(C1) There exists an $\alpha_2 \in (0, 1]$, such that

$$\sup_{b\in\{1,\ldots,p\}} |||(\mathbf{C}_k)_{S_2^c(b)S_2(b)}[(\mathbf{C}_k)_{S_2(b)S_2(b)}]^{-1}|||_\infty \leq 1 - \alpha_2,$$

where $(\mathbf{C}_k)_{S_2^c(b)S_2(b)}$ is a sub-matrix of $\mathbf{C}_k$ whose rows and columns are indexed by the elements of $S_2^c(b)$ and $S_2(b)$, respectively, where $S_2^c(b)$ is the complement set of $S_2(b)$, defined in Section 4.2. As it is mentioned in Yin and Li (2013), this condition is the matrix version of the irrepresentability condition used in the ℓ_1 penalized regression setting of Zhao and Yu (2006).

(C2) There exists a constant $C_{\max}$, such that the largest eigenvalue

$$\Lambda_{\max}\left([(\mathbf{C}_k \otimes \mathbf{I}_p)_{S_2S_2}]^{-1}(\mathbf{C}_k \otimes \mathbf{\Sigma})_{S_2S_2}[(\mathbf{C}_k \otimes \mathbf{I}_p)_{S_2S_2}]^{-1}\right) \leq C_{\max}.$$

(C3) For all $n_k > 0$, the largest eigenvalue of $\mathbf{C}_k$ has a common upper bound Λ_3, that is $\Lambda_{\max}(\mathbf{C}_k) \leq \Lambda_3$.

Now to prove Lemma 4.4.1, using (4.4.4), we have

$$\begin{aligned}
\|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty &= \sqrt{n_k}\|-(\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p)(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}) - (\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{W}_k'\mathbf{\Theta} - \mathbf{\Theta}\mathbf{W}_k''\mathbf{\Theta}\|_\infty \\
&\leq \sqrt{n_k}\|\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p\|_\infty |||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}|||_\infty + \sqrt{n_k}|||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}|||_\infty \|\mathbf{W}_k'\mathbf{\Theta}\|_\infty \\
&\quad + \sqrt{n_k}\|\mathbf{\Theta}\mathbf{W}_k''\mathbf{\Theta}\|_\infty.
\end{aligned} \tag{4.9.6}$$

To find an upper bound on $\|\hat{\boldsymbol{\Theta}}_k\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \mathbf{I}_p\|_\infty$, we write

$$\begin{aligned}
\|\hat{\boldsymbol{\Theta}}_k\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \mathbf{I}_p\|_\infty &= \|\hat{\boldsymbol{\Theta}}_k\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \hat{\boldsymbol{\Theta}}_k\boldsymbol{\Sigma} + \hat{\boldsymbol{\Theta}}_k\boldsymbol{\Sigma} - \mathbf{I}_p\|_\infty \\
&= \|\hat{\boldsymbol{\Theta}}_k(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}) + \hat{\boldsymbol{\Theta}}_k\boldsymbol{\Sigma} - \boldsymbol{\Theta}\boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma} - \mathbf{I}_p\|_\infty \\
&= \|\hat{\boldsymbol{\Theta}}_k(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}) + (\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta})\boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma} \\
&\quad - \boldsymbol{\Theta}\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} + \boldsymbol{\Theta}\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \mathbf{I}_p\|_\infty \\
&= \|\hat{\boldsymbol{\Theta}}_k(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}) + (\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta})\boldsymbol{\Sigma} - \boldsymbol{\Theta}(\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}) \\
&\quad + \boldsymbol{\Theta}\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Theta}\boldsymbol{\Sigma}\|_\infty \\
&\leq |||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty\|\mathbf{W}'_k\|_\infty + \|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}\|_\infty|||\boldsymbol{\Sigma}|||_\infty + \|\boldsymbol{\Theta}\mathbf{W}'_k\|_\infty.
\end{aligned} \tag{4.9.7}$$

Substituting (4.9.7) in (4.9.6), we have

$$\|\mathbf{R}_{k,\boldsymbol{\Theta}}\|_\infty \leq \sqrt{n_k}\{2\|\mathbf{R}^1_{k,\boldsymbol{\Theta}}\|_\infty + \|\mathbf{R}^2_{k,\boldsymbol{\Theta}}\|_\infty + \|\mathbf{R}^3_{k,\boldsymbol{\Theta}}\|_\infty + \|\mathbf{R}^4_{k,\boldsymbol{\Theta}}\|_\infty\}, \tag{4.9.8}$$

where

$$\begin{aligned}
\|\mathbf{R}^1_{k,\boldsymbol{\Theta}}\|_\infty &= \|\boldsymbol{\Theta}\mathbf{W}'_k\|_\infty|||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty, \\
\|\mathbf{R}^2_{k,\boldsymbol{\Theta}}\|_\infty &= |||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||^2_\infty\|\mathbf{W}'_k\|_\infty, \\
\|\mathbf{R}^3_{k,\boldsymbol{\Theta}}\|_\infty &= \|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}\|_\infty|||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty\kappa_{\boldsymbol{\Sigma}}, \\
\|\mathbf{R}^4_{k,\boldsymbol{\Theta}}\|_\infty &= \|\boldsymbol{\Theta}\mathbf{W}''_k\boldsymbol{\Theta}\|_\infty.
\end{aligned}$$

Under assumption (B1) from Section 4.4 for the matrix $\boldsymbol{\Theta}$, it holds that

$$|||\boldsymbol{\Theta}|||_\infty = \max_{a\in\mathcal{V}}\|\boldsymbol{\Theta}_a\|_1 \leq \sqrt{d_1}\Lambda_{\max}(\boldsymbol{\Theta}) = O(\sqrt{d_1}), \tag{4.9.9}$$

where $\boldsymbol{\Theta}_a$ is the a-th row of $\boldsymbol{\Theta}$. Under (B2) and the additional assumptions (C1)-(C3), the conditions of result 1 from Theorem 2 of Yin and Li (2013) are fulfilled, and for the k-th sub-sample we have

$$\|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}\|_\infty = O_p\left(\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a\boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}}\}\sqrt{\frac{\log(4p^\tau)}{n_k}}\right), \tag{4.9.10}$$

where $\max_a \boldsymbol{\Sigma}_{aa}$ is the maximal diagonal element of the covariance matrix $\boldsymbol{\Sigma}$, $\tau > 2$ is a constant and γ is a common sub-Gaussian parameter for Gaussian random variables $\varepsilon^1, \ldots, \varepsilon^p$, where $\gamma = O(1)$. Moreover, based on result 2 of the aforementioned theorem, the edge set $\mathcal{E}(\hat{\boldsymbol{\Theta}}_k)$, i.e. the edge set created based on the estimated $\hat{\boldsymbol{\Theta}}_k$, is a subset of the true edge set $\mathcal{E}(\boldsymbol{\Theta})$ with high probability. As

such, $\hat{\mathbf{\Theta}}_k$ has at most d_1 non-zero entries per row, and we get

$$\begin{aligned}\|\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}\|_\infty = \max_a \sum_{b=1}^{p} |\hat{\mathbf{\Theta}}_{ab,k} - \mathbf{\Theta}_{ab}| &= \max_a \sum_{b=1}^{d_1} |\hat{\mathbf{\Theta}}_{ab,k} - \mathbf{\Theta}_{ab}| \\ &\leq \max_a \sum_{b=1}^{d_1} \max_b |\hat{\mathbf{\Theta}}_{ab,k} - \mathbf{\Theta}_{ab}| \\ &= d_1 \|\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}\|_\infty,\end{aligned}$$

where $\hat{\mathbf{\Theta}}_{ab,k}$ is the (a,b)-th element of $\hat{\mathbf{\Theta}}_k$. Thus, using (4.9.10),

$$\|\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}\|_\infty \leq O_p\bigg(d_1\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \mathbf{\Sigma}_{aa}\kappa_{\mathbf{H}}\}\sqrt{\frac{\log(4p^\tau)}{n_k}}\bigg).$$

Therefore in (4.9.8), we have

$$\|\mathbf{R}_{k,\mathbf{\Theta}}\|_\infty \leq 5\sqrt{n_k}\max\{\|\mathbf{R}^1_{k,\mathbf{\Theta}}\|_\infty, \|\mathbf{R}^2_{k,\mathbf{\Theta}}\|_\infty, \|\mathbf{R}^3_{k,\mathbf{\Theta}}\|_\infty, \|\mathbf{R}^4_{k,\mathbf{\Theta}}\|_\infty\}, \qquad (4.9.11)$$

where

$$\begin{aligned}\|\mathbf{R}^1_{k,\mathbf{\Theta}}\|_\infty &\leq \sqrt{d_1}\Lambda_{\max}(\mathbf{\Theta})\|\mathbf{W}'_k\|_\infty \\ &\quad \times O_p\bigg(d_1\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \mathbf{\Sigma}_{aa}\kappa_{\mathbf{H}}\}\sqrt{\frac{\log(4p^\tau)}{n_k}}\bigg), \\ \|\mathbf{R}^2_{k,\mathbf{\Theta}}\|_\infty &\leq \|\mathbf{W}'_k\|_\infty O_p\bigg(d_1^2\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \mathbf{\Sigma}_{aa}\kappa_{\mathbf{H}}\}^2\frac{\log(4p^\tau)}{n_k}\bigg), \\ \|\mathbf{R}^3_{k,\mathbf{\Theta}}\|_\infty &\leq \kappa_{\mathbf{\Sigma}} O_p\bigg(d_1\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \mathbf{\Sigma}_{aa}\kappa_{\mathbf{H}}\}^2\frac{\log(4p^\tau)}{n_k}\bigg).\end{aligned}$$

To obtain an upper bound on $\|\mathbf{W}'_k\|_\infty$, under assumptions (C1)-(C3), and using Lemma 2 of Yin and Li (2013), we can write

$$\|\mathbf{W}'_k\|_\infty = O_p\bigg(\sqrt{\frac{\log(4p^\tau)}{C_2 n_k}}\bigg),$$

where $C_2 = [128(1+4\gamma^2)^2\max_a^2 \mathbf{\Sigma}_{aa}]^{-1}$. Moreover,

$$\|\mathbf{R}^4_{k,\mathbf{\Theta}}\|_\infty = \|\mathbf{\Theta}\{(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k - \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k\}\mathbf{\Theta}\|_\infty.$$

Adding and subtracting $\mathbf{X}_k\mathbf{\Gamma}$ to the first term in $\mathbf{R}^4_{k,\mathbf{\Theta}}$, we get that

$$\|\mathbf{R}^4_{k,\mathbf{\Theta}}\|_\infty \leq \|\mathbf{\Theta}\|^2_\infty\bigg\{2\|(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top\mathbf{X}_k^\top\boldsymbol{\xi}_k\|_\infty/n_k + \|\mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\|_F^2/n_k\bigg\}.$$

Under assumption (B1) and by conditioning on the event $\mathcal{F}_k(n_k, p, q)$, we have

$$\|\mathbf{R}^4_{k,\boldsymbol{\Theta}}\|_\infty \leq d_1\{\rho_k\|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_1 + \|\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})\|_F^2/n_k\}.$$

Recall that S_2 and $S_2(b)$ denote the support of the coefficient matrix $\boldsymbol{\Gamma}$ and its b-th column, $b = 1, \ldots, p$, with cardinalities s_2 and $s_2(b)$, respectively. Under assumption (C1), by a similar argument as in the proof of Theorem 6.1 of Bühlmann and van de Geer (2011), one can show that for the fixed design matrix $\mathbf{X}_k$,

$$\rho_k\|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_1 + \|\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})\|_F^2/n_k \leq 4\rho_k^2 s_2/\phi^2.$$

where $\phi^2 \in (0, \infty)$ is the compatibility constant which has a similar role to μ_b from the RE condition (4.10.5) for the fixed design matrix $\mathbf{X}_k$. It is noteworthy that as it is shown in Theorem 7.2 of Bühlmann and van de Geer (2011), under the fixed design setting, the irrepresentability condition (C1) is enough to reach the compatibility condition and the restricted eigenvalue condition (4.10.5) is not needed. The reader can refer to Chapter 7 of Bühlmann and van de Geer (2011) for more details. Since $\phi^2 \in (0, \infty)$, there exists $L = O(1)$, such that $1/\phi^2 \leq L$. Combining this with $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and substituting in $\mathbf{R}^4_{k,\boldsymbol{\Theta}}$, we get

$$\|\mathbf{R}^4_{k,\boldsymbol{\Theta}}\|_\infty = O_p(d_1 s_2 \log(pq)/n_k).$$

Substituting the obtained bounds in (4.9.11) and considering $\kappa_{\boldsymbol{\Sigma}} = O(1)$, $\kappa_{\mathbf{H}} = O(1)$, $1/\alpha_1 = O(1)$, and $\gamma = O(1)$, we get

$$\|\mathbf{R}_{k,\boldsymbol{\Theta}}\|_\infty = O_p\Big(\sqrt{n_k}\max\{d_1^{3/2}\log(p)/n_k, d_1^2(\log(p)/n_k)^{3/2}, d_1\log(p)/n_k, d_1 s_2\log(pq)/n_k\}\Big),$$

and under the sparsity assumptions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$ the result $\|\mathbf{R}_{k,\boldsymbol{\Theta}}\|_\infty = o_p(1)$ follows. □

4.9.4 Proof of Theorem 4.4.2

The proof of this theorem is an extension of the proof of Theorem 1 from Jankova and van de Geer (2015) by considering a non-zero mean structure. Under the assumptions of Lemma 4.4.1 from the main text, it is shown that $\|\mathbf{R}_{k,\boldsymbol{\Theta}}\|_\infty = o_p(1)$. As such, using (4.4.3), we have

$$\begin{aligned}\sqrt{n_k}(\hat{\boldsymbol{\Theta}}^d_{ab,k} - \boldsymbol{\Theta}_{ab}) &= -\sqrt{n_k}\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab} + o_p(1)\\ &= -\frac{1}{\sqrt{n_k}}\sum_{l=1}^{n_k}(\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b\\ &\quad - \boldsymbol{\Theta}_{ab}) + o_p(1), \qquad (4.9.12)\end{aligned}$$

where $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$ and $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ are the l-th row of the sub-matrices $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively, and $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ are p-dimensional vectors coming from the a-th row and the b-th row of $\boldsymbol{\Theta}$. The second equality in (4.9.12) follows directly since $\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta} = \boldsymbol{\Theta}\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{\Gamma}}\boldsymbol{\Theta} - \boldsymbol{\Theta}$. To show the normality of the term in (4.9.12), define $\mathbf{Z}_{ab,l,k} := \boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$, $l = 1, \ldots, n_k$. Then,

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_{ab,l,k}) &= \boldsymbol{\Theta}_a^\top\mathbb{E}\{(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\}\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab} \\ &= \mathbf{e}_a^\top\boldsymbol{\Theta}\mathbf{e}_b - \boldsymbol{\Theta}_{ab} = 0,\end{aligned}$$

where $\mathbf{e}_b$ is a p-dimensional unit column vector of zeros with one at position b and similarly for $\mathbf{e}_a$. On the other hand, since $\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k}$ is a Gaussian random vector, the variance

$$\sigma_{ab}^2 = \mathbb{V}\mathrm{ar}(\mathbf{Z}_{ab,l,k}) = \mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b)$$

is finite. Denote by $\mathcal{Z}_{n_k} = \sum_{l=1}^{n_k}\mathbf{Z}_{ab,l,k}$. Then, $z_{n_k}^2 := \mathbb{V}\mathrm{ar}(\mathcal{Z}_{n_k}) = n\sigma_{ab}^2$. Dividing (4.9.12) by $\sigma_{ab} > 0$, we get

$$\sqrt{n_k}(\hat{\boldsymbol{\Theta}}_{ab,k}^d - \boldsymbol{\Theta}_{ab})/\sigma_{ab} = \mathcal{Z}_{n_k}/z_{n_k} + o_p(1)/\sigma_{ab}.$$

It is enough to show that $\mathcal{Z}_{n_k}/z_{n_k} \xrightarrow{d} \mathcal{N}(0,1)$, where $\xrightarrow{d}$ denotes convergence in distribution. By substituting $\mathbf{Z}_{ab,l,k}$ in the proof of Theorem 1 from Jankova and van de Geer (2015), with the same argument, the normality of $\mathcal{Z}_{n_k}/z_{n_k}$ follows. The reader can refer to Jankova and van de Geer (2015) for more details.

To show $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$, under the multivariate Gaussian distribution of $\dot{\boldsymbol{\varepsilon}}_{l,k}$, which is the l-th row of $\boldsymbol{\xi}_k$, we have $\dot{\boldsymbol{\varepsilon}}_{l,k} = \dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, and as such $\boldsymbol{\Theta}(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Theta})$. Using a similar argument as in the proof of Lemma 2 from Jankova and van de Geer (2015), it holds that $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$ and $1/\sigma_{ab} = O(1)$. □

4.9.5 Proof of Theorem 4.5.1

a) The proof of this theorem relies on Lemma 4.10.9, which shows the negligibility of the remainder term $\mathbf{R}_{a,\boldsymbol{\Gamma}}$ as $n_\dagger$ and K grow. To show the asymptotic normality of $\mathbf{W}_{a,\boldsymbol{\Gamma}}$, define

$$\zeta_a := \sum_{k=1}^K \sqrt{\frac{n_k[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

where $\mathbf{T}_{a,k}$ is the a-th element of $\mathbf{T}_k$ from (4.3.3). As $\sqrt{\frac{\sum_{k=1}^K n_k/[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^K n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \xrightarrow{p} 1$, see the proof of Lemma 4.10.9, using Slutsky's theorem, it is enough to show that ζ_a converges in distribution to $\mathcal{N}(0,1)$. Defining $\zeta_a' := \sum_{k=1}^K\sqrt{\frac{n_k}{n}} \times$

$\frac{\mathbf{T}_{a,k}}{\sqrt{[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$, we show in Lemma 4.10.10 that $|\zeta_a - \zeta_a'| \xrightarrow{p} 0$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, and then convergence in distribution of ζ_a follows by convergence in distribution of ζ_a'.

Now to show the asymptotic normal distribution of ζ_a', denoting by $\mathbf{Z}_{a,k} := \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$, we use the Lindeberg theorem (see for instance, Theorem 2.1 in Chapter 7 of Gut, 2005). We have

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_{a,k}) &= \mathbb{E}\bigg\{\mathbb{E}\bigg(\sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \mid \mathbf{X}_k\bigg)\bigg\} \\ &= \mathbb{E}\bigg\{\sqrt{\frac{n_k}{n}} \times \frac{1}{\sqrt{[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\mathbb{E}(\mathbf{T}_{a,k} \mid \mathbf{X}_k)\bigg\} = 0,\end{aligned}$$

where the last equality is deduced from the conditional normal distribution of $\mathbf{T}_{a,k}$ shown in Theorem 4.3.1. Moreover,

$$\sum_{k=1}^{K}\mathbb{E}(\mathbf{Z}_{a,k}^2) = \sum_{k=1}^{K}\mathbb{E}\bigg\{\frac{n_k}{n} \times \frac{1}{[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbb{E}(\mathbf{T}_{a,k}^2 \mid \mathbf{X}_k)\bigg\} = \sum_{k=1}^{K}\frac{n_k}{n} = 1,$$

where the second equality is deduced using the conditional variance of $\mathbf{T}_{a,k}$, which is equal to $[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$. Now to check the Lindeberg condition, for every $\epsilon > 0$, we can write

$$\begin{aligned}&\mathbb{E}(\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}| > \epsilon) \mid \mathbf{X}_k) \\ &\quad= \frac{n_k}{n[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbb{E}\bigg\{\mathbf{T}_{a,k}^2\mathbb{I}(|\mathbf{T}_{a,k}| > \epsilon\sqrt{\frac{n[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{n_k}}) \mid \mathbf{X}_k\bigg\},\end{aligned} \tag{4.9.13}$$

where $\mathbb{I}(\cdot)$ is the indicator function. As was done in Chapter 2, we use next the relation

$$\mathbb{E}(X^2\mathbb{I}(|X| > a)) = a^2\mathbb{P}(|X| > a) + 2\int_a^\infty u\mathbb{P}(|X| > u)\, du$$

for a positive constant a. Applying this equality to (4.9.13), we get

$$\begin{aligned}&\mathbb{E}(\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}| > \epsilon) \mid \mathbf{X}_k) \\ &\quad= \epsilon^2\mathbb{P}(|\mathbf{T}_{a,k}| > \epsilon\sqrt{\frac{n[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k) \\ &\quad+ \frac{2n_k}{n[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\int_{\epsilon\sqrt{n[\mathbf{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}/n_k}}^\infty u\mathbb{P}(|\mathbf{T}_{a,k}| > u \mid \mathbf{X}_k)\, du.\end{aligned} \tag{4.9.14}$$

Applying the concentration inequality $\mathbb{P}(|X-\mu|> t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$, $t \in \mathbb{R}$, for a normal random variable X with mean μ and variance σ^2, we get

$$\mathbb{P}\bigg(|\mathbf{T}_{a,k}|> \epsilon\sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k\bigg) \leq 2e^{-\frac{n\epsilon^2}{2n_k}}.$$

In the integral, by substituting $t := \frac{\sqrt{n_k}u}{\epsilon\sqrt{n[\boldsymbol{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$, we have

$$\begin{aligned}
&\int_{\epsilon\sqrt{n[\boldsymbol{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}/n_k}}^{\infty} u\mathbb{P}(|\mathbf{T}_{a,k}|> u \mid \mathbf{X}_k)\, du \\
&= \frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}\epsilon^2}{n_k} \int_1^\infty t\mathbb{P}\bigg(|\mathbf{T}_{a,k}|> \epsilon t\sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k\bigg)\, dt \\
&\leq \frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}\epsilon^2}{n_k} \int_1^\infty 2te^{-\frac{n\epsilon^2t^2}{2n_k}}\, dt = 2[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}e^{-\frac{n\epsilon^2}{2n_k}}.
\end{aligned}$$

As such, in (4.9.14),

$$\mathbb{E}(\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}|> \epsilon) \mid \mathbf{X}_k) \leq 2e^{-\frac{n\epsilon^2}{2n_k}}\{\epsilon^2 + 2n_k/n\}.$$

Thus, in the Lindeberg condition, we get

$$\begin{aligned}
&\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^K \mathbb{E}\bigg\{\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}|> \epsilon)\bigg\} \\
&= \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^K \mathbb{E}\bigg\{\mathbb{E}(\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}|> \epsilon) \mid \mathbf{X}_k)\bigg\} \\
&\leq \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^K 2e^{-\frac{n\epsilon^2}{2n_k}}\{\epsilon^2 + 2n_k/n\} \\
&\leq \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^K 2e^{-\frac{Kn_\dagger\epsilon^2}{2n}}\{\epsilon^2 + 2n_k/n\} \\
&= \lim_{K\to\infty} \bigg\{\bigg(\lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} 2e^{-\frac{Kn_\dagger\epsilon^2}{2n}}\bigg)\bigg(\lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^K \{\epsilon^2 + 2n_k/n\}\bigg)\bigg\} \\
&\leq \lim_{K\to\infty} (2\epsilon^2 + 4)Ke^{-Kc_\dagger\epsilon^2/2} = 0,
\end{aligned}$$

where the last inequality is deduced due to the fact that $n_k < n$, $k = 1, \ldots, K$, and the last equality follows by l'Hôpital's rule.

b) By basic algebra it can be shown that

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}(\mathrm{vec}(\tilde{\boldsymbol{\Gamma}}_{\mathrm{wAvg}})_a-\mathrm{vec}(\boldsymbol{\Gamma})_a)=\mathbf{W}'_{a,\boldsymbol{\Gamma}}+\mathbf{R}'_{a,\boldsymbol{\Gamma}},$$

where

$$\mathbf{W}'_{a,\boldsymbol{\Gamma}}=\sqrt{\frac{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\mathbf{T}_{a,k},$$

$$\mathbf{R}'_{a,\boldsymbol{\Gamma}}=\sqrt{\frac{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\mathbf{R}_{a,k,\boldsymbol{\Gamma}},$$

where $\mathbf{T}_{a,k}$ and $\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$ are respectively the a-th element of $\mathbf{T}_k$ and $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ defined in (4.3.3). Consider the random variable $\zeta_a=\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\mathbf{T}_{a,k}$. In part a), it is shown that $\zeta'_a:=\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\times\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\mathbf{T}_{a,k}$ converges in distribution to $\mathcal{N}(0,1)$. On the other hand, one can easily show that $|\zeta_a-\zeta'_a|=O_p(Kd_2\sqrt{\log(pq)\log(q)/n})$, and then the $o_p(1)$ result follows by the mentioned sparsity condition on d_2 and with $K=O(n^{1/3}/(\sqrt{\log(pq)\log(q)}\max\{s_2,d_2\}))$. As such, the asymptotic normality of ζ_a follows directly. Moreover, the $o_p(1)$ result of the remainder term $\mathbf{R}'_{a,\boldsymbol{\Gamma}}$ is reached by the same technique as in part a).

c) By basic algebra it can be shown that

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^{K}1/n_k)}}(\mathrm{vec}(\tilde{\boldsymbol{\Gamma}}_{\mathrm{sAvg}})_a-\mathrm{vec}(\boldsymbol{\Gamma})_a)$$
$$=\mathbf{W}''_{a,\boldsymbol{\Gamma}}+\mathbf{R}''_{a,\boldsymbol{\Gamma}},$$

where

$$\mathbf{W}''_{a,\boldsymbol{\Gamma}}=\frac{\sqrt{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\times\frac{1}{\sqrt{([\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^{K}1/n_k)}}$$
$$\times\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\mathbf{T}_{a,k},$$

$$\mathbf{R}''_{a,\boldsymbol{\Gamma}}=\frac{\sqrt{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\times\frac{1}{\sqrt{([\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^{K}1/n_k)}}.$$
$$\times\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\mathbf{R}_{a,k,\boldsymbol{\Gamma}}.$$

Considering $\zeta_a := \frac{1}{\sqrt{([\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^K 1/n_k)}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}}\mathbf{T}_{a,k}$ and $\zeta'_a := \sum_{k=1}^K \frac{1}{\sqrt{n_k[\boldsymbol{\Sigma}\otimes(\mathbf{M_k}\mathbf{C_k}\mathbf{M_k}^\top)]_{aa}(\sum_{k=1}^K 1/n_k)}}\mathbf{T}_{a,k}$, by similar techniques as in part a), using the Lindeberg theorem, one can show that ζ'_a is asymptotically normal with mean zero and variance 1. Moreover, $|\zeta_a - \zeta'_a| = O_p(d_2\sqrt{nK\log(pq)\log(q)}/n_\dagger)$ and $|\mathbf{R}''_{a,\boldsymbol{\Gamma}}| = O_p(s_2\sqrt{nK\log(pq)\log(q)}/n_\dagger)$, which are $o_p(1)$ under the mentioned sparsity conditions and with $\sqrt{K} = O(n_\dagger^{1/3}/(\sqrt{\log(pq)\log(q)}\max\{s_2, d_2\}))$. □

4.9.6 Proof of Theorem 4.5.2

a) The proof of this theorem relies on Lemma 4.10.11 from Section 4.10, which shows the negligibility of the remainder term $\mathbf{R}_{ab,\boldsymbol{\Theta}}$ as $n_\dagger$ and K both grow. To show the asymptotic normality, define

$$\zeta_{ab} := \sum_{k=1}^K \frac{1}{\sqrt{n}\sigma_{ab}} \times \frac{\sigma^2_{ab}}{\hat{\sigma}^2_{ab,k}} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

where $\dot{\mathbf{Y}}_{l,k}$ and $\dot{\mathbf{X}}_{l,k}$ are the l-th row, $l = 1, \ldots, n_k$, of $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively. Similarly to the proof of Theorem 4.5.1, by defining the sequence

$$\zeta'_{ab} := \sum_{k=1}^K \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

we have

$$\zeta'_{ab} = \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l})(\dot{\mathbf{Y}}_{l} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}),$$

where $\dot{\mathbf{Y}}_l \in \mathbb{R}^p$ and $\dot{\mathbf{X}}_l \in \mathbb{R}^q$ are the l-th sample, $l = 1, \ldots, n$, of $\dot{\mathbf{Y}}$ and $\dot{\mathbf{X}}$, respectively. The sequence ζ'_{ab} converges to $\mathcal{N}(0,1)$ as it is shown in Theorem 2 of Jankova and van de Geer (2015). As such, we only need to show that $|\zeta_{ab} - \zeta'_{ab}| \xrightarrow{p} 0$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, and then convergence in distribution of ζ'_{ab} yields convergence in distribution of ζ_{ab}. As it is shown in the proof of Lemma 4.10.11, the term $(1/\hat{\sigma}^2_{ab,k}) = O_p(1)$. Using Remark 4.2, we have

$$\begin{aligned}|\zeta_{ab} - \zeta'_{ab}| \leq \sum_{k=1}^K \frac{1}{\sqrt{n}\sigma_{ab}} \times O_p(\max\{\log(p)/n_k, \sqrt{d_1\log(p)/n_k}\}) \\ \times \left|\sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})\right|. \quad (4.9.15)\end{aligned}$$

Due to the definition of $\mathbf{W}_k$ in Section 4.4, we have that $|\sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab})| = n_k|\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab}|$. Under assumption (B2)

and using Lemma 2 of Lam and Fan (2009), for which the conditions are fulfilled, we have

$$n_k|\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab}| \leq n_k\|\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\|_\infty \leq n_k\|\boldsymbol{\Theta}\|_\infty^2\|\mathbf{W}_k\|_\infty \leq d_1\sqrt{n_k\log(p)}.$$

Substituting this bound in (4.9.15), we get

$$|\zeta_{ab}-\zeta'_{ab}| \leq O_p(\frac{K}{\sqrt{n}}\max\{d_1(\log(p))^{3/2}/\sqrt{n_\dagger}, d_1^{3/2}\log(p)\}),$$

and under the conditions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $K = O(n^{1/3}/(d_1\log(p)))$, the result $|\zeta_{ab}-\zeta'_{ab}| = o_p(1)$ follows.

b) Similarly to a).

c) By a similar technique as for the proof of part c) of Theorem 4.5.1, it can be shown that

$$|\mathbf{R}''_{ab,\boldsymbol{\Theta}}| = O_p\left(\sqrt{nK}\max\{d_1^{3/2}\log(p)/n_\dagger, d_1^2(\log(p)/n_\dagger)^{3/2}, d_1 s_2\log(pq)/n_\dagger\}\right),$$

which is $o_p(1)$ under the sparsity conditions $d_1^{3/2} = o(n_\dagger^{1/3}/\log(p))$ and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, and $K = O(n_\dagger^{4/3}/n)$. The proof of asymptotic normality is similar to the one from part c) of Theorem 2.4.2 from Chapter 2, and we do not repeat it here. □

4.10 Appendix B: Some technical lemmas and their proofs

Lemma 4.10.1. Consider the regression model (4.2.2) with random noise matrix $\boldsymbol{\xi}_k$ and random Gaussian design matrix $\mathbf{X}_k$. Supposing $[\mathbf{C}_k]_{aa} = 1$, $a = 1,\ldots,q$, where $[\mathbf{C}_k]_{aa}$ is the a-th diagonal element of $\mathbf{C}_k = \mathbf{X}_k^\top\mathbf{X}_k/n_k$, and choosing $\rho_0 = (\max_{1\leq b\leq p}\sqrt{\boldsymbol{\Sigma}_{bb}})\sqrt{A\log(pq)/n_k}$, where $A > 4$ is a constant, we have

$$\mathbb{P}(\mathcal{F}_k(n_k,p,q)) \geq 1 - 2/(pq)^{A/2-1}.$$

Proof. According to the definition of the elementwise ℓ_∞ norm, the event $\mathcal{F}_k$ is equivalent to

$$\mathcal{F}_k(n_k,p,q) = \left\{\max_{1\leq a\leq q}\max_{1\leq b\leq p}|\sum_{l=1}^{n_k}\varepsilon_{l,k}^b X_{l,k}^a|/n_k \leq \rho_0\right\},$$

where $X_{l,k}^a$ and $\varepsilon_{l,k}^b$ are the a-th and the b-th components of the vectors $\dot{\mathbf{X}}_{l,k}$ and $\dot{\boldsymbol{\varepsilon}}_{l,k}$ which are the l-th row of $\mathbf{X}_k$ and $\boldsymbol{\xi}_k$, respectively. Conditioning on $X_{1,k}^a,\ldots,X_{n_k,k}^a$, the random variable $\sum_{l=1}^{n_k}\varepsilon_{l,k}^b X_{l,k}^a$ is a linear transformation of $\varepsilon_{1,k}^b,\ldots,\varepsilon_{n_k,k}^b$ and it

is normally distributed with mean zero and variance $n_k\boldsymbol{\Sigma}_{bb}[\mathbf{C}_k]_{aa}$. Supposing that $[\mathbf{C}_k]_{aa}=1$, we get $V_{ab}\mid(X^a_{1,k},\ldots,X^a_{n_k,k})^\top\sim\mathcal{N}(0,1)$, where $V_{ab}:=\sum_{l=1}^{n_k}\varepsilon^b_{l,k}X^a_{l,k}/\sqrt{n_k\boldsymbol{\Sigma}_{bb}}$. With our choice of ρ_0, we have that,

$$\begin{aligned}\mathbb{P}\Big(\max_{\substack{1\le a\le q\\1\le b\le p}}|\sum_{l=1}^{n_k}\varepsilon^b_{l,k}X^a_{l,k}|/n_k\ge\rho_0\Big)&=\mathbb{P}\Big(\max_{1\le a\le q}\frac{\max_{1\le b\le p}|\sum_{l=1}^{n_k}\varepsilon^b_{l,k}X^a_{l,k}|}{n_k\max_{1\le b\le p}\sqrt{\boldsymbol{\Sigma}_{bb}}}\\&\qquad>\sqrt{A\log(pq)/n_k}\Big)\\&\le\mathbb{P}\Big(\max_{\substack{1\le a\le q\\1\le b\le p}}\Big|\frac{\sum_{l=1}^{n_k}\varepsilon^b_{l,k}X^a_{l,k}}{\sqrt{n_k\boldsymbol{\Sigma}_{bb}}}\Big|>\sqrt{A\log(pq)}\Big)\\&\le\sum_{a=1}^{q}\sum_{b=1}^{p}\mathbb{P}(|V_{ab}|\ge\sqrt{A\log(pq)}).\end{aligned}\tag{4.10.1}$$

To obtain a bound on $\mathbb{P}(|V_{ab}|\ge\sqrt{A\log(pq)})$, by conditioning on $\mathbf{X}^a:=(X^a_{1,k},\ldots,X^a_{n_k,k})^\top$, $a=1,\ldots,q$, and the Gaussian concentration inequality $\mathbb{P}(|X-\mu|>\sigma t)\le 2\exp(-t^2/2)$, $t\in\mathbb{R}$, for a normal random variable X with mean μ and variance σ^2, we get

$$\begin{aligned}\mathbb{P}(|V_{ab}|\ge\sqrt{A\log(pq)})&=\int\mathbb{P}(|V_{ab}|\ge\sqrt{A\log(pq)}\mid\mathbf{X}^a=\mathbf{x}^a)f_{\mathbf{X}^a}(\mathbf{x}^a)\,d\mathbf{x}^a\\&\le 2\exp\{-A\log(pq)/2\}\int f_{\mathbf{X}^a}(\mathbf{x}^a)\,d\mathbf{x}^a=\frac{2}{(pq)^{A/2}}.\end{aligned}\tag{4.10.2}$$

Substituting (4.10.2) in (4.10.1), the result of the lemma follows directly. □

Lemma 4.10.2. *Under the assumptions of Lemma 4.10.1, it holds that*

$$\mathbb{P}\Big(\bigcap_{k=1}^{K}\mathcal{F}_k(n_k,p,q)\Big)\ge 1-\frac{2K}{(pq)^{A/2-1}},$$

and if K grows at the rate $K=o((pq)^{A/2-1})$, then this probability tends to one with increasing n.

Proof. We have

$$\mathbb{P}\Big(\bigcap_{k=1}^{K}\mathcal{F}_k(n_k,p,q)\Big)=1-\mathbb{P}\Big((\bigcap_{k=1}^{K}\mathcal{F}_k(n_k,p,q))^c\Big)=1-\mathbb{P}\Big(\bigcup_{k=1}^{K}\mathcal{F}^c_k(n_k,p,q)\Big).\tag{4.10.3}$$

We have as well that

$$\mathbb{P}\Big(\bigcup_{k=1}^{K}\mathcal{F}^c_k(n_k,p,q)\Big)\le\sum_{k=1}^{K}\mathbb{P}(\mathcal{F}^c_k(n_k,p,q))\le\sum_{k=1}^{K}\frac{2}{(pq)^{A/2-1}}=\frac{2K}{(pq)^{A/2-1}},\tag{4.10.4}$$

where the second inequality follows using Lemma 4.10.1. By substituting (4.10.4) in (4.10.3), the result of this Lemma follows directly. □

Before providing Lemmas 4.10.6 and 4.10.7, we need some preliminary notation. The results are provided under the Restricted Eigenvalue (RE) condition. The following definition from Raskutti et al. (2010) defines the RE condition for the population covariance matrix and is essentially equivalent to the RE condition of Bickel et al. (2009). For a given subset $G \subset \{1, \ldots, q\}$ with cardinality $g = \#G$ and a constant $\delta \geq 1$, define the set

$$C(G; \delta) := \{\boldsymbol{\nu} \in \mathbb{R}^q : \quad \|\boldsymbol{\nu}_{G^c}\|_1 \leq \delta \|\boldsymbol{\nu}_G\|_1\},$$

where $\boldsymbol{\nu}_G$ is the restriction of vector $\boldsymbol{\nu}$ to G and has zeros outside the set G. The symbol G^c denotes the complement set of G.

Definition 4.10.3. (Raskutti et al., 2010) A deterministic covariance matrix $\mathbf{Q}$ satisfies the RE condition over G of order g, if there exist constants $(\delta, \mu) \in [1, \infty) \times (0, \infty)$ such that

$$\|\mathbf{Q}^{1/2}\boldsymbol{\nu}\|_2 \geq \mu \|\boldsymbol{\nu}\|_2, \quad \text{for all } \boldsymbol{\nu} \in C(G; \delta),$$

where $\mathbf{Q}^{1/2}$ is the square root matrix of $\mathbf{Q}$.

Note that Definition 4.10.3 provides a lower bound on the ℓ_2 norm of the product between $\mathbf{Q}^{1/2}$ and the vector $\boldsymbol{\nu}$. As in the multivariate regression we have a matrix of coefficients not a vector anymore, we need to provide a lower bound on the product of $\mathbf{Q}^{1/2}$ and each column of the coefficient matrix. To this end, consider an arbitrary matrix $\mathbf{V} \in \mathbb{R}^{q \times p}$ in general, and denote its vectorized form with $\boldsymbol{\nu} := \text{vec}(\mathbf{V})$ such that $\boldsymbol{\nu} = (\boldsymbol{\nu}_{(1)}^\top, \ldots, \boldsymbol{\nu}_{(p)}^\top)^\top \in \mathbb{R}^{qp}$, where $\boldsymbol{\nu}_{(b)} \in \mathbb{R}^q$ is the b-th column of $\mathbf{V}$, $b = 1, \ldots, p$. To simplify the notation and having one index for the elements of $\boldsymbol{\nu}$, denote the index set of $\boldsymbol{\nu}_{(b)}$ as $\{q(b-1)+1, \ldots, qb\}$, $b = 1, \ldots, p$. Consider $G_b \subset \{q(b-1)+1, \ldots, qb\}$ with cardinality $g_b = \#G_b$ and complement set G_b^c. For a constant $\delta_b \geq 1$, define the set

$$C(G_b; \delta_b) := \{\boldsymbol{\nu}_{(b)} \in \mathbb{R}^q : \quad \|[\boldsymbol{\nu}_{(b)}]_{G_b^c}\|_1 \leq \delta_b \|[\boldsymbol{\nu}_{(b)}]_{G_b}\|_1\},$$

where $[\boldsymbol{\nu}_{(b)}]_{G_b}$ is the restriction of $\boldsymbol{\nu}_{(b)}$ to G_b which has zeros outside the subset G_b.

Definition 4.10.4. A deterministic covariance matrix $\mathbf{Q}$ satisfies the multivariate restricted eigenvalue (MRE) condition over $G = \{G_1, \ldots, G_p\}$ of order $g = \sum_{b=1}^p g_b$, if there exist constants $(\delta_b, \mu_b) \in [1, \infty) \times (0, \infty)$, such that for every $b = 1, \ldots, p$,

$$\|\mathbf{Q}^{1/2}\boldsymbol{\nu}_{(b)}\|_2 \geq \mu_b \|\boldsymbol{\nu}_{(b)}\|_2, \quad \text{for all} \quad \boldsymbol{\nu}_{(b)} \in C(G_b; \delta_b).$$

Definition 4.10.5. The sample covariance matrix $\mathbf{C}_k = \mathbf{X}_k^\top \mathbf{X}_k / n_k$ of the design matrix $\mathbf{X}_k$ satisfies the MRE condition over $G = \{G_1, \ldots, G_p\}$ of order

$g = \sum_{b=1}^{p} g_b$, if there exist constants $(\delta_b, \mu_b) \in [1, \infty) \times (0, \infty)$, such that for every $b = 1, \ldots, p$,

$$\boldsymbol{\nu}_{(b)}^{\top} \mathbf{C}_k \boldsymbol{\nu}_{(b)} \equiv \|\mathbf{X}_k \boldsymbol{\nu}_{(b)}\|_2^2 / n_k \geq \mu_b^2 \|\boldsymbol{\nu}_{(b)}\|_2^2, \quad \text{for all} \quad \boldsymbol{\nu}_{(b)} \in C(G_b; \delta_b). \tag{4.10.5}$$

To introduce Lemma 4.10.6, let $S_2(b)$ be the support of the b-th column of $\boldsymbol{\Gamma}$. By vectorizing $\boldsymbol{\Gamma}$ and rewriting its columns as explained for the general matrix $\mathbf{V}$, we have $S_2(b) \subset \{q(b-1)+1, \ldots, qb\}$, with cardinality $s_2(b) = \#S_2(b)$ and complement set $S_2^c(b)$. For a constant $\delta_b \geq 1$, denote the set

$$C(S_2(b); \delta_b) := \{\boldsymbol{\nu}_{(b)} \in \mathbb{R}^q : \quad \|[\boldsymbol{\nu}_{(b)}]_{S_2^c(b)}\|_1 \leq \delta_b \|[\boldsymbol{\nu}_{(b)}]_{S_2(b)}\|_1\},$$

where $[\boldsymbol{\nu}_{(b)}]_{S_2(b)}$ is the restriction of $\boldsymbol{\nu}_{(b)}$ to $S_2(b)$ which has zeros outside the subset $S_2(b)$.

Lemma 4.10.6. Consider the regression model (4.2.2) for the k-th sub-sample with random Gaussian design $\mathbf{X}_k$ which has independent and identical $\mathcal{N}_q(\mathbf{0}, \mathbf{Q})$ rows, and denote the maximum diagonal element of $\mathbf{Q}$ by $\mathbf{Q}_{\max}$. Suppose that $\mathbf{Q}$ satisfies the MRE condition over S_2 of order s_2 for all vectors in $C(S_2(b); \delta_b)$ with parameters (δ_b, μ_b), $b = 1, \ldots, p$. On the event $\mathcal{F}_k(n_k, p, q)$, for some positive constants c, c' and c'', if the sub-sample size n_k, $k = 1, \ldots, K$, satisfies

$$n_k \geq c'' \frac{\mathbf{Q}_{\max}^2 (1 + \delta_{\max})^2}{\mu_{\min}^2} s_2 \log(q), \tag{4.10.6}$$

where $\mu_{\min} = \min_{1 \leq b \leq p} \mu_b$ and $\delta_{\max} = \max_{1 \leq b \leq p} \delta_b$, then we have

$$\frac{\|\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})\|_F^2}{n_k} \geq (\mu_{\min}/8)^2 \|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_F^2, \tag{4.10.7}$$

with probability at least $1 - c' \exp(-c n_k)$.

Proof. In general, consider the matrix $\mathbf{V} \in \mathbb{R}^{q \times p}$ with the b-th column $\boldsymbol{\nu}_{(b)}$, $b = 1, \ldots, p$, and its vectorized form as $\boldsymbol{\nu} \in \mathbb{R}^{qp}$. We have

$$\frac{\|\mathbf{X}_k \mathbf{V}\|_F}{\sqrt{n_k}} = \frac{\|\mathrm{vec}(\mathbf{X}_k \mathbf{V})\|_2}{\sqrt{n_k}} = \frac{\|(\mathbf{I}_p \otimes \mathbf{X}_k)\boldsymbol{\nu}\|_2}{\sqrt{n_k}} = \sqrt{\frac{\sum_{b=1}^{p} \|\mathbf{X}_k \boldsymbol{\nu}_{(b)}\|_2^2}{n_k}}. \tag{4.10.8}$$

Using Theorem 1 of Raskutti et al. (2010), under the Gaussianity of the design matrix $\mathbf{X}_k$, we have

$$\frac{\|\mathbf{X}_k \boldsymbol{\nu}_{(b)}\|_2}{\sqrt{n_k}} \geq \frac{1}{4} \|\mathbf{Q}^{1/2} \boldsymbol{\nu}_{(b)}\|_2 - 9 \mathbf{Q}_{\max} \sqrt{\frac{\log(q)}{n_k}} \|\boldsymbol{\nu}_{(b)}\|_1, \quad \text{for all} \quad \boldsymbol{\nu}_{(b)} \in \mathbb{R}^q,$$

with probability at least $1 - c' \exp(-cn_k)$. Using the relation between ℓ_1 and ℓ_2 norms, for all vectors $\boldsymbol{\nu}_{(b)} \in C(S_2(b); \delta_b)$ with parameters (μ_b, δ_b), $b = 1, \ldots, p$, we have

$$\begin{aligned}\|\boldsymbol{\nu}_{(b)}\|_1 = \|[\boldsymbol{\nu}_{(b)}]_{S_2(b)}\|_1 + \|[\boldsymbol{\nu}_{(b)}]_{S_2^c(b)}\|_1 &\leq (1+\delta_b)\|[\boldsymbol{\nu}_{(b)}]_{S_2(b)}\|_1 \\ &\leq (1+\delta_b)\sqrt{s_2(b)}\|\boldsymbol{\nu}_{(b)}\|_2. \end{aligned} \tag{4.10.9}$$

Under the MRE condition of $\mathbf{Q}$ for all vectors $\boldsymbol{\nu}_{(b)} \in C(S_2(b); \delta_b)$ with parameters (μ_b, δ_b), $b = 1, \ldots, p$, together with (4.10.9) and the fact that $s_2(b) \leq s_2$, $b = 1, \ldots, p$, we get

$$\frac{\|\mathbf{X}_k \boldsymbol{\nu}_{(b)}\|_2^2}{n_k} \geq \left\{ \frac{1}{4}\mu_{\min} - 9\mathbf{Q}_{\max}(1+\delta_{\max})\sqrt{\frac{s_2 \log(q)}{n_k}} \right\}^2 \|\boldsymbol{\nu}_{(b)}\|_2^2. \tag{4.10.10}$$

Substituting (4.10.10) in (4.10.8),

$$\frac{\|\mathbf{X}_k \mathbf{V}\|_F}{\sqrt{n_k}} \geq \left\{ \frac{1}{4}\mu_{\min} - 9\mathbf{Q}_{\max}(1+\delta_{\max})\sqrt{\frac{s_2 \log(q)}{n_k}} \right\} \sqrt{\sum_{b=1}^{p} \|\boldsymbol{\nu}_{(b)}\|_2^2}. \tag{4.10.11}$$

Applying the lower bound (4.10.6) in (4.10.11) for some constant c'', we get

$$\|\mathbf{X}_k \mathbf{V}\|_F / \sqrt{n_k} \geq (\mu_{\min}/8)\|\mathbf{V}\|_F. \tag{4.10.12}$$

Note that using (4.10.15) from the proof of Lemma 4.10.7, we have

$$\|\hat{\boldsymbol{\Gamma}}_{S_2^c,k} - \boldsymbol{\Gamma}_{S_2^c}\|_1 \leq 4\|\hat{\boldsymbol{\Gamma}}_{S_2,k} - \boldsymbol{\Gamma}_{S_2}\|_1, \tag{4.10.13}$$

where $\boldsymbol{\Gamma}_{S_2}$ and $\hat{\boldsymbol{\Gamma}}_{S_2,k}$ are the matrices $\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\Gamma}}_k$ which have zeros outside the set S_2, respectively. By the same argument as in the proof of Lemma 4.10.7 for univariate linear regression (see for example, Bickel et al., 2009), a similar inequality to (4.10.13) can be written for the b-th column of $\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}$ and it can be deduced that the b-th column of $\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}$ is in $C(S_2(b); \delta_b)$. As such, by replacing $\mathbf{V}$ with $\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}$ in (4.10.12), the result in (4.10.7) follows directly. □

The following Lemma provides an ℓ_1-error bound on $\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}$, by a similar approach to that of Cai et al. (2013) for the Dantzig selector.

Lemma 4.10.7. Under the conditions of Lemma 4.10.6 and the lower bound (4.10.6) on the sub-sample size n_k, with probability at least $1 - c' \exp(-cn_k)$, we have

$$\|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_1 = O_p(s_2 \sqrt{\log(pq)/n_k}). \tag{4.10.14}$$

Proof. Consider the regression model (4.2.2) and recall the Lasso solution (4.3.1). Due to the fact that $\hat{\boldsymbol{\Gamma}}_k$ minimizes the loss in (4.3.1),

$$\|\mathbf{X}_k(\boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_k)\|_F^2/(2n_k) + \rho_k\|\hat{\boldsymbol{\Gamma}}_k\|_1 \leq \operatorname{trace}((\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top \mathbf{X}_k^\top \boldsymbol{\xi}_k)/n_k + \rho_k\|\boldsymbol{\Gamma}\|_1,$$

where

$$\begin{aligned}\operatorname{trace}((\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \boldsymbol{\xi}_k) &\leq \left|\operatorname{trace}((\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \boldsymbol{\xi}_k)\right| \\ &= \left|(\operatorname{vec}(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}))^\top \operatorname{vec}(\mathbf{X}_k^\top \boldsymbol{\xi}_k)\right| \\ &\leq \|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}\|_1 \|\operatorname{vec}(\mathbf{X}_k^\top \boldsymbol{\xi}_k)\|_\infty,\end{aligned}$$

where the last inequality holds due to Hölder's inequality. Now, on the event $\mathcal{F}_k(n_k, p, q)$, by a similar argument as in the univariate regression (see for example Bühlmann and van de Geer (2011), chapter 6), we get

$$\frac{\|\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)\|_F^2}{2n_k} + \frac{\rho_k \|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}\|_1}{2} \leq 2\rho_k \|\hat{\mathbf{\Gamma}}_{S_2,k} - \mathbf{\Gamma}_{S_2}\|_1, \tag{4.10.15}$$

where $\mathbf{\Gamma}_{S_2}$ and $\hat{\mathbf{\Gamma}}_{S_2,k}$ are the matrices $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}_k$ which have zeros outside the set S_2, respectively. Due to the relation between the elementwise ℓ_1 and Frobenius norms of a matrix, we have

$$\|\hat{\mathbf{\Gamma}}_{S_2,k} - \mathbf{\Gamma}_{S_2}\|_1 \leq \sqrt{s_2} \|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}\|_F. \tag{4.10.16}$$

Substituting (4.10.16) in (4.10.15) and then applying Lemma 4.10.6 under the lower bound (4.10.6) on the sub-sample size n_k, we get

$$\frac{\|\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)\|_F^2}{2n_k} + \frac{\rho_k \|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}\|_1}{2} \leq 2\rho_k \sqrt{s_2} \left(\frac{8}{\mu_{\min}}\right) \frac{\|\mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\|_F}{\sqrt{n_k}},$$

with probability at least $1 - c' \exp(-cn_k)$, and

$$\|\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)\|_F^2 / n_k \leq 16\rho_k^2 s_2 (8/\mu_{\min})^2, \quad \|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}\|_1 \leq 16\rho_k s_2 (8/\mu_{\min})^2. \tag{4.10.17}$$

Since $\mu_{\min} \in (0, \infty)$, there exists $L = O(1)$ such that $(8/\mu_{\min})^2 \leq L$, and by considering $\rho_k \asymp \sqrt{\log(pq)/n_k}$, the result of (4.10.14) follows. □

Lemma 4.10.8. Consider the regression model (4.2.2) where $\mathbf{X}_k$ satisfies assumptions (A1) and (A2). Suppose that the eigenvalues of $\mathbf{\Sigma}$, the covariance matrix of the noise, are uniformly bounded. On the event $\mathcal{F}_k(n_k, p, q)$ with regularization parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$, we have

$$\|\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}\|_\infty = O_p(\max\{\sqrt{\log(p)/n_k}, s_2 \log(pq)/n_k\}), \tag{4.10.18}$$

and under the sparsity condition $s_2 = o(n_\dagger^{\pi_1}/(\log(q)\log(pq)))$, $0 < \pi_1 \leq 1/2$, we get $\|\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}\|_\infty = o_p(1)$.

Proof. Recalling that $\boldsymbol{\xi}_k = \mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\Gamma}$, by adding and subtracting $\boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k$ to $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}$, we get

$$\|\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\Sigma}\|_\infty \leq \|\boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k - \boldsymbol{\Sigma}\|_\infty + \|\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k\|_\infty. \tag{4.10.19}$$

Under uniformly boundedness of the eigenvalues of $\boldsymbol{\Sigma}$, using Lemma 2 of Lam and Fan (2009),

$$\|\boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k - \boldsymbol{\Sigma}\|_\infty = O_p(\sqrt{\log(p)/n_k}). \tag{4.10.20}$$

To find an upper bound on the second term of (4.10.19), by adding and subtracting $\mathbf{X}_k\boldsymbol{\Gamma}$ to $\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k$, we get

$$\begin{aligned}
&\|\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} - \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k\|_\infty \\
&\quad = \|(\boldsymbol{\xi}_k - \mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}))^\top(\boldsymbol{\xi}_k - \mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}))/n_k - \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_k/n_k\|_\infty \\
&\quad \leq 2\|(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top\mathbf{X}_k^\top\boldsymbol{\xi}_k/n_k\|_\infty + \|(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top\mathbf{X}_k^\top\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})/n_k\|_\infty.
\end{aligned} \tag{4.10.21}$$

Using Lemma 4.10.7, on the event $\mathcal{F}_k(n_k, p, q)$,

$$2\|(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top\mathbf{X}_k^\top\boldsymbol{\xi}_k/n_k\|_\infty \leq 2\|\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}\|_1\|\mathbf{X}_k^\top\boldsymbol{\xi}_k\|_\infty/n_k = O_p(s_2\log(pq)/n_k). \tag{4.10.22}$$

On the other hand,

$$\|(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top\mathbf{X}_k^\top\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})/n_k\|_\infty \leq \|\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})\|_F^2/n_k = O_p(s_2\log(pq)/n_k), \tag{4.10.23}$$

where the last equality is due to the fact that in (4.10.17), $\mu_{\min} \in (0, \infty)$ and $\rho_k \asymp \sqrt{\log(pq)/n_k}$. Substituting (4.10.22) and (4.10.23) in (4.10.21) and then substituting (4.10.21) and (4.10.20) in (4.10.19), the result in (4.10.18) follows. □

Lemma 4.10.9. Consider the regression model (4.2.2) where $\mathbf{X}_k$ satisfies assumptions (A1) and (A2), and consider the maximum row sparsity of $\mathbf{Q}^{-1}$ as $d_2 = o(\sqrt{n_\dagger}/\log(q))$. Moreover, consider the coefficient matrix $\boldsymbol{\Gamma}$ with sparsity condition $s_2 = o(n_\dagger^{\pi_1}/(\log(pq)\log(q)))$, $0 < \pi_1 \leq 1/2$. Suppose that the event $\mathcal{F}_k(n_k, p, q)$ holds jointly in $k = 1, \ldots, K$. Moreover, suppose that $\hat{\boldsymbol{\Gamma}}_k^d$, $k = 1, \ldots, K$, is the k-th debiased estimator in (4.3.3) with tuning parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$ and $\tilde{\boldsymbol{\Gamma}}_{\text{owAvg}}$ is the pooled estimator in (4.5.4). Let $n_k/n \to c_k \in (0, 1)$ as n_k grows, such that $\lim_{K\to\infty}\sum_{k=1}^K c_k = 1$, and consider the remainder term $\mathbf{R}_{a,\boldsymbol{\Gamma}}$ defined in (4.5.5). Then, it follows that

$$|\mathbf{R}_{a,\boldsymbol{\Gamma}}| = O_p(Ks_2\sqrt{\log(pq)\log(q)/n}), \tag{4.10.24}$$

and if K grows at the rate $K = O(n^{1/4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\}))$, we have $|\mathbf{R}_{a,\boldsymbol{\Gamma}}| = o_p(1)$.

Proof. First note that using Lemma 4.10.8, $|[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}]_{aa}| = O_p(\max\{\sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\})$, where we recall that $\mathbf{A}_{aa}$ is the a-th diagonal element of an arbitrary matrix $\mathbf{A}$. Moreover, under the sparsity condition $d_2 = o(\sqrt{n_\dagger}/\log(q))$, with tuning parameter $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$ in the nodewise Lasso regression (4.9.1), using Lemma 5.4 of van de Geer et al. (2014), we get $|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top - \mathbf{Q}^{-1}]_{aa}| = O_p(\sqrt{d_2\log(q)/n_k})$. As such,

$$|[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} - [\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}| = O_p(\max\{\sqrt{d_2\log(q)/n_k}, \sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\}). \tag{4.10.25}$$

It can be shown that

$$\left|\frac{1}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} - \frac{1}{[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}\right| = O_p(\max\{\sqrt{d_2\log(q)/n_k}, \sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\}). \tag{4.10.26}$$

As such,

$$\begin{aligned}
&\left|\sum_{k=1}^K \{\frac{n_k}{n} \times \frac{[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\} - 1\right| \\
&= [\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}\left|\sum_{k=1}^K \frac{n_k}{n}\{\frac{1}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} - \frac{1}{[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}\}\right| \\
&\leq [\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}\sum_{k=1}^K \frac{n_k}{n}\left|\frac{1}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} - \frac{1}{[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}\right| \\
&= O_P\bigg(K\max\{\sqrt{d_2\log(q)/n}, \sqrt{\log(p)/n}, s_2\log(pq)/\sqrt{nn_\dagger}\}\bigg),
\end{aligned}$$

where the last equality holds using (4.10.26), and the fact that $n_k < n$, $k = 1,\ldots,K$. As such, $\left|\sum_{k=1}^K \{\frac{n_k}{n} \times \frac{[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\} - 1\right| = o_p(1)$ as K grows at the rate $K = O_P\bigg(n^{1/4}/(\sqrt{\log(q)\log(pq)}\max\{s_2, \sqrt{d_2}\})\bigg)$, and as a result $\frac{\sum_{k=1}^K n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{\sum_{k=1}^K n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}} \xrightarrow{p} 1$. By considering the continuous map $g(x) = 1/\sqrt{x}$, the sequence $\sqrt{\frac{\sum_{k=1}^K n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^K n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$ converges in probability to 1 as K and n_k, $k = 1,\ldots,K$, grow. As such, due to the definition of $\mathbf{R}_{a,\mathbf{\Gamma}}$, it is enough to show the bound (4.10.24) for $\sum_{k=1}^K \frac{\sqrt{n_k[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\mathbf{\Sigma}}_k \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{R}_{a,k,\mathbf{\Gamma}}$, with $\mathbf{R}_{a,k,\mathbf{\Gamma}}$ the a-th element, $a = 1,\ldots,qp$, of $\mathbf{R}_{k,\mathbf{\Gamma}}$ from (4.3.3).

We first show that $1/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} = O_p(1)$. Note that $[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$ is nothing else than the a-th element of the outer product of the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ and $\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top$. Thus, it is just needed to find a lower bound for the diagonal elements of these two matrices. By construction, the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ are always positive. Combining Theorem 2.2 of van de Geer et al. (2014) and the reversed triangle inequality, for each $a \in \{1, \ldots, q\}$ we get

$$|\mathbf{Q}_{aa}^{-1}| - |[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| \leq |\mathbf{Q}_{aa}^{-1} - [\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| = o_p(1). \tag{4.10.27}$$

Since $\mathbf{Q}^{-1}$ is positive definite, using assumption (A2), we get $\mathbf{Q}_{aa}^{-1} \geq \Lambda_{\min}(\mathbf{Q}^{-1}) > 1/\Lambda_1$. As such, for sufficiently large n_k, using (4.10.27) we have that $|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| > 0$. Thus, for every $a, b, c \in \{1, \ldots, q\}$, there exists a bound J_k such that

$$\frac{1}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \leq \frac{1}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}]_{bb}|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{cc}|} \leq \frac{1}{J_k} = O_p(1). \tag{4.10.28}$$

By the fact that $[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}$ does not grow with n based on assumption, using Theorem 4.3.1 and combining it with (4.10.28), we have that

$$\begin{aligned} |\sum_{k=1}^{K} \frac{\sqrt{n_k[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \mathbf{R}_{a,k,\boldsymbol{\Gamma}}| &\leq \sum_{k=1}^{K} \sqrt{n_k/n}|\mathbf{R}_{a,k,\boldsymbol{\Gamma}}| \times O_p(1) \\ &\leq O_p(K s_2 \sqrt{\log(q)\log(pq)/n}). \end{aligned}$$

Considering the sparsity conditions $s_2 = o(n_\dagger^{\pi_1}/(\log(pq)\log(q)))$, $0 < \pi_1 \leq 1/2$, $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $K = O(n^{1/4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\}))$, the $o_p(1)$ result follows immediately. □

Lemma 4.10.10. Under the assumptions of Theorem 4.5.1, by considering

$$\zeta_a = \sum_{k=1}^{K} \sqrt{\frac{n_k[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{n[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

where $\mathbf{T}_{a,k}$ is the a-th element of $\mathbf{T}_k$ from (4.3.3), and

$$\zeta_a' = \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

we have $|\zeta_a - \zeta_a'| \xrightarrow{p} 0$, as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$.

Proof. Denoting the sequence $\zeta_a'' = \sum_{k=1}^K \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}}$, we show that $|\zeta_a - \zeta_a''| \xrightarrow{p} 0$ and $|\zeta_a'' - \zeta_a'| \xrightarrow{p} 0$, and then automatically, $|\zeta_a - \zeta_a'| \xrightarrow{p} 0$. We have

$$\begin{aligned}
|\zeta_a - \zeta_a''| &\leq \sum_{k=1}^K \sqrt{\frac{n_k}{n}} \left| \sqrt{\frac{[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}}} - 1 \right| \times \frac{|\mathbf{T}_{a,k}|}{\sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}}} \\
&= \sum_{k=1}^K \sqrt{\frac{n_k}{n}} \frac{\left| \sqrt{[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} - \sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}} \right|}{\sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}}} \\
&\quad \times \frac{|\mathbf{T}_{a,k}|}{\sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}}} \\
&\leq O_p(1) \sum_{k=1}^K \sqrt{\frac{n_k}{n}} \left| \sqrt{[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} - \sqrt{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes \mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top]_{aa}} \right| \times |\mathbf{T}_{a,k}|,
\end{aligned}$$

where the last inequality follows by the fact that $1/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa} = O_p(1)$ using (4.10.28). Moreover, using (4.10.25) we get

$$\begin{aligned}
|\zeta_a - \zeta_a''| \leq \sum_{k=1}^K \sqrt{\frac{n_k}{n}} |\mathbf{T}_{a,k}| \times O_p(\max\{&\sqrt{d_2 \log(q)/n_k}, \sqrt{\log(p)/n_k}, \\
&s_2 \log(pq)/n_k\}).
\end{aligned} \tag{4.10.29}$$

On the other hand,

$$\mathbf{T}_k = \mathrm{vec}(\mathbf{M}_k \mathbf{X}_k^\top \boldsymbol{\xi}_k / \sqrt{n_k}) = (\mathbf{I}_p \otimes \mathbf{M}_k) \mathrm{vec}(\mathbf{X}_k^\top \boldsymbol{\xi}_k) / \sqrt{n_k},$$

and working on the event $\mathcal{F}_k(n_k, p, q)$, by considering $\rho_k \asymp \sqrt{\log(pq)/n_k}$,

$$\|\mathbf{T}_k\|_\infty \leq \sqrt{n_k} O_p(\sqrt{\log(pq)/n_k}) \|\mathbf{I}_p \otimes \mathbf{M}_k\|_\infty. \tag{4.10.30}$$

As was shown in the proof of Lemma 4.4.1,

$$\|\mathbf{I}_p \otimes \mathbf{M}_k\|_\infty = \|\mathbf{M}_k\|_\infty = \max_{j \in \{1,\ldots,q\}} \|\mathbf{M}_{j,k}\|_1 = O_p(\sqrt{d_2}),$$

where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$. As such, in (4.10.30), we get $\|\mathbf{T}_k\|_\infty \leq O_p(\sqrt{d_2 \log(pq)})$. Substituting this result in (4.10.29), we have

$$\begin{aligned}
|\zeta_a - \zeta_a''| &\leq \sum_{k=1}^K \sqrt{\frac{n_k}{n}} O_p(\sqrt{d_2 \log(pq)} \times \max\{\sqrt{d_2 \log(q)/n_k}, \sqrt{\log(p)/n_k}, \\
&\qquad s_2 \log(pq)/n_k\}) \\
&\leq O_p(K \sqrt{d_2 \log(pq)/n} \times \max\{\sqrt{d_2 \log(q)}, \sqrt{\log(p)}, \\
&\qquad s_2 \log(pq)/\sqrt{n_\dagger}\}),
\end{aligned}$$

where by considering the sparsity conditions $s_2 = o(n_\dagger^{\pi_1}/(\log(pq)\log(q)))$, $0 < \pi_1 \leq 1/2$, $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $K = O(n^{1/4}/(\sqrt{\log(pq)\log(q)} \times \max\{s_2, \sqrt{d_2}\}))$, the $o_p(1)$ result follows immediately.

Now, to show $|\zeta_a'' - \zeta_a'| \xrightarrow{p} 0$, by the same argument,

$$
\begin{aligned}
|\zeta_a'' - \zeta_a'| &\leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \left| \sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} - \sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \right| \\
&\quad \times |\mathbf{T}_{a,k}| \\
&\leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \left\{ \left| \sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} - \sqrt{[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} \right| \right. \\
&\quad \left. + \left| \sqrt{[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} - \sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \right| \right\} |\mathbf{T}_{a,k}|.
\end{aligned}
$$

Similarly to the proof of $|\zeta_a - \zeta_a''| \xrightarrow{p} 0$, by considering the mentioned sparsity conditions, we conclude that $|\zeta_a'' - \zeta_a'| \xrightarrow{p} 0$. □

Lemma 4.10.11. Consider the regression model (4.2.2), and suppose that assumptions (B1)-(B2) and (C1)-(C3) from Appendix 4.9.3 hold. Consider the coefficient matrix $\boldsymbol{\Gamma}$ with sparsity condition $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, and the precision matrix $\boldsymbol{\Theta}$ with maximum node degree $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$. Suppose that the event $\mathcal{F}_k(n_k, p, q)$ holds jointly in $k = 1, \ldots, K$. Moreover, suppose that $\hat{\boldsymbol{\Theta}}_k^d$, $k = 1, \ldots, K$, is the k-th debiased estimator in (4.4.3) with tuning parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$ and $\hat{\boldsymbol{\Theta}}_{\text{owAvg}}$ is the pooled estimator in (4.5.6). Consider the remainder term $\mathbf{R}_{ab,\boldsymbol{\Theta}}$ defined in (4.5.7). Then it follows that

$$
|\mathbf{R}_{ab,\boldsymbol{\Theta}}| = O_p\left(\frac{K}{\sqrt{n}} \max\left\{ d_1^{3/2} \log(p), d_1^2 (\log(p))^{3/2} / \sqrt{n_\dagger}, d_1 s_2 \log(pq) \right\} \right), \quad (4.10.31)
$$

and if K grows as $K = O(n^{1/3}/(d_1 \log(p)))$, we have $|\mathbf{R}_{ab,\boldsymbol{\Theta}}| = o_p(1)$.

Proof. By a similar argument as in the proof of Lemma 4.10.9 and using Remark 4.2, $\left| \frac{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}{\sum_{k=1}^K n_k/\sigma_{ab}^2} - 1 \right| = O_p\left(K \max\{ \log(p)/n, \sqrt{d_1 \log(p)/n} \} \right)$, which is of order $o_p(1)$ as K grows at the rate $K = O(n^{1/3}/(d_1 \log(p)))$. As a result, by considering the continuous map $g(x) = 1/\sqrt{x}$, the sequence $\sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}}$ converges in probability to 1 as K and n_k, $k = 1, \ldots, K$, grow. As such, we just need to show the boundedness of $\sum_{k=1}^K \frac{\sqrt{n_k}\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2} \times \mathbf{R}_{ab,k,\boldsymbol{\Theta}}$. Due to the positive definiteness of the graphical Lasso estimator defined in (4.4.1), there exists a positive constant L_k such that with high probability $\hat{\sigma}_{ab,k}^2 \geq \Lambda_{\min}^2(\hat{\boldsymbol{\Theta}}_k) > L_k$, where $\Lambda_{\min}(\hat{\boldsymbol{\Theta}}_k)$ is the

minimum eigenvalue of $\hat{\mathbf{\Theta}}_k$ and then the term $1/\hat{\sigma}^2_{ab,k} = O_p(1)$. Moreover, using (4.4.5),

$$|\sum_{k=1}^{K} \frac{\sqrt{n_k}\sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}} \mathbf{R}_{ab,k,\mathbf{\Theta}}| \leq \sum_{k=1}^{K} \frac{1}{\sqrt{n}} O_p(\max\{d_1^{3/2}\log(p), d_1^2(\log(p))^{3/2}/\sqrt{n_\dagger}, d_1 s_2 \log(pq)\}),$$

and the result in (4.10.31) follows directly. By considering the mentioned sparsity and $K = O(n^{1/3}/(d_1\log(p)))$, one can reach the $o_p(1)$ rate. □

4.11 Appendix C: Extra simulation results

Table 4.4: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets, for the proposed estimators and different competitors, when $n = 25000$.

			Active set				Non-active set			
			K				K			
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Debiased	Full	1.00				6.50			
			(.01)				(.01)			
		owAvg		1.03	1.08	1.18		6.72	7.15	7.79
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		Top1		1.39	1.46	1.19		9.12	9.60	7.85
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		sAvg		1.59	2.88	3.95		10.27	16.91	20.15
				(.01)	(.02)	(.03)		(.03)	(.02)	(.03)
		wAvg		1.04	1.25	1.81		6.80	8.00	10.10
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Sparse	SFull	2.74				.35			
			(.01)				(.00)			
		STop1		2.82	2.84	2.88		1.31	1.56	2.28
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SsAvg		2.54	2.29	2.04		4.29	7.37	8.83
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SwAvg		2.59	2.45	2.22		1.93	2.77	4.01
				(.01)	(.01)	(.01)		(.00)	(.00)	(.01)
$\boldsymbol{\Gamma}$	Debiased	Full	1.55				15.43			
			(.01)				(.02)			
		owAvg		1.62	1.74	1.55		16.18	17.30	14.28
				(.02)	(.02)	(.01)		(.02)	(.02)	(.02)
		Top1		2.21	2.34	2.67		22.08	23.33	26.65
				(.02)	(.02)	(.03)		(.03)	(.03)	(.03)
		sAvg		2.42	3.91	7.09		24.17	39.04	70.18
				(.02)	(.04)	(1.45)		(.03)	(.06)	(14.36)
		wAvg		1.63	1.90	3.38		16.30	18.93	33.51
				(.02)	(.02)	(.61)		(.02)	(.02)	(6.03)
	Sparse	SFull	4.79				1.82			
			(.00)				(.00)			
		STop1		4.80	4.80	4.80		1.91	1.93	1.99
				(.00)	(.00)	(.00)		(.00)	(.00)	(.01)
		SsAvg		4.67	4.23	4.17		2.03	2.82	3.11
				(.12)	(.02)	(.02)		(.02)	(.01)	(.01)
		SwAvg		4.75	4.59	4.51		1.82	1.84	1.98
				(.04)	(.01)	(.01)		(.02)	(.00)	(.00)

Table 4.5: Average coverage probability and average length of the confidence intervals over 500 repetitions for the proposed estimators and different competitors, when $n = 25000$.

			Avg.Cov				Avg.Len			
			K				K			
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Active set	Full	.94				.02			
		owAvg		.94	.94	.95		.02	.03	.03
		Top1		.94	.94	.94		.03	.04	.03
		sAvg		.92	.89	.90		.04	.06	.08
		wAvg		.95	.98	.98		.03	.04	.06
	Non-active set	Full	.95				.02			
		owAvg		.95	.96	.96		.02	.03	.03
		Top1		.96	.96	.95		.03	.04	.03
		sAvg		.94	.94	.97		.04	.06	.08
		wAvg		.97	.98	.98		.03	.04	.06
$\boldsymbol{\Gamma}$	Active set	Full	.95				.08			
		owAvg		.95	.95	.92		.08	.09	.07
		Top1		.95	.95	.95		.11	.12	.14
		sAvg		.94	.92	.90		.12	.17	.27
		wAvg		.96	.97	.98		.09	.12	.20
	Non-active set	Full	.95				.08			
		owAvg		.95	.95	.94		.08	.09	.07
		Top1		.95	.95	.95		.11	.12	.14
		sAvg		.94	.92	.90		.12	.17	.27
		wAvg		.96	.97	.98		.09	.12	.20

CHAPTER 5

The DistributedGGL R package

This chapter provides an implementation of the procedures developed in Chapters 2-4 in R. To make the implementation convenient and user-friendly, an R package called **DistributedGGL** is constructed, and it is extensively explained in this chapter.

5.1 Introduction

The R package **DistributedGGL** provides tools for estimation and inference of Gaussian graphical models under the distributed setting. The package is provided with documentation and illustration of the available functions. This package supports different distributed estimation methods of the precision matrix of a multivariate Gaussian distribution including optimally weighted, sample-size weighted and simple average methods. Parallel computing with different sub-samples is implemented with the R package **foreach** (Microsoft and Weston, 2009). Moreover, graphical Lasso estimation for each sub-sample is performed using the R package **huge** (Jiang et al., 2010). Statistical inference including asymptotic confidence intervals and a data driven threshold for the hypothesis testing procedure based on the proposed methods are also implemented in this package.

Furthermore, estimation and inference for the coefficient and precision matrices for covariate adjusted Gaussian graphical models are also implemented in this package. Estimation of the coefficient matrix for each sub-sample is implemented with the package **MRCE** (Rothman, 2012). Moreover, the nodewise Lasso regression when estimating the inverse sample covariance of the covariates is implemented with the **glment** package (Friedman et al., 2008). Finally, linear algebra operations are implemented with C++, and integrated in R via the **Rccp** package (Eddelbuettel and Balamuta, 2018).

5.2 Implementation of the optimally weighted average estimator in Gaussian graphical models

Before explaining the function, we recall the model and the object of estimation briefly for the reader's convenience.

To estimate the precision matrix based on the optimally weighted average method, we first estimate the precision matrix based on K parallel distributed jobs and then we aggregate estimates to the final estimate based on the estimator

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \hat{\mathbf{\Theta}}^d_{ab,k},$$

for every $(a,b) \in \mathcal{V} \times \mathcal{V}$, where $\hat{\mathbf{\Theta}}^d_{ab,k}$ is the (a,b)-th entry of the k-th debiased estimator $\hat{\mathbf{\Theta}}^d_k$, $k = 1, \ldots, K$, defined by Jankova and van de Geer (2015), of the form

$$\hat{\mathbf{\Theta}}^d_k = 2\hat{\mathbf{\Theta}}_k - \hat{\mathbf{\Theta}}_k \hat{\mathbf{\Sigma}}_k \hat{\mathbf{\Theta}}_k,$$

where $\hat{\mathbf{\Theta}}_k$ is the graphical Lasso estimator, defined by Friedman et al. (2008), of the form

$$\hat{\mathbf{\Theta}}_k = \arg \min_{\mathbf{\Theta} \in S^p_{++}} \left\{ \text{trace}(\hat{\mathbf{\Sigma}}_k \mathbf{\Theta}) - \log \det(\mathbf{\Theta}) + \lambda_k \|\mathbf{\Theta}\|_{1,\text{off}} \right\},$$

based on the k-th sub-sample, where $\lambda_k > 0$ is the regularization parameter, and $\hat{\mathbf{\Sigma}}_k$ is the sample covariance matrix obtained based on the k-th sub-sample of the form $\hat{\mathbf{\Sigma}}_k = \mathbf{Y}_k^\top \mathbf{Y}_k / n_k$, with $\mathbf{Y}_k$ a matrix of dimension $n_k \times p$, and n_k the sub-sample size in the k-th sub-sample. Note that, $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k} \hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k}$.

5.2.1 Arguments and output

The function `owAvg()` implements the owAvg estimator in `R`, and the output of this function can be obtained by calling

```
owAvg(data, K, lambda, SubSize, alpha).
```

The argument `data` is the input data matrix $\mathbf{Y}$ of dimension $n \times p$, where n is the total sample size and p is the number of variables (nodes). The argument `K` is the number of sub-samples (machines) for splitting the dataset. The parameter `lambda` is a sequence of regularization parameters of length K, for the graphical Lasso optimization problem, the same for each sub-sample. The argument `SubSize` is a vector of length K which indicates the sub-sample sizes n_k, $k = 1, \ldots, K$, for K parallel jobs. Finally, the argument `alpha` is the significance level for providing

$100(1-\alpha)\%$ asymptotic confidence intervals for every element of the precision matrix $\mathbf{\Theta}$, which is of the form

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}},$$

where $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution.

The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ containing the final aggregated estimates corresponding to $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum $\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}$ which is needed to perform inference for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `RTime`: The total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Weight`: A list of K matrices corresponding to the estimated weights. The weight of the k-th machine to every pair (a,b) of the precision matrix is $(1/\sum_{l=1}^{K} \frac{n_l}{\hat{\sigma}^2_{ab,l}}) \times (\frac{n_k}{\hat{\sigma}^2_{ab,k}})$.

- `Lower`: A $p \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure, for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure, for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval for every entry $\mathbf{\Theta}_{ab}$, $(a,b) \in \mathcal{V} \times \mathcal{V}$, based on the distributed procedure which is of the form $2\Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}$.

- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$, based on the distributed procedure. The difference between this output and `Lower` is in the quantile $\Phi^{-1}(\cdot)$ which is of the form $\Phi^{-1}(1-\alpha/(p(p-1)))$ with the Bonferroni correction.

- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$, based on the distributed procedure.

- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the distributed procedure with Bonferroni correction which is of the form $2\Phi^{-1}(1-\alpha/(p(p-1)))/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}$, for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

5.2.2 Illustration

We start by loading the packages required in this example.

```
R> library(huge)
R> library(mvtnorm)
R> library(DistributedGGL)
R> library(igraph)
```

Throughout this illustration, we generate a dataset from a multivariate normal distribution with mean vector $\mathbf{0}$, covariance matrix $\mathbf{\Sigma}$, and precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. To generate a precision matrix, we consider a random graph structure using the function `huge.generator()` from the **huge** package. Other graph structures such as "hub", "cluster", and "band" can be also considered. The number of nodes and the probability of connection between every pair of nodes are set to $p = 300$ and prob $= 0.05$. The following code generate a positive definite covariance matrix and precision matrix of the random graph.

```
R> p = 300
R> prob = 0.05

R> set.seed(123)
R> obj1 = huge.generator(prob = prob, d = p, graph = "random", v=1,
+                         vis = FALSE, verbose=FALSE)

R> mysigma = obj1$sigma
R> myTheta = obj1$omega

R> par(mar = c(4, 5.2, 0.5, 1))
R> x <- (1:nrow(as.matrix(obj1$theta)))
R> y <- (1:ncol(as.matrix(obj1$theta)))
R> mat1 <- (as.matrix(obj1$theta))
R> image(x, y, mat1, col = c(0,1), axes = F, xlab = "NODE",
+        ylab = "NODE", cex.lab = 3, bty = "n")
R> axis(2, c(1,seq(100,500, by=100)), las = 2, cex.axis = 1.6,
+       tck = -.01)
R> axis(1, c(1,seq(100,500, by=100)), cex.axis = 1.6, tck = -.01)

R> par(mar = c(0, 0, 0, 0))
```

```
R> g = graph.adjacency(as.matrix(obj1$theta), mode = "undirected",
+                       diag = FALSE)
R> layout.grid = layout.fruchterman.reingold(g)
R> plot(g, layout = layout.grid, edge.color = "gray50",
+       vertex.color = "red", vertex.size = 3, vertex.label = NA)
```

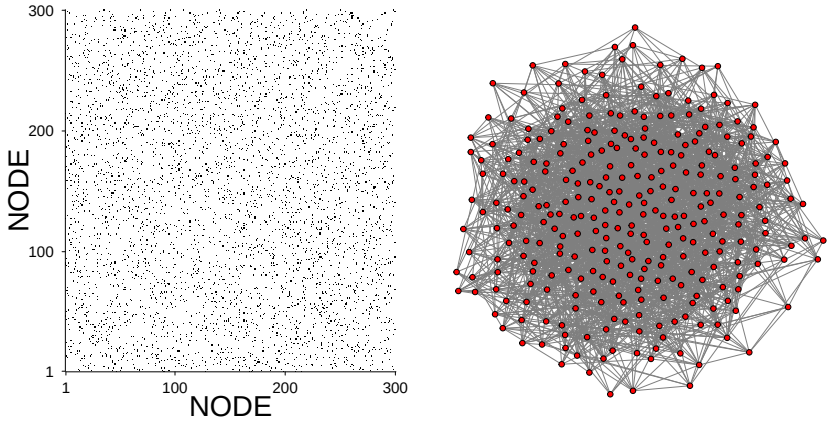


Figure 5.1: Visual representation of the adjacency matrix of the precision matrix $\boldsymbol{\Theta}$ and its graph structure for $p = 300$ nodes. A black dot on the left hand figure denotes an edge between the corresponding nodes.

As it is observed in the code above, using the function `image()` from the `R` package **igraph** (Csardi et al., 2005) and the function `plot()`, we can visualize the adjacency matrix obtained from the structure of graph. The output of these two functions is presented in Figure 5.1. To generate $n = 10000$ sample data from a p-dimensional normal distribution with this graph structure, we use the function `rmvnorm()` from the package **mvtnorm** (Genz et al., 2023) as follows.

```
R> n=10000
R> set.seed(123)
R> data = rmvnorm(n, sigma = mysigma)
```

To perform estimation and inference in the distributed setting, one needs the sub-sample sizes for K sub-samples. The number of sub-samples is set to $K = 10$ in this example and the splitting setting is considered the same as in Nezakati and Pircalabelu (2023c), see Chapter 2, which is as follows. Among all available machines, two of them are considered as powerful and the remaining ones are considered as ordinary. The first one is the most powerful and $(55 - K)\%$ of the dataset is distributed on this machine. The second one is less powerful than the first and $(60 - K)\%$ of the remaining dataset is distributed on this one. The remaining

dataset is distributed roughly equally on the remaining machines. The following code gives the size of these K sub-samples.

```
R> K = 10
R> size1 = floor(((55/100)-(K/100))*n)
R>   aa = (55/100)-(K/100)
R>   SubSize = c()
R>   if ( K==1 ) { SubSize = n
+   } else if ( K==2 ) {
+     SubSize[1] = size1
+     SubSize[2] = n-SubSize[1]
+   } else if ( K==3 ) {
+     set.seed(2020)
+     SubSize[1] = size1
+     SubSize[2] = floor( ((60/100)-(K/100))*(1-aa)*n )
+     SubSize[3] = n-sum(SubSize[1],SubSize[2])
+   } else {
+     SubSize[1] = size1
+     SubSize[2] = floor( ((60/100)-(K/100))*(1-aa)*n  )
+     SubSize[3:(K-1)] = floor((n-(SubSize[1]+SubSize[2]))/(K-2))
+     SubSize[K] = n-sum(SubSize[1:(K-1)])
+   }

R> SubSize
[1] 4500 2749  343  343  343  343  343  343  343  350
```

In the sequel, the regularization parameter is considered as $\lambda = \sqrt{\log(p)/n}$ for all K sub-samples, and the significance level α is set to $\alpha = 0.05$.

```
R> lambda = rep(sqrt(log(p)/n), K); alpha = 0.05
R> obj2 = owAvg(data, K, lambda, SubSize, alpha)

R> obj2$ThetaTilde[1:4, 1:4]
             [,1]         [,2]        [,3]         [,4]
[1,]  1.520939446  0.004047655 -0.01018492 -0.002265978
[2,]  0.004047282  1.760001265  0.01936385 -0.001634396
[3,] -0.010183839  0.019364070  1.93312186  0.019853876
[4,] -0.002266144 -0.001633657  0.01985434  1.683757419

R> obj2$Lower[1:4, 1:4]
            [,1]        [,2]         [,3]         [,4]
[1,]  1.48427561 -0.02347224 -0.038780947 -0.029224410
[2,] -0.02347262  1.71872519 -0.010973330 -0.030246595
[3,] -0.03877987 -0.01097311  1.888575811 -0.009854104
[4,] -0.02922458 -0.03024586 -0.009853644  1.644153120
```

```
R> obj2$Upper[1:4, 1:4]
           [,1]       [,2]       [,3]       [,4]
[1,] 1.55760328 0.03156755 0.01841111 0.02469245
[2,] 0.03156718 1.80127734 0.04970104 0.02697780
[3,] 0.01841219 0.04970125 1.97766791 0.04956186
[4,] 0.02469229 0.02697854 0.04956232 1.72336172

R> obj2$Len[1:4, 1:4]
           [,1]       [,2]       [,3]       [,4]
[1,] 0.07332767 0.05503980 0.05719206 0.05391686
[2,] 0.05503980 0.08255214 0.06067437 0.05722440
[3,] 0.05719205 0.06067437 0.08909210 0.05941596
[4,] 0.05391686 0.05722440 0.05941596 0.07920860
```

The results presented above include the final aggregated estimate, the lower bound, the upper bound, and the length of asymptotic confidence intervals for $\boldsymbol{\Theta}_{ab}$, $a, b = 1, \ldots, 4$, respectively.

5.3 Implementation of edge detection with false discovery rate control

The following algorithm is a summary of the edge detection procedure using the optimally weighted average estimator, described in Chapter 3, for the multiple hypothesis testing $H_{0,ab} = \boldsymbol{\Theta}_{ab} = 0$ vs $H_{1,ab} = \boldsymbol{\Theta} \neq 0$, $1 \leq a < b \leq p$.

Algorithm 1: Edge detection procedure with false discovery control

Step1. Construct the test statistic $\hat{U}_{ab} = \sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}}\tilde{\boldsymbol{\Theta}}_{ab,\text{owAvg}}$ using the optimally weighted average estimator $\tilde{\boldsymbol{\Theta}}_{\text{owAvg}}$.

Step2. For a pre-specified level $0 < q < 1$, let $t_p = (2\log(p_1) - 2\log(\log(p_1)))^{1/2}$, and compute

$$t_0' = \inf\left\{0 \leq t \leq t_p : \frac{p_1 G(t)}{\max\{R'(t), 1\}} \leq q\right\},$$

where $R'(t) = \sum_{1 \leq a < b \leq p} \mathbb{I}(|\hat{U}_{ab}| \geq t)$, $G(t) = 2(1 - \Phi(t))$ with $\Phi(\cdot)$ as the cumulative distribution function of the standard normal distribution, and $p_1 = p(p-1)/2$. If t_0' does not exist, set $t_0' = \sqrt{2\log(p_1)}$.

Step3. Reject $H_{0,ab}$ if $|\hat{U}_{ab}| \geq t_0'$.

The functions `Distrib.dGL()` and `FDR()` implement Algorithm 1 in `R`, for which a comprehensive explanation is provided in Section 5.3.1.

5.3.1 Arguments and output

The function `Distrib.dGL()` implements the test statistic $\hat{U}_{ab}$ in `R`, and the output of this function can be obtained by calling

```
Distrib.dGL(DeTheta, VarInv).
```

The argument `DeTheta` is a $p \times p$ dimensional matrix, whose (a,b)-th entry is the owAvg estimator $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ for every $(a,b) \in \mathcal{V} \times \mathcal{V}$. This estimate can be obtained in `R` using the function `owAvg()`, as explained in Section 5.2.1. The argument `VarInv` is a matrix of dimension $p \times p$, whose (a,b)-th entry is the inverse of the variance of $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ for every $(a,b) \in \mathcal{V} \times \mathcal{V}$, which is of the form $\sum_{k=1}^{K} n_k / \hat{\sigma}^2_{ab,k}$. This matrix is also an output of the function `owAvg()` in `R`.

The output of this function is a list including

- `a`: The row index of $\hat{U}_{ab}$, which is equal to $a = 1, \ldots, p-1$.
- `b`: The column index of $\hat{U}_{ab}$, which is equal to $b = a, \ldots, p$
- `stat.ab`: The test statistic $\hat{U}_{ab}$ for every $1 \leq a < b \leq p$.
- `debiased.est.ab`: The debiased estimator $\hat{\mathbf{\Theta}}^{d}_{ab,k}$ for every $\leq a < b \leq p$.

After taking the absolute value of the test statistic $\hat{U}_{ab}$, we find the threshold t_0' using the function `FDR()`. The usage of this function is summarized as follows.

```
FDR(stat, q)
```

The argument `stat` is a vector of length $p(p-1)/2$ containing the absolute values of $\hat{U}_{ab}$ for every $1 \leq a < b \leq p$, which can be constructed using `Distrib.dGL()`. The argument `q` is a pre-specified level between zero and one from Step 2 of Algorithm 1, which should be specified by the user. This function searches for the value of t_0', as explained in Algorithm 1. The output of this function is a list including

- `which.rejected`: A vector containing the absolute values of the test statistics $\hat{U}_{ab}$, $1 \leq a < b \leq p$, for which $H_{0,ab}$ is rejected (corresponding to Step 3 of Algorithm 1).
- `which.t0`: An index number (1 or 0) which shows whether the threshold t_0' defined by the procedure has been chosen or the alternative $t_0' = \sqrt{2\log(p_1)}$.
- `t.hat`: The value of t_0' which has been chosen by the procedure.

5.3.2 Illustration

As a continuation of the illustration in Section 5.2.2, after estimating $\mathbf{\Theta}$, to select the edges using Algorithm 1, we first construct the test statistic using the function `Distrib.dGL()` as follows

```
R> dbGL <- Distrib.dGL(DeTheta = obj2[[1]], VarInv= obj2[[2]])
```

where we substitute the estimation of $\mathbf{\Theta}$ and its variance when using the function `owAvg()`. Using the third output of `dbGL`, which is the absolute value of the test statistic $\hat{U}_{ab}$, we can find the threshold t_0' and the test statistics correspond to the rejected set, as follows

```
R> q = 0.1
R> FDR.FUNC = FDR(stat = abs(dbGL[,3]), q = q)

R> FDR.FUNC$which.t0
[1] 1
R> FDR.FUNC$t.hat
[1] 2.81

R> dGL.pairs<- dbGL[FDR.FUNC[[1]],1:2]

R> dGL.pairs[1:5,]
     [,1] [,2]
[1,]    1   47
[2,]    1   90
[3,]    1   91
[4,]    1  128
[5,]    1  168
```

The first and second output show that the threshold t_0' has been chosen by the procedure defined in Step 2 of Algorithm 1 (not the alternative value), and its value is 2.81, respectively. The last output represents the indices of 5 pairs $\mathbf{\Theta}_{ab}$, for which $H_{0,ab} = 0$ is rejected. It means that there are estimated edges between pairs $(1, 47), (1, 90), \ldots, (1, 168)$ in this special example.

5.4 Implementation of the optimally weighted average estimator for covariate adjusted graphical models

In this section, we implement the estimation of the coefficient matrix $\mathbf{\Gamma}$ and of the precision matrix $\mathbf{\Theta}$, under the distributed setting, in the multivariate regression

model

$$\dot{\mathbf{Y}} = \mathbf{\Gamma}^\top \dot{\mathbf{X}} + \dot{\varepsilon},$$

where $\mathbf{\Theta}$ is the inverse of the covariance matrix $\mathbf{\Sigma}$, such that $\dot{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The implemented functions in R, compute $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ introduced in Chapter 4. For the reader's convenience, we summarize briefly these two estimators in this section.

The final aggregated estimator for the a-th element of $\text{vec}(\mathbf{\Gamma})$, $a = 1, \ldots, pq$, is of the form

$$\begin{aligned}\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a = &\left(\sum_{k=1}^{K} \frac{n_k}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \right)^{-1} \\ &\times \sum_{k=1}^{K} \frac{n_k}{[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \text{vec}(\hat{\mathbf{\Gamma}}_k^d)_a,\end{aligned}$$

where $\text{vec}(\hat{\mathbf{\Gamma}}_k^d)_a$ is the a-th element of the vectorized form of $\hat{\mathbf{\Gamma}}_k^d$, the k-th debiased estimator, which is of the form

$$\hat{\mathbf{\Gamma}}_k^d = \hat{\mathbf{\Gamma}}_k + \mathbf{M}_k \mathbf{X}_k^\top (\mathbf{Y}_k - \mathbf{X}_k \hat{\mathbf{\Gamma}}_k)/n_k,$$

where $\hat{\mathbf{\Gamma}}_k$ is the solution to the following optimization problem

$$\hat{\mathbf{\Gamma}}_k = \arg\min_{\mathbf{\Gamma}} \left\{ \frac{1}{2n_k} \text{trace}\{(\mathbf{Y}_k - \mathbf{X}_k \mathbf{\Gamma})^\top (\mathbf{Y}_k - \mathbf{X}_k \mathbf{\Gamma})\} + \rho_k \|\mathbf{\Gamma}\|_1 \right\}, \tag{5.4.1}$$

where $\rho_k > 0$ is the regularization parameter. Note that $\mathbf{C}_k = \mathbf{X}_k^\top \mathbf{X}_k / n_k$ is the sample covariance matrix of the covariates $\mathbf{X}_k$, and $\mathbf{M}_k$ is an approximation for the inverse of $\mathbf{C}_k$, which can be found using nodewise Lasso regressions. Moreover, $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k \hat{\mathbf{\Gamma}}_k)^\top (\mathbf{Y}_k - \mathbf{X}_k \hat{\mathbf{\Gamma}}_k)/n_k$.

After estimating $\hat{\mathbf{\Gamma}}_k$, the final aggregated estimator of $\mathbf{\Theta}_{ab}$, $(a, b) \in \mathcal{V} \times \mathcal{V}$, can be constructed as

$$\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}} = \left(\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_{ab,k}^2} \right)^{-1} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_{ab,k}^2} \hat{\mathbf{\Theta}}_{ab,k}^d,$$

where $\hat{\mathbf{\Theta}}_{ab,k}^d$ is the (a, b)-th pair of the debiased estimator of $\mathbf{\Theta}$ in the k-th subsample, which is of the form

$$\hat{\mathbf{\Theta}}_k^d = 2\hat{\mathbf{\Theta}}_k - \hat{\mathbf{\Theta}}_k \hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \hat{\mathbf{\Theta}}_k,$$

and $\hat{\mathbf{\Theta}}_k$ is the graphical Lasso estimator of Friedman et al. (2008), which is the solution to the following optimization problem

$$\hat{\mathbf{\Theta}}_k = \arg\min_{\mathbf{\Theta} \in \mathcal{S}_{++}^p} \left\{ \text{trace}(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_k \|\mathbf{\Theta}\|_{1,\text{off}} \right\}, \tag{5.4.2}$$

where $\lambda_k > 0$ is the regularization parameter.

As it is shown in Chapter 4, the aggregated estimators $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ and $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ are asymptotically normal. As such, performing inference in the form of hypothesis testing and confidence intervals is feasible based on these estimators. In the next section the functions `owAvgCondGamma()` and `owAvgCondTheta()` are introduced for estimating and performing inference for the elements of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively.

5.4.1 Function and output

The function `owAvgCondGamma()` implements the owAvg estimator $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ of $\text{vec}(\mathbf{\Gamma})_a$, $a = 1, \ldots, qp$, in `R` and the output of this function can be obtained by calling

```
owAvgCondGamma(Y, X, K, rho, lambda, rhotilde, SubSize, alpha).
```

The arguments `Y` and `X` are input response and covariate matrices of dimension $n \times p$ and $n \times q$, respectively, where n is the total sample size, and p and q are the number of responses (nodes) and covariates, respectively. The parameters `rho` and `lambda` are respectively the regularization parameters in the optimization problem (5.4.1) and (5.4.2), and both are vectors of length K. The argument `rhotilde` is a sequence of regularization parameters of length K, for the nodewise Lasso regression to estimate $\mathbf{M}_k$, $k = 1, \ldots, K$, which is an approximation for the inverse of $\mathbf{C}_k$ based on the k-th sub-sample, when $q > n_k$. The argument `SubSize` is a vector of length K which indicates the sub-sample sizes n_k, $k = 1, \ldots, K$, for K parallel jobs. Finally, the argument `alpha` is the significance level for providing $100(1-\alpha)\%$ confidence intervals for every element of the vectorized form of the coefficient matrix $\mathbf{\Gamma}$, which is of the form

$$\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}.$$

The output of this function is a list containing

- `GammaTilde`: The $q \times p$ dimensional matrix $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ containing the final aggregated estimates for $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ for every element $a = 1, \ldots, qp$.
- `SigmaHat`: A $q \times p$ dimensional matrix containing the sum $\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}$ needed for performing inference for every element $a = 1, \ldots, qp$.
- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.
- `Lower`: A $q \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure for every element $a = 1, \ldots, qp$.

- `Upper`: A $q \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure for every element $a = 1, \ldots, qp$.

- `Len`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the distributed procedure which is of the form $2\Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}$ for every element $a = 1, \ldots, qp$.

- `LowerB`: A $q \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with the Bonferroni correction based on the distributed procedure for every element $a = 1, \ldots, qp$. The difference between this output and `Lower` is in the quantile Φ^{-1} which is of the form $\Phi^{-1}(1-\alpha/(p(p-1)))$ with Bonferroni correction.

- `UpperB`: A $q \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the distributed procedure for every element $a = 1, \ldots, qp$.

- `LenB`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the distributed procedure with Bonferroni correction which is of the form $2\Phi^{-1}(1-\alpha/(p(p-1)))/\sqrt{\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}$ for every element $a = 1, \ldots, qp$.

The function `owAvgCondTheta()` implements the owAvg estimator $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ of $\mathbf{\Theta}_{ab}$, $(a,b) \in \mathcal{V} \times \mathcal{V}$, in `R` and the output of this function can be obtained by calling

```
owAvgCondTheta(Y, X, K, rho, lambda, SubSize, alpha).
```

The arguments of this function are the same as those of `owAvgCondGamma()` except `rhotilde` which is not needed in estimation of $\mathbf{\Theta}$. The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ containing the final aggregated estimates for $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum $\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}$ needed to perform inference for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Lower`: A $p \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ based on the distributed procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the distributed procedure which is of the form $2\Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with the Bonferroni correction based on the distributed procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$. The difference between this output and `Lower` is in the quantile Φ^{-1} which is of the form $\Phi^{-1}(1-\alpha/(p(p-1)))$ with Bonferroni correction.

- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound in the mentioned asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the distributed procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the distributed procedure with Bonferroni correction which is of the form $2\Phi^{-1}(1-\alpha/(p(p-1)))/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

5.4.2 Illustration

We first start with generating a dataset from the covariate adjusted Gaussian graphical model. To this end, we followed the simulation setup of Yin and Li (2013) for generating sparse matrices $\mathbf{\Gamma}$ and $\mathbf{\Theta}$.

First, to generate the precision matrix $\mathbf{\Theta}$, we randomly generated a link between all pairs $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, with probability of connection of 0.01. Then, the corresponding entry in the precision matrix is generated uniformly from $[-1,-0.5] \bigcup [0.5,1]$, for each link. After that, for each row, each entry except the diagonal one is divided by the sum of the absolute values of the off-diagonal entries multiplied by 1.5. Finally, the matrix is symmetrized and the diagonal entries are fixed to 1. To generate the regression coefficient matrix, we first generate a sparse indicator matrix with non-zero elements having a probability of 0.01 of occurring for every pair $(a,b); a = 1,\ldots,q, b = 1,\ldots,p$. Then, corresponding to the non-zero entries of this indicator matrix, we generate uniformly the entries of $\mathbf{\Gamma}$ from $[-1,-\nu_m] \bigcup [\nu_m,1]$, where ν_m is the minimum absolute non-zero value of the generated precision matrix.

To generate covariates $\dot{\mathbf{X}}$ from a multivariate Gaussian distribution, we first generate the covariance matrix $\mathbf{Q}$ from the Toeplitz structure $\varrho^{|a-b|}$, $a,b \in \{1,\ldots,q\}$ with $\varrho = 0.9$. The function `GenerateModel()` implements the generation of $\mathbf{\Theta}$, $\mathbf{\Gamma}$ and $\mathbf{Q}$ in `R` as follows.

```
R> p = 300
R> q = 150
R> sparse_F =  0.01
R> sparse_S = 0.01
R> rho = 0.9

R> set.seed(123)
R> obj1 = GenerateModel(p = p, q = q, sparse_F = sparse_F,
+                        sparse_S = sparse_S, rho = rho)

R> mysigma = obj1$Sig
R> mygamma = obj1$F
R> mysigmaX = obj1$Sigx
```

The arguments `sparse_F` and `sparse_S` are the sparsity parameters of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively, and the argument `rho` is the Toeplitz parameter of $\mathbf{Q}$. To generate a dataset, we first generate $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top$ from a q-dimensional Gaussian distribution with mean vector zero and covariance matrix $\mathbf{Q}$. Finally, conditionally on $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, we generate $\dot{\mathbf{Y}}$ from a p-dimensional Gaussian distribution with mean $\mathbf{\Gamma}^\top \dot{\mathbf{x}}$ and covariance matrix $\mathbf{\Theta}^{-1}$. The function `GenerateData()` implements this data generation in `R`.

```
R> n = 10000
R> data = GenerateData(n = n, p = p, q = q, Sig = mysigma,
+                       Sigx = mysigmaX, F = mygamma)

R> Y = data$y
R> X = data$x
R> colnames(X) <- c(1:q)
R> colnames(Y) <- c(1:p)
R> dim(X)
[1] 10000   150
R> dim(Y)
[1] 10000   300
```

The output of this function is a list of two matrices `Y` and `X` of dimensions $n \times p$ and $n \times q$, respectively.

To perform estimation and inference in the distributed setting, one needs the sub-sample sizes for K sub-samples. The number of sub-samples is set to $K = 10$ in this example and the splitting setting is considered the same as in Section 5.2.2. We provide the code here again for the reader's convenience.

```
R> K = 10

R> size1=floor(((55/100)-(K/100))*n)
```

```
R>   aa=(55/100)-(K/100)
R>   SubSize=c()
R>   if ( K==1 ) {
+     SubSize=n
+   } else if ( K==2 ) {
+     SubSize[1]=size1
+     SubSize[2]=n-SubSize[1]
+   } else if ( K==3 ) {
+     set.seed(2020)
+     SubSize[1]=size1
+     SubSize[2]=floor( ((60/100)-(K/100))*(1-aa)*n )
+     SubSize[3]=n-sum(SubSize[1],SubSize[2])
+   } else {
+     SubSize[1]=size1
+     SubSize[2] = floor( ((60/100)-(K/100))*(1-aa)*n  )
+     SubSize[3:(K-1)] = floor((n-(SubSize[1]+SubSize[2]))/(K-2))
+     SubSize[K]=n-sum(SubSize[1:(K-1)])
+   }
```

In what follows, the regularization parameters are set to $\rho_k = \rho = \sqrt{\log(pq)/n}$, $\lambda_k = \lambda = \sqrt{\log(p)/n}$, and $\tilde{\rho}_{j,k} = \tilde{\rho} = \sqrt{\log(q)/n}$ in the optimization problems (5.4.1), (5.4.2), and the nodewise Lasso regression, respectively, for all K subsamples, and the significance level α is set to $\alpha = 0.05$.

```
R> lambda = rep(sqrt(log(p)/n), K)
R> rho = rep(sqrt((log(p)+log(q))/n), K)
R> rhotilde = rep(sqrt(log(q)/n), K)

R> obj2 = owAvgCondGamma(Y = Y, X = X, K = K, rho = rho, lambda =
+                        lambda, rhotilde = rhotilde, SubSize =
+                        SubSize, alpha)

R> obj2$GammaTilde[1:4, 1:4]
              [,1]          [,2]         [,3]         [,4]
[1,]  0.0076616357 -0.014994208 -0.006984971  0.001241561
[2,] -0.0007496598  0.010508778 -0.013141730 -0.019238793
[3,] -0.0488435493  0.007391558  0.007968575 -0.001090420
[4,]  0.0124639815 -0.016727065  0.000837673 -0.003272885

R> obj2$Lower[1:4, 1:4]
            [,1]        [,2]        [,3]        [,4]
[1,] -0.04556247 -0.06676384 -0.05844017 -0.04971139
[2,] -0.07185692 -0.05868826 -0.08194074 -0.08732383
[3,] -0.11951562 -0.06134597 -0.06035818 -0.06872920
```

```
[4,] -0.05762693 -0.08489587 -0.06690242 -0.07035413

R> obj2$Upper[1:4, 1:4]
           [,1]       [,2]       [,3]       [,4]
[1,] 0.06088574 0.03677542 0.04447022 0.05219451
[2,] 0.07035760 0.07970581 0.05565728 0.04884625
[3,] 0.02182852 0.07612909 0.07629533 0.06654836
[4,] 0.08255490 0.05144174 0.06857776 0.06380836

R> obj2$Len[1:4, 1:4]
          [,1]      [,2]      [,3]      [,4]
[1,] 0.1064482 0.1035393 0.1029104 0.1019059
[2,] 0.1422145 0.1383941 0.1375980 0.1361701
[3,] 0.1413441 0.1374751 0.1366535 0.1352776
[4,] 0.1401818 0.1363376 0.1354802 0.1341625
```

Finally, to estimate the precision matrix $\mathbf{\Theta}$, we use the following function in R.

```
R> obj3 = owAvgCondTheta(Y = Y, X = X, K = K, rho = rho, lambda =
+                        lambda, SubSize = SubSize, alpha)

R> obj3$ThetaTilde[1:4, 1:4]
             [,1]         [,2]         [,3]        [,4]
[1,]  1.129852087  0.010842049 -0.025321714 0.001521015
[2,]  0.010842049  1.137501044 -0.001379203 0.002711373
[3,] -0.025321714 -0.001379203  1.180100515 0.003849879
[4,]  0.001521015  0.002711373  0.003849879 1.148853884

R> obj3$Lower[1:4, 1:4]
             [,1]         [,2]        [,3]        [,4]
[1,]  1.101568437 -0.009267171 -0.04565525 -0.01867077
[2,] -0.009267171  1.108947615 -0.02180696 -0.01758195
[3,] -0.045655250 -0.021806956  1.15091042 -0.01666970
[4,] -0.018670770 -0.017581946 -0.01666970  1.12004012

R> obj3$Upper[1:4, 1:4]
             [,1]       [,2]         [,3]       [,4]
[1,]  1.158135737 0.03095127 -0.004988178 0.02171280
[2,]  0.030951270 1.16605447  0.019048550 0.02300469
[3,] -0.004988178 0.01904855  1.209290612 0.02436946
[4,]  0.021712799 0.02300469  0.024369457 1.17766765

R> obj3$Len[1:4, 1:4]
           [,1]       [,2]       [,3]       [,4]
```

```
[1,] 0.05656730 0.04021844 0.04066707 0.04038357
[2,] 0.04021844 0.05710686 0.04085551 0.04058664
[3,] 0.04066707 0.04085551 0.05838019 0.04103916
[4,] 0.04038357 0.04058664 0.04103916 0.05762753
```

5.5 Additional functions

The package contains several additional functions that serve for the estimation and inference using the weighted and simple average methods. As the usage of these functions is similar to those introduced in the previous sections, we do not provide detailed examples here.

5.5.1 Implementation of the weighted average estimator for Gaussian graphical models

The function `wAvg()` in `R` implements the weighted average estimator

$$\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\mathbf{\Theta}}^d_{ab,k},$$

for every $(a, b) \in \mathcal{V} \times \mathcal{V}$. The output of this function can be obtained by calling

`wAvg(data, K, lambda, SubSize, alpha).`

The arguments of this function are the same as those of `owAvg()`, and they are provided in Section 5.2.2. The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$ containing the final aggregated estimates for $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.
- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k}$ needed for performing inference for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.
- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.
- `Lower`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1 - \alpha$ based on the wAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} - \Phi^{-1}(1 - \alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})/(nK)}$, for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ based on the wAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} + \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})/(nK)}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the wAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the wAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the wAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the wAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

5.5.2 Implementation of the simple average estimator for Gaussian graphical models

The function `sAvg()` in `R` implements the simple average estimator

$$\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{\Theta}}^{d}_{ab,k},$$

for every $(a,b) \in \mathcal{V} \times \mathcal{V}$. The output of this function can be obtained by calling

```
sAvg(data, K, lambda, SubSize, alpha).
```

The arguments of this function are the same as those of `owAvg()` and `wAvg()`. The reader can refer to Section 5.2.2 for more information. The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\tilde{\mathbf{\Theta}}_{\text{sAvg}}$ containing the final aggregated estimates for $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k}$ needed for performing inference for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Lower`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} - \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)/K^3}$, for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} + \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)/K^3}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

5.5.3 Implementation of the weighted average estimator for the coefficient matrix for covariate adjusted graphical models

The function `wAvgCondGamma()` implements the wAvg estimator

$$\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a = \sum_{k=1}^{K} \frac{n_k}{n} \text{vec}(\hat{\mathbf{\Gamma}}^d_k)_a$$

of $\text{vec}(\mathbf{\Gamma})_a$, $a = 1, \ldots, qp$, in `R` and the output of this function can be obtained by calling

```
wAvgCondGamma(Y, X, K, rho, lambda, rhotilde, SubSize, alpha).
```

The arguments of this function are the same as those of `owAvgConGamma()`. The reader can refer to Section 5.4.2 for more information. The output of this function is a list containing

- `GammaTilde`: The $q \times p$ dimensional matrix $\tilde{\mathbf{\Gamma}}_{\text{wAvg}}$ containing the final aggregated estimates for $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ for every element $a = 1, \ldots, qp$.

- `SigmaHat`: A $q \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$, where $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k$, needed for performing inference for every element $a = 1, \ldots, qp$.

- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Lower`: A $q \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ based on the wAvg procedure, which is of the form $\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{wAvg}})_a - \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})/(nK)}$, for every element $a = 1, \ldots, qp$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $q \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ based on the wAvg procedure, which is of the form $\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{wAvg}})_a + \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})/(nK)}$ for every element $a = 1, \ldots, qp$.

- `Len`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the wAvg procedure for every element $a = 1, \ldots, qp$.

- `LowerB`: A $q \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1 - \alpha$ with Bonferroni correction based on the wAvg procedure for every element $a = 1, \ldots, qp$.

- `UpperB`: A $q \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1 - \alpha$ with Bonferroni correction based on the wAvg procedure for every element $a = 1, \ldots, qp$.

- `LenB`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the wAvg procedure for every element $a = 1, \ldots, qp$.

5.5.4 Implementation of the weighted average estimator for the precision matrix for covariate adjusted graphical models

The function `wAvgCondTheta()` implements the wAvg estimator

$$\tilde{\mathbf{\Theta}}_{ab,\mathrm{wAvg}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\mathbf{\Theta}}^d_{ab,k}$$

of $\mathbf{\Theta}_{ab}$, $(a, b) \in \mathcal{V} \times \mathcal{V}$, in `R` and the output of this function can be obtained by calling

`wAvgCondTheta(Y, X, K, rho, lambda, SubSize, alpha).`

The arguments of this function are the same as those of `wAvgCondTheta()`, and they are not provided here. The reader can refer to Section 5.4.2 for more information. The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\mathbf{\Theta}_{\text{wAvg}}$ containing the final aggregated estimates for $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k}$, where $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k}$, needed for performing inference for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Lower`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1 - \alpha$ based on the wAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} - \Phi^{-1}(1 - \alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})/(nK)}$, for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1 - \alpha$ based on the wAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} + \Phi^{-1}(1 - \alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k})/(nK)}$ for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the wAvg procedure for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1 - \alpha$ with Bonferroni correction based on the wAvg procedure for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1 - \alpha$ with Bonferroni correction based on the wAvg procedure for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the wAvg procedure for every pair $(a, b) \in \mathcal{V} \times \mathcal{V}$.

5.5.5 Implementation of the simple average estimator for the coefficient matrix for covariate adjusted graphical models

The function `sAvgCondGamma()` implements the sAvg estimator

$$\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a = \frac{1}{K}\sum_{k=1}^{K}\text{vec}(\hat{\mathbf{\Gamma}}_k^d)_a$$

of $\text{vec}(\mathbf{\Gamma})_a$, $a = 1, \ldots, qp$, in `R` and the output of this function can be obtained by calling

```
sAvgCondGamma(Y, X, K, rho, lambda, rhotilde, SubSize, alpha).
```

The arguments of this function are the same as those of `owAvgCondGamma()`. The reader can refer to Section 5.4.2 for more information. The output of this function is a list containing

- `GammaTilde`: The $q \times p$ dimensional matrix $\tilde{\mathbf{\Gamma}}_{\text{sAvg}}$ containing the final aggregated estimates for $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a$ for every element $a = 1, \ldots, qp$.

- `SigmaHat`: A $q \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$, where $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k$, needed for performing inference for every element $a = 1, \ldots, qp$.

- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.

- `Lower`: A $q \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a - \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^{K}1/n_k)/K^3}$, for every element $a = 1, \ldots, qp$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $q \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a + \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^{K}1/n_k)/K^3}$ for every element $a = 1, \ldots, qp$.

- `Len`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the sAvg procedure for every element $a = 1, \ldots, qp$.

- `LowerB`: A $q \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every element $a = 1, \ldots, qp$.
- `UpperB`: A $q \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every element $a = 1, \ldots, qp$.
- `LenB`: A $q \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the sAvg procedure for every element $a = 1, \ldots, qp$.

5.5.6 Implementation of the simple average estimator for the precision matrix for covariate adjusted graphical models

The function `sAvgCondTheta()` implements the sAvg estimator

$$\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{\Theta}}^{d}_{ab,k}$$

of $\mathbf{\Theta}_{ab}$, $(a,b) \in \mathcal{V} \times \mathcal{V}$, in `R` and the output of this function can be obtained by calling

```
sAvgCondTheta(Y, X, K, rho, lambda, SubSize, alpha).
```

The arguments of this function are the same as those of `owAvgCondTheta()`. The reader can refer to Section 5.4.2 for more information. The output of this function is a list containing

- `ThetaTilde`: The $p \times p$ dimensional matrix $\mathbf{\Theta}_{\text{sAvg}}$ containing the final aggregated estimates for $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `SigmaHat`: A $p \times p$ dimensional matrix containing the sum of variance estimates in the sub-samples of the form $\sum_{k=1}^{K} \hat{\sigma}^2_{ab,k}$, where $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k}$, needed for performing inference for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `RTime`: Total running time obtained as the sum of the maximum time among all parallel jobs and the time to combine the results.
- `Lower`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} - \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K} 1/n_k)/K^3}$, for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

- `Upper`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ based on the sAvg procedure, which is of the form $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} + \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K}\hat{\sigma}^2_{ab,k})(\sum_{k=1}^{K}1/n_k)/K^3}$ for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `Len`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `LowerB`: A $p \times p$ dimensional matrix containing the lower bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `UpperB`: A $p \times p$ dimensional matrix containing the upper bound of the asymptotic confidence interval at level $1-\alpha$ with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.
- `LenB`: A $p \times p$ dimensional matrix containing the length of the asymptotic confidence interval with Bonferroni correction based on the sAvg procedure for every pair $(a,b) \in \mathcal{V} \times \mathcal{V}$.

5.6 Discussion

In this chapter, we showcased the usage of the **DistributedGGL** package, which performs estimation and inference for Gaussian graphical models under a distributed setting. This is done by using the functions `owAvg()` and `owAvgCondTheta()`, respectively. Moreover, estimation of the coefficient matrix for covariate adjusted Gaussian graphical models is performed with the function `owAvgCondGamma()`. Performing hypothesis testing for the entries of the precision matrix is also performed with the functions `Distrib.dGL()` and `FDR()`. The advantage of this package is the simplicity of using and implementing the theoretical part of papers Nezakati and Pircalabelu (2023a), Nezakati and Pircalabelu (2023b) and Nezakati and Pircalabelu (2023c) in practice.

CHAPTER 6

Conclusions and extensions

The objective of this thesis was to study Gaussian graphical models from a new perspective: the distributed setting. This divide and conquer paradigm proves to be helpful in addressing contemporary challenges, particularly preserving the security and privacy of datasets. We investigated the asymptotic properties, such as convergence rate, consistency, and asymptotic distribution of the proposed estimators. The context of this thesis can be extended in various directions, some of which we propose here.

(1) The main assumption throughout this thesis was Gaussianity. All the models and estimators were built under the assumption that the dataset comes from a multivariate Gaussian distribution. The most natural extension of this thesis is to explore for cases where data are non-Gaussian. For instance, one can consider the non-paranormal distribution, where the Gaussian random vector $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top$ is replaced by the random vector $\boldsymbol{f}(\dot{\mathbf{Y}}) = (f_1(Y^1), \ldots, f_p(Y^p))^\top$, and $\boldsymbol{f}(\dot{\mathbf{Y}})$ is assumed to be multivariate Gaussian. This model is well studied in Liu et al. (2009), where they proposed a non-parametric estimator for the functions $\{f_j\}$, $j = 1, \ldots, p$, and showed that how the graphical Lasso can be used to estimate the graph in the high-dimensional setting. Combining this model with the distributed setting is more challenging due to the unknown form of $\{f_j\}$, $j = 1, \ldots, p$, and the complicated form of the KKT condition.

(2) As it was observed in this thesis, the graphical Lasso estimator was biased and it produced a lot of false negatives. This method was equipped with one regularization parameter, λ. An alternative approach would be to impose different λ's, not just one, to shrink similar estimators towards each other, and hopefully get less false negatives.

(3) The model in this thesis was built for a Gaussian random vector. However, in

some situations, we deal with a matrix variate Gaussian data where, instead of a p-dimensional vector, we have a $p \times q$ dimensional matrix. Considering the distributed setting from this perspective is also of interest.

(4) The distributed setting in this thesis was constructed by splitting observations into different non-overlapping independent sub-samples. Another extension of this approach can be applied to cases where variables are split into different groups rather than observations. Since there are dependencies between variables, this approach is more challenging and requires greater delicacy.

In this chapter, we delve into some of these extensions in detail as follows.

Section 6.1 explores in detail idea (2), while idea (3) is investigated in Section 6.2. Finally, Section 6.3 provides a detailed explanation of idea (4). The remaining ideas will be investigated in further research.

6.1 Sorted L-One Penalized Estimation (SLOPE)

As an extension of the Lasso penalty, one can consider SLOPE, defined by Bogdan et al. (2015), where the penalty is defined by the sorted ℓ_1 norm (SL1)

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \sum_{i=1}^{p} \lambda_i |\beta|_{(i)},$$

used in a regression model, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of model parameters, whose absolute values are sorted in descending order as $|\beta|_{(1)} \geq \ldots \geq |\beta|_{(p)}$, whereas $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ is the sequence of corresponding tuning parameters, which satisfy the condition $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. The SLOPE penalty is based on a decaying sequence of tuning parameters, which allows for assigning exactly the same estimated regression coefficients to the groups of variables with a similar influence on the loss function. In general, SLOPE exhibits two levels of shrinkage of regression coefficients: (1) shrinking towards zero; and (2) shrinking the similar estimates towards each other (Bogdan et al., 2015; Bellec et al., 2018).

In connection with graphical models, Riccobello et al. (2022) proposed SLOPE for the Gaussian graphical models (Gslope) by first vectorizing the elements on the upper triangle of $\boldsymbol{\Theta}$, to create a new vector $\boldsymbol{\Theta}^\star = (\boldsymbol{\Theta}_1^\star, \ldots, \boldsymbol{\Theta}_m^\star)^\top = (\boldsymbol{\Theta}_{12}, \ldots, \boldsymbol{\Theta}_{1p}, \boldsymbol{\Theta}_{23}, \ldots, \boldsymbol{\Theta}_{2p}, \ldots, \boldsymbol{\Theta}_{(p-1)p})^\top$ of length m, where $m = p(p-1)/2$. Then, the authors defined the following SL1 penalty to create

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\Theta}) = \sum_{i=1}^{m} \lambda_i |\boldsymbol{\Theta}^\star|_{(i)},$$

where $|\boldsymbol{\Theta}^\star|_{(i)}$, $i = 1, \ldots, m$, are the absolute values of $\boldsymbol{\Theta}^\star$ entries in descending order, and which assigns to each of them a specific tuning parameter, such that

$\lambda_1 \geq \ldots \geq \lambda_m \geq 0$. By considering the above penalty function, Gslope solves the following optimization problem

$$\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} = \arg\min_{\boldsymbol{\Theta} \in S^p_{++}} \{\mathrm{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\Theta})\}, \tag{6.1.1}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. Similarly to graphical Lasso, Gslope is a convex optimization problem, which has a unique solution. However, the optimal result is biased. To debias this estimator, we propose to use a similar technique to Jankova and van de Geer (2015). To this end, rewrite the optimization problem (6.1.1) as

$$\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} = \arg\min_{\boldsymbol{\Theta} \in S^p_{++}} \{\mathrm{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \sum_{a \neq b} \Lambda_{ab}|\boldsymbol{\Theta}_{ab}|\}, \tag{6.1.2}$$

where

$$\Lambda_{ab} = \begin{cases} \lambda_{\sigma(a,b)}; & a < b, \\ \lambda_{\sigma(b,a)}; & a > b, \end{cases}$$

and $\sigma(a,b)$ is the order of $\boldsymbol{\Theta}_{ab}$ in the upper triangle of $\boldsymbol{\Theta}$. For example in a precision matrix with $p = 3$, suppose that $|\boldsymbol{\Theta}_{13}| > |\boldsymbol{\Theta}_{23}| > |\boldsymbol{\Theta}_{12}|$. Then, $\sigma(1,2) = 3$, $\sigma(1,3) = 1$, $\sigma(2,3) = 2$, and as a result, $\Lambda_{12} = \lambda_3$, $\Lambda_{13} = \lambda_1$, and $\lambda_{23} = \lambda_2$. Note that, by the structure of SLOPE, $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

We propose to use next the KKT condition of the optimization problem (6.1.2), which is characterized as

$$\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Theta}}^{-1}_{\mathrm{GS}} + \boldsymbol{\Lambda} \odot \hat{\mathbf{D}} = \mathbf{0}, \tag{6.1.3}$$

where $\boldsymbol{\Lambda}$ is a $p \times p$ dimensional matrix with entries Λ_{ab}, the symbol $\odot$ is the Hadamard product (elementwise product) between two matrices, and $\hat{\mathbf{D}}$ is a matrix which belongs to the sub-differential of the ℓ_1 off-diagonal norm evaluated at $\hat{\boldsymbol{\Theta}}_{\mathrm{GS}}$. Denote $\mathbf{W} := \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$. By inverting (6.1.3), and performing some algebra, we get

$$\hat{\boldsymbol{\Theta}}^d_{\mathrm{GS}} - \boldsymbol{\Theta} = -\boldsymbol{\Theta}\mathbf{W}\boldsymbol{\Theta} + \boldsymbol{\Delta},$$

where

$$\hat{\boldsymbol{\Theta}}^d_{\mathrm{GS}} := \hat{\boldsymbol{\Theta}}_{\mathrm{GS}} + \hat{\boldsymbol{\Theta}}_{\mathrm{GS}}(\boldsymbol{\Lambda} \odot \hat{\mathbf{D}}) = 2\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} - \hat{\boldsymbol{\Theta}}_{\mathrm{GS}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}_{\mathrm{GS}},$$
$$\boldsymbol{\Delta} := -(\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} - \boldsymbol{\Theta})\mathbf{W}\boldsymbol{\Theta} - (\hat{\boldsymbol{\Theta}}_{\mathrm{GS}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}_p)(\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} - \boldsymbol{\Theta}).$$

To show the asymptotic normality of $\hat{\boldsymbol{\Theta}}^d_{\mathrm{GS}}$, we need to show that the remainder term $\boldsymbol{\Delta}$ is negligible, and $\boldsymbol{\Theta}\mathbf{W}\boldsymbol{\Theta} \xrightarrow{d} \mathcal{N}(0,1)$. To this end, we need to bound $\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} - \boldsymbol{\Theta}$. By following the proof of Theorem 1 of Ravikumar et al. (2011), under some regularity assumptions, it can be shown that

$$\|\hat{\boldsymbol{\Theta}}_{\mathrm{GS}} - \boldsymbol{\Theta}\|_\infty \leq 2(1 + 8/\alpha_1)\kappa_{\mathbf{H}}\bar{\delta}_f(n, p^\tau),$$

where α_1 and $\kappa_{\mathbf{H}}$ are defined in Chapter 2, and

$$\bar{\delta}_f(n, p^\tau) := \arg\max\{\delta : f(n, \delta) \leq p^\tau\},$$

for some $\tau > 2$, and $f(n, \delta) : \mathbb{N} \times (0, \infty) \to (0, \infty)$ is a tail function such that for every $(a, b) \in \mathcal{V} \times \mathcal{V}$,

$$\mathbb{P}(|\hat{\mathbf{\Sigma}}_{ab} - \mathbf{\Sigma}_{ab}| \geq \delta) \leq 1/f(n, \delta), \quad \text{for all } \delta \in (0, \infty).$$

The tail function f can have different forms. In the case of Gaussian graphical models, when $\dot{\mathbf{Y}}$ is a multivariate Gaussian vector, the deviation of the sample covariance matrix from the target has an exponential-type tail function of the form $f(n, \delta) = \exp(cn\delta^2)$ for some constant $c > 0$. After some calculations, it can be shown that in this case $\bar{\delta}_f(n, p^\tau) = \sqrt{\log(p^\tau)/(cn)}$.

The difference of this result with the one of Ravikumar et al. (2011) is that we need more conditions on the regularization parameter λ_i, $i = 1, \ldots, m$. The regularization parameter λ_i should be of the form $\lambda_i = (8/\alpha_1)\bar{\delta}_f(n, p^\tau/i)$ for some $\tau > 2$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$, and moreover, $1 \leq \frac{\lambda_1}{\lambda_m} < \frac{4}{(1-\alpha_1)(4+\alpha_1)+\alpha_1}$. The negligibility of the remainder term $\mathbf{\Delta}$ is left for future research.

By showing the asymptotic normality of the debiased Gslope estimator $\hat{\mathbf{\Theta}}^d_{\mathrm{GS}}$, one can apply the same procedure as in Chapter 2 in the distributed setting. By splitting the dataset into K non-overlapping sub-samples, performing debiasing procedure on each sub-sample, one can benefit from the asymptotic normality, and by maximizing a pseudo log-likelihood function constructed with the asymptotic density of the debiased estimators, the aggregated estimator could be defined and further investigated.

6.2 Matrix variate graphical models

As an extension of the multivariate Gaussian distribution for vector data, one can consider the matrix variate normal distribution for a matrix variable $\mathbf{X} \in \mathbb{R}^{p \times q}$. We assume that the matrix variable $\mathbf{X} \in \mathbb{R}^{p \times q}$ follows a matrix variate Gaussian distribution with probability density function

$$f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{\Sigma}, \mathbf{\Psi}) = (2\pi)^{-qp/2}\det(\mathbf{\Sigma})^{-q/2}\det(\mathbf{\Psi})^{-p/2}\exp\left\{ -\frac{1}{2}\mathrm{trace}(\mathbf{x}\mathbf{\Psi}^{-1}\mathbf{x}^\top\mathbf{\Sigma}^{-1})\right\}, \tag{6.2.1}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ and $\mathbf{\Psi} \in \mathbb{R}^{q \times q}$ are the row and column variance matrices. Such a model is frequently encountered in many applications such as financial market analysis, genomic data analysis, and brain imaging studies. The Matrix variate Gaussian distribution is useful for characterizing conditional independence between the variables (Dawid, 1981; Gupta and Nagar, 2018). Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{ij})$ and $\mathbf{\Gamma} = \mathbf{\Psi}^{-1} = (\gamma_{ij})$ be the precision matrices for the row and the column vectors,

respectively. Formally, in a matrix normal distribution, two arbitrary entries X_{ij} and X_{kl} are conditionally independent given the remaining entries if and only if: (1) at least one of the entries ω_{ij} and γ_{kl} when $i \neq k, j \neq l$ is zero; (2) $\omega_{ik} = 0$ when $i \neq k, j = l$; (3) $\gamma_{jl} = 0$ when $i = k, j \neq l$. The partial correlation between variables X_{ij} and X_{kl} is then defined as

$$\rho_{ij,kl} = -\frac{\omega_{ik}}{\sqrt{\omega_{ii}\omega_{kk}}} \times \frac{\gamma_{jl}}{\sqrt{\gamma_{jj}\gamma_{ll}}}.$$

The partial correlation $\rho_{ij,kl}$ is not zero only if both ω_{ik} and γ_{jl} are non-zero.

To build sparse estimators for $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, consider an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the matrix variable $\mathbf{X}$. One can implement the penalization problem of Leng and Tang (2012) as

$$\min_{\mathbf{\Omega},\mathbf{\Gamma}} \left\{ \frac{1}{npq} \sum_{i=1}^{n} \text{trace}(\mathbf{X}_i \mathbf{\Gamma} \mathbf{X}_i^\top \mathbf{\Omega}) - \frac{1}{p} \log\det(\mathbf{\Omega}) - \frac{1}{q} \log\det(\mathbf{\Gamma}) + p_{\lambda_1}(\mathbf{\Omega}) + p_{\lambda_2}(\mathbf{\Gamma}) \right\}, \tag{6.2.2}$$

where $p_\lambda(\mathbf{A}) = \sum_{i \neq j} p_\lambda(|a_{ij}|)$ for a square matrix $\mathbf{A} = (a_{ij})$ with a suitably defined penalty function p_λ, such as Lasso, defined as $p_\lambda(s) = \lambda|s|$. As it was observed in this thesis, Lasso produces bias estimators. To combine distributed approach with matrix variate Gaussian graphical models, one first needs to debias the estimators in the sub-samples, and then by using the asymptotic normality of the debiased estimators, one constructs the aggregated estimators. The optimization problem (6.2.2) with the Lasso penalty is a non-convex problem as the trace produces interaction terms between the two precision matrices. However, it is biconvex as convexity holds only with respect to either precision matrix individually but not jointly. As such, by fixing one precision matrix, say $\mathbf{\Gamma}$, the optimization problem (6.2.2) reduces to a convex graphical Lasso problem on $\mathbf{\Omega}$, as

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega} \in S_{++}^p} \left\{ \frac{1}{p} \text{trace}(\hat{\mathbf{\Sigma}}_1 \mathbf{\Omega}) - \frac{1}{p} \log\det(\mathbf{\Omega}) + \lambda_1 \|\mathbf{\Omega}\|_{1,\text{off}} \right\},$$

$\hat{\mathbf{\Sigma}}_1 = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{\Gamma} \mathbf{X}_i^\top / p$, and the KKT condition is of the form

$$\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Omega}} + \lambda_1 \hat{\mathbf{D}}_1 = \mathbf{0}, \tag{6.2.3}$$

where $\hat{\mathbf{D}}_1$ is a matrix which belongs to the sub-differential of the Lasso penalty evaluated at $\hat{\mathbf{\Omega}}$. Similarly, by fixing $\mathbf{\Omega}$, the optimization problem (6.2.2) reduces to the convex problem

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma} \in S_{++}^q} \left\{ \frac{1}{q} \text{trace}(\mathbf{\Gamma} \hat{\mathbf{\Sigma}}_2) - \frac{1}{q} \log\det(\mathbf{\Gamma}) + \lambda_2 \|\mathbf{\Gamma}\|_{1,\text{off}} \right\},$$

on $\boldsymbol{\Gamma}$, where $\hat{\boldsymbol{\Sigma}}_2 = \sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{\Omega} \mathbf{X}_i^\top / q$, and the KKT condition is of the form

$$\hat{\boldsymbol{\Sigma}}_2 - \hat{\boldsymbol{\Gamma}} + \lambda_2 \hat{\mathbf{D}}_2 = \mathbf{0}, \tag{6.2.4}$$

where $\hat{\mathbf{D}}_2$ is a matrix which belongs to the sub-differential of the Lasso penalty evaluated at $\hat{\boldsymbol{\Gamma}}$. By solving the KKT conditions (6.2.3) and (6.2.4), and performing some algebra, one can construct the debiased estimators for $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$. This investigation can be a challenging problem and is left for future work.

6.3 Feature splitting via a decorrelated procedure

The distributed framework in this thesis was based on splitting sample data into different, non-overlapping independent sub-samples. When $p > n$, splitting on the variable space is more effective. However, when the variables are correlated, by splitting on the variable space one might miss important information, simply due to the fact that not all variables are present in the model at the same time. Wang et al. (2016) proposed a decorrelation approach (DECO) in penalized regression models, where they first decorrelated variables by a transformation based on singular value decomposition (SVD), and then they partitioned the space of variables onto different sub-groups.

In view of graphical models, one can consider nodewise graphical Lasso (Meinshausen and Bühlmann, 2006) in connection with a DECO approach. Throughout this thesis, we considered estimation of the precision matrix using the graphical Lasso method which is based on regularization of the maximum likelihood in terms of the ℓ_1 penalty. Another common approach to precision matrix estimation is based on projections. This approach reduces the problem to a series of regression problems and estimates each column of the precision matrix using a Lasso estimator or the Dantzig selector (Candès and Tao, 2007). The idea was first introduced by Meinshausen and Bühlmann (2006) as neighborhood selection for Gaussian graphical models and further studied in Yuan (2010), Cai et al. (2011) and Sun and Zhang (2013) among many others.

In the framework of nodewise graphical Lasso, consider the linear regression model

$$\mathbf{X}_j = \mathbf{X}_{-j} \boldsymbol{\beta} + \boldsymbol{\epsilon}_j, \tag{6.3.1}$$

for each $j = 1, \ldots, p$, where $\mathbf{X}_j$, of length n, and $\mathbf{X}_{-j}$, of dimension $n \times (p-1)$, are the j-th column and the sub-matrix of $\mathbf{X}_{n \times p}$ obtained by removing the j-th column, respectively. Moreover, $\boldsymbol{\epsilon}_j$ is the noise vector of length n. To decorrelate the columns of $\mathbf{X}_{-j}$, one can implement the method of Wang et al. (2016) as follows.

When $p \leq n$, the most intuitive way is to orthogonalize variables via the singular value decomposition (SVD) of the design matrix $\mathbf{X}_{-j} = \mathbf{U}_j \mathbf{D}_j \mathbf{V}_j^\top$, where $\mathbf{U}_j$ is

an $n \times (p-1)$ matrix, $\mathbf{D}_j$ is an $(p-1) \times (p-1)$ diagonal matrix and $\mathbf{V}_j$ is an $(p-1) \times (p-1)$ orthogonal matrix. By multiplying both sides of (6.3.1) by $\sqrt{p-1}\mathbf{D}_j^{-1}\mathbf{U}_j^\top$, we get

$$\sqrt{p-1}\mathbf{D}_j^{-1}\mathbf{U}_j^\top \mathbf{X}_j = \sqrt{p-1}\mathbf{V}_j^\top \boldsymbol{\beta} + \sqrt{p-1}\mathbf{D}_j^{-1}\mathbf{U}_j^\top \boldsymbol{\epsilon}_j.$$

Now the new features (the columns of $\sqrt{p-1}\mathbf{V}_j^\top$) are mutually orthogonal. An alternative approach to avoid doing SVD is to replace $\sqrt{p-1}\mathbf{D}_j^{-1}\mathbf{U}_j^\top$ by $\sqrt{p-1}\mathbf{U}_j\mathbf{D}_j^{-1}\mathbf{U}_j^\top = (\mathbf{X}_{-j}\mathbf{X}_{-j}^\top/(p-1))^{+/2}$, where $\mathbf{A}^+$ denotes the Moore-Penrose pseudo-inverse. Then, we have

$$(\mathbf{X}_{-j}\mathbf{X}_{-j}^\top/(p-1))^{+/2}\mathbf{X}_j = \sqrt{p-1}\mathbf{U}_j\mathbf{V}_j^\top \boldsymbol{\beta} + (\mathbf{X}_{-j}\mathbf{X}_{-j}^\top/(p-1))^{+/2}\boldsymbol{\epsilon}_j, \quad (6.3.2)$$

where the new features (the columns of $\sqrt{p-1}\mathbf{U}_j\mathbf{V}_j^\top$) are mutually orthogonal. By defining the new data as $\tilde{\mathbf{X}}_j = (\mathbf{X}_{-j}\mathbf{X}_{-j}^\top/(p-1))^{+/2}\mathbf{X}_j$ and $\tilde{\mathbf{X}}_{-j} = \sqrt{p-1}\mathbf{U}_j\mathbf{V}_j^\top$, the mutually orthogonal property allows to decompose $\tilde{\mathbf{X}}_{-j}$ column-wisely to K subsets $\tilde{\mathbf{X}}_{-j}^{(k)}$, $k = 1, \ldots, K$.

When $p > n$, the SVD decomposition on $\mathbf{X}_{-j}$ induces a different form on the three matrices, i.e., $\mathbf{U}_j$ is now an $n \times n$ orthogonal matrix, $\mathbf{D}_j$ is an $n \times n$ diagonal matrix, and $\mathbf{V}_j$ is an $n \times (p-1)$ matrix.

Now for each $j = 1, \ldots, p$, consider the transformed data $(\tilde{\mathbf{X}}_j, \tilde{\mathbf{X}}_{-j}^{(k)})$, such that $\tilde{\mathbf{X}}_j$ is of length n and $\tilde{\mathbf{X}}_{-j}^{(k)}$ is of dimension $n \times p_k$, for $k = 1, \ldots, K$, such that $\sum_{k=1}^K p_k = p-1$. Define the estimators of the regression coefficients, $\hat{\boldsymbol{\beta}}_j^k = \{\hat{\beta}_{j,j'}; j' = 1, \ldots, p_k, j' \neq j\} \in \mathbb{R}^{p_k}$, as follows

$$\hat{\boldsymbol{\beta}}_j^k := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p_k}} \|\tilde{\mathbf{X}}_j - \tilde{\mathbf{X}}_{-j}^{(k)}\boldsymbol{\beta}\|_2^2/n + 2\lambda_j\|\boldsymbol{\beta}\|_1. \quad (6.3.3)$$

where λ_j is the regularization parameter for the j-th penalized regression problem. Further, define the column vectors

$$\hat{\Gamma}_j^k := (-\hat{\beta}_{j,1}, \ldots, -\hat{\beta}_{j,j-1}, 1, -\hat{\beta}_{j,j+1}, \ldots, -\hat{\beta}_{j,p_k})^\top,$$

and estimators of the noise level

$$\hat{\tau}_j^k := \|\tilde{\mathbf{X}}_j - \tilde{\mathbf{X}}_{-j}^{(k)}\hat{\boldsymbol{\beta}}_j^k\|_2^2/n + \lambda_j\|\hat{\boldsymbol{\beta}}_j^k\|_1,$$

for $j = 1, \ldots, p$, and $k = 1, \ldots, K$. Finally, define the j-th column of the nodewise Lasso estimator $\hat{\boldsymbol{\Theta}}$ as

$$\hat{\boldsymbol{\Theta}}_j := (\hat{\Gamma}_j^1/\hat{\tau}_j^1, \ldots, \hat{\Gamma}_j^K/\hat{\tau}_j^K).$$

By applying the debiasing procedure of Jankova and van de Geer (2017) for nodewise Lasso regression, one can debias this estimator and benefit from asymptotic normality.

References

Akbani, R., K. C. Akdemir, B. A. Aksoy, M. Albert, A. Ally, S. B. Amin, H. Arachchi, A. Arora, J. T. Auman, B. Ayala, et al. (2015). Genomic classification of cutaneous melanoma. *Cell 161*(7), 1681–1696.

Arroyo, J. and E. Hou (2016). Efficient distributed estimation of inverse covariance matrices. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5. IEEE.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research 9*(3), 485–516.

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085.

Barber, R. F. and E. J. Candès (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics 47*(5), 2504–2537.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics 46*(3), 1352–1382.

Bellec, P. C., G. Lecué, and A. B. Tsybakov (2018). Slope meets Lasso: improved oracle bounds and optimality. *The Annals of Statistics 46*(6B), 3603–3642.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association 70*(350), 428–434.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bogdan, M., E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics 9*(3), 1103–1140.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cai, T., W. Liu, and H. Zhou (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics 44*(2), 455–488.

Cai, T. T., H. Li, W. Liu, and J. Xie (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika 100*(1), 139–156.

Cancer Genome Atlas Research Network (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature 543*(7645), 378–384.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 551–577.

Candès, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Cardoso-Cachopo, A. (2007). Improving methods for single-label text categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Chen, M., Z. Ren, H. Zhao, and H. Zhou (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *Journal of the American Statistical Association 111*(513), 394–406.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*(4), 1655–1684.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting 32*(3), 754–762.

Clarke, L., I. Glendinning, and R. Hempel (1994). The MPI message passing interface standard. In *Programming Environments for Massively Parallel Distributed Systems: Working Conference of the IFIP WG 10.3, April 25–29, 1994*, pp. 213–218. Springer.

Corbett, J. C., J. Dean, M. Epstein, A. Fikes, C. Frost, J. J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, et al. (2013). Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems (TOCS) 31*(3), 1–22.

Csardi, G., T. Nepusz, V. Traag, S. Horvat, F. Zanini, D. Noom, and K. Muller (2005). *igraph: Network Analysis and Visualization*. R package version 1.5.0.1.

Cui, C., J. Jia, Y. Xiao, and H. Zhang (2021). Directional FDR control for sub-Gaussian sparse GLMs. *arXiv preprint arXiv:2105.00393*.

Curtiss, M., I. Becker, T. Bosman, S. Doroshenko, L. Grijincu, T. Jackson, S. Kunnatur, S. Lassen, P. Pronin, S. Sankar, et al. (2013). Unicorn: A system for searching the social graph. *Proceedings of the VLDB Endowment 6*(11), 1150–1161.

d'Aspremont, A., O. Banerjee, and L. El Ghaoui (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications 30*(1), 56–66.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika 68*(1), 265–274.

Dean, J. and S. Ghemawat (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM 51*(1), 107–113.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers 23*(2), 229–236.

Eddelbuettel, D. and J. J. Balamuta (2018). Extending *R* with *C++*: A brief introduction to *Rcpp*. *The American Statistician 72*(1), 28–36.

Fan, J. and J. Chen (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(4), 927–943.

Fan, Y., E. Demirkaya, G. Li, and J. Lv (2019). Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association 115*(529), 362—-379.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics 9*(3), 432–441.

Friedman, J., T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, J. Qian, and J. Yang (2008). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.1-7.

Gelman, A. and F. Tuerlinckx (2000). Type s error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics 15*(3), 373–390.

Genz, A., F. Bretz, T. Miwa, and X. Mi (2023). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.2-2.

Golosnoy, V., B. Gribisch, and M. I. Seifert (2022). Sample and realized minimum variance portfolios: Estimation, statistical inference, and tests. *Wiley Interdisciplinary Reviews: Computational Statistics 14*(5), 1–18.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Gupta, A. K. and D. K. Nagar (2018). *Matrix variate distributions*. Chapman and Hall/CRC.

Gut, A. (2005). *Probability: a graduate course*, Volume 200. Springer.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the Lasso and generalizations*. CRC press.

He, X., J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. (2014). Practical lessons from predicting clicks on ads at Facebook. In *Proceedings of the eighth international workshop on data mining for online advertising*, pp. 1–9.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Hsieh, C. J., M. A. Sustik, I. S. Dhillon, and P. Ravikumar (2014). Quic: quadratic approximation for sparse inverse covariance estimation. *The Journal of Machine Learning Research 15*(1), 2911–2947.

Huo, X. and S. Cao (2019). Aggregated inference. *Wiley Interdisciplinary Reviews: Computational Statistics 11*(1), e1451.

Jankova, J. and S. van de Geer (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics 9*(1), 1205–1229.

Jankova, J. and S. van de Geer (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test 26*, 143–162.

Javanmard, A. and H. Javadi (2019). False discovery rate control via debiased Lasso. *Electronic Journal of Statistics 13*(1), 1212–1253.

Javanmard, A. and A. Montanari (2013). Model selection for high-dimensional regression under the generalized irrepresentability condition. *Advances in neural information processing systems 26*, 1–9.

Javanmard, A. and A. Montanari (2018). Debiasing the Lasso: Optimal sample size for gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Jia, J., K. Rohe, and B. Yu (2013). The Lasso under Poisson-like heteroscedasticity. *Statistica Sinica 23*(1), 99–118.

Jiang, H., X. Fei, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, X. Li, and T. Zhao (2010). *huge: High-Dimensional Undirected Graph Estimation*. R package version 1.3.5.

Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Kallenberg, O. (1997). *Foundations of modern probability*, Volume 2. Springer.

Kempf, A. and C. Memmel (2006). Estimating the global minimum variance portfolio. *Schmalenbach Business Review 58*, 332–348.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics 37*(6B), 4254–4278.

Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.

Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research 18*(1), 115–144.

Leng, C. and C. Y. Tang (2012). Sparse matrix graphical models. *Journal of the American Statistical Association 107*(499), 1187–1200.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association 00*(0), 1–13. In press.

Liu, D., R. Y. Liu, and M. Xie (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association 110*(509), 326–340.

Liu, H., F. Han, and C.-h. Zhang (2012). Transelliptical graphical models. *Advances in neural information processing systems 25*, 1–9.

Liu, H., J. Lafferty, and L. Wasserman (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research 10*(10), 2295–2328.

Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems 24*(2), 1432–1440.

Liu, J., T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell 173*(2), 400–416.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics 41*(6), 2948–2978.

Loh, P.-L. and X. L. Tan (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics 12*(1), 1429–1467.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR.

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

Meng, X., J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al. (2016). Mllib: Machine learning in Apache Spark. *The Journal of Machine Learning Research 17*(1), 1235–1241.

Microsoft and S. Weston (2009). *foreach: Provides Foreach Looping Construct*. R package version 1.5.2.

Nezakati, E. and E. Pircalabelu (2023a). Directional false discovery rate control via debiased and distributed procedures in Gaussian graphical models. pp. 1–34. Technical report.

Nezakati, E. and E. Pircalabelu (2023b). Estimation and inference in sparse multivariate regression and conditional Gaussian graphical models under an unbalanced distributed setting. pp. 1–50. Technical report.

Nezakati, E. and E. Pircalabelu (2023c). Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models. *Statistics and Computing 33*, 1–14.

Ng, B., G. Varoquaux, J. B. Poline, and B. Thirion (2013). A novel sparse group Gaussian graphical model for functional connectivity estimation. In *Information Processing in Medical Imaging: 23rd International Conference, IPMI 2013, Asilomar, CA, USA, June 28–July 3, 2013. Proceedings 23*, pp. 256–267. Springer.

Obozinski, G., M. J. Wainwright, and M. I. Jordan (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics 39*(1), 1–47.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics 4*(1), 53–77.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research 11*, 2241–2259.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Riccobello, R., M. Bogdan, G. Bonaccolto, P. J. Kremer, S. Paterlini, and P. Sobczyk (2022). Sparse graphical modelling via the sorted ℓ_1-norm. *arXiv preprint arXiv:2204.10403*.

Rothman, A. J. (2012). *MRCE: Multivariate Regression with Covariance Estimation*. R package version 2.4.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4*(1), 1–30.

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research 14*(1), 3385–3418.

Taigman, Y., M. Yang, M. Ranzato, and L. Wolf (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708.

Tang, L., L. Zhou, and P. X. K. Song (2020). Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis 176*, 104567.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory 52*(3), 1030–1051.

Vagata, P. and K. Wilfong (2017). Scaling the Facebook data warehouse to 300 pb. https://engineering.fb.com/2014/04/10/core-data/scaling-the-facebook-data-warehouse-to-300-pb/.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Wang, G. P. and H. J. Cui (2021). Efficient distributed estimation of high-dimensional sparse precision matrix for transelliptical graphical models. *Acta Mathematica Sinica, English Series 37*(5), 689–706.

Wang, H. (2014). Coordinate descent algorithm for covariance graphical Lasso. *Statistics and Computing 24*(4), 521–529.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica 25*(3), 831–851.

Wang, L., X. Ren, and Q. Gu (2016). Precision matrix estimation in high dimensional Gaussian graphical models with faster rates. In *Artificial Intelligence and Statistics*, pp. 177–185.

Wang, X., D. B. Dunson, and C. Leng (2016). Decorrelated feature space partitioning for distributed sparse regression. *Advances in Neural Information Processing Systems 29*, 1–9.

Wang, X., L. Qin, H. Zhang, Y. Zhang, L. Hsu, and P. Wang (2015). A regularized multivariate regression approach for eQTL analysis. *Statistics in Biosciences 7*(1), 129–146.

Xia, Y., T. Cai, and T. T. Cai (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika 102*(2), 247–266.

Xie, M., K. Singh, and W. E. Strawderman (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association 106*(493), 320–333.

Xu, G., Z. Shang, and G. Cheng (2019). Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. *Journal of Computational and Graphical Statistics 28*(4), 891–908.

Xue, J. and F. Liang (2019). Double-parallel monte carlo for bayesian analysis of big data. *Statistics and Computing 29*(1), 23–32.

Yin, J. and H. Li (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*(4), 2630–2650.

Yin, J. and H. Li (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis 116*, 365–381.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhang, T. and H. Zou (2014). Sparse precision matrix estimation via Lasso penalized d-trace loss. *Biometrika 101*(1), 103–120.

Zhang, Y., J. Duchi, and M. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research 16*(1), 3299–3340.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1533.