UNBALANCED DISTRIBUTED ESTIMATION AND INFERENCE FOR THE PRECISION MATRIX IN GAUSSIAN GRAPHICAL MODELS

Ensiyeh Nezakati, Eugen Pircalabelu

REPRINT | 2023 / 16

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

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

Ensiyeh Nezakati[1] and Eugen Pircalabelu[1*]

[1] Institute of Statistics, Biostatistics and Actuarial Sciences, Voie du Roman Pays 20, Louvain-la-Neuve, 1348, Belgium.

*Corresponding author(s). E-mail(s): eugen.pircalabelu@uclouvain.be;
Contributing authors: nezakati.ensiyeh@uclouvain.be;

Abstract

This paper 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 in one single machine. In this paper, 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.

Keywords: Gaussian graphical models, Precision matrix, Lasso penalization, Unbalanced distributed setting, De-biased estimator, Pseudo log-likelihood.

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 paper 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 average. 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 exhibit the efficiency of the model. Recently, Tang et al. [2020] developed a strategy using an aggregating method based on confidence distributions to combine de-biased 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 $\mathcal{O}(n^{1/2-\epsilon})$, $\epsilon \in (0, 1/2]$, the combined estimator achieves the same estimation efficiency as the centralized maximum likelihood estimator. In this paper, a pseudo likelihood is constructed based on the asymptotic distribution of the de-biased 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 $\mathcal{O}(n^{1/2-\epsilon}/(d \log p))$, $\epsilon \in [1/6, 1/2)$, which is a function of the total sample size, the sparsity d and the number of variables, p.

The rest of the paper is organized as follows. In Section 2, notation and preliminaries are introduced. A methodology for performing distributed estimation based on sub-samples is presented in Section 3. We continue in Section 4 by proposing a final estimator that pools together separate estimators. Theoretical properties of this estimator are also investigated in this section. In Section 5, the performance of the estimator is evaluated at the hand of a controlled simulation study and in Section 6, the performance on a real data set is illustrated. We close with a discussion on the method and possible extensions in Section 7. Auxiliary proofs and more simulation results can be found in the Supplementary materials. An `R`-package, called `DistributedGGL`, is also provided on the website of the corresponding author for general use.

2 Notation and preliminaries

The model in this paper is defined under the sub-Gaussianity assumption. A zero-mean random variable X is sub-Gaussian if there exists a constant γ such that $\mathbb{E}\big(\exp(tX)\big) \leqslant \exp(\gamma^2 t^2/2)$ for all $t \in \mathbb{R}$. This upper bound on the moment generating function implies the tail bound $\mathbb{P}(|X| > x) \leqslant 2\exp(-x^2/2\gamma^2)$ for all $x > 0$. Similarly, a zero-mean random vector $\dot{\mathbf{X}} = (X^1, \ldots, X^p)^\top$ with covariance matrix $\boldsymbol{\Sigma}$ is sub-Gaussian if all normalized components $X^a/\sqrt{\boldsymbol{\Sigma}_{aa}}, a = 1, \ldots, p$, are sub-Gaussian random variables with a common parameter $\gamma > 0$, where by $\boldsymbol{\Sigma}_{aa}$ we denote the a-th diagonal element of $\boldsymbol{\Sigma}$. A prime example of a sub-Gaussian random vector is a multivariate zero-mean Gaussian vector. Consider the Gaussian graphical model, where a p-dimensional random vector $\dot{\mathbf{X}} = (X^1, \ldots, X^p)^\top$ follows a multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with mean vector zero and covariance matrix $\boldsymbol{\Sigma}$. The components of $\dot{\mathbf{X}}$ 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 $X^1, \ldots, X^p$. A pair (a, b) is included in the edge set $\mathcal{E}$ if and only if the variables X^a and X^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 $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ 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 $\mathbf{\Theta}$ per row, i.e., the maximal node degree, with d and the index set of non-zero off-diagonal entries of $\mathbf{\Theta}$ with S, formally, $S = \{(a,b) \mid \mathbf{\Theta}_{ab} \neq 0, a \neq b\}$, having cardinality s. The set of non-active components is denoted by S^c.

Before starting the discussion, we introduce some notation which is needed later in the paper. 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 $\{a_n;\ n \geqslant 1\}$ and $\{b_n;\ n \geqslant 1\}$, $b_n = \mathcal{O}(a_n)$ if there exist positive numbers M_0 and N_0 such that $|\frac{b_n}{a_n}| \leqslant M_0$ for all $n \geqslant N_0$. We write $b_n \asymp a_n$ if both $b_n = \mathcal{O}(a_n)$ and $a_n = \mathcal{O}(b_n)$ hold. Similarly, for a random sequence $\{X_n;\ n \geqslant 1\}$, we write $X_n = \mathcal{O}_p(a_n)$ if for every $\epsilon > 0$, there exist finite numbers $M_0 > 0$ and $N_0 > 0$ such that $\mathbb{P}(|\frac{X_n}{a_n}| > M_0) < \epsilon$ for all $n \geqslant N_0$. Furthermore, $b_n = o(a_n)$ if $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$. In the case of a random sequence $\{X_n;\ n \geqslant 1\}$, we write $X_n = o_p(a_n)$ if $\frac{X_n}{a_n} \xrightarrow{p} 0$, 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 $\mathbf{x}$.

3 Distributed penalized estimation

In this section, we propose distributed estimators of the precision matrix $\mathbf{\Theta}$ using only sub-sample information. First, denote the Hessian of the negative log-likelihood function $\ell(\mathbf{\Theta}) = \text{trace}(\hat{\mathbf{\Sigma}}_k \mathbf{\Theta}) - \log\det(\mathbf{\Theta})$, where $\hat{\mathbf{\Sigma}}_k = \mathbf{X}_k^\top \mathbf{X}_k / n_k$ is the sample covariance in the k-th sub-sample (properly defined later in the section), evaluated at the true precision matrix $\mathbf{\Theta}$ by $\mathbf{\Gamma}$, which can be shown to be $\mathbf{\Gamma} = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ (see Ravikumar et al. [2011]). By definition, $\mathbf{\Gamma}$ is a $p^2 \times p^2$ matrix indexed by the pair of elements from the vertex set, such that $\mathbf{\Gamma} = [\mathbf{\Gamma}_{(a,b),(c,d)}]$, where $(a,b),(c,d) \in \mathcal{V} \times \mathcal{V}$. Let $S_1 = \{S \cup \{(1,1),(2,2),\ldots,(p,p)\}\}$ with cardinality s_1, which is equal to $s_1 = s + p$, and its complement set with S_1^c. The following regularity assumptions (see for example, Ravikumar et al. [2011], Jankova and van de Geer [2015]) are considered for the theoretical guarantees in estimating $\mathbf{\Theta}$:

(A1) The irrepresentability condition holds for the true precision matrix $\mathbf{\Theta}$, i.e., there exists $\alpha \in (0,1]$ such that $\max_{e \in S_1^c} \|\mathbf{\Gamma}_{eS_1} \left(\mathbf{\Gamma}_{S_1 S_1}\right)^{-1}\|_1 \leqslant 1 - \alpha$, where $\mathbf{\Gamma}_{S_1 S_1} \in \mathbb{R}^{s_1 \times s_1}$ is a sub-matrix of $\mathbf{\Gamma}$ whose rows and columns are

indexed by the elements of S_1. Moreover, e is a pair $(a, b) \in S_1^c$ such that $\mathbf{\Gamma}_{eS_1}$ is an s_1-dimensional column vector with elements $\mathbf{\Gamma}_{e,(c,d)}$, where $(c, d) \in S_1$.

Assumption (A1) is necessary and sufficient for the graphical Lasso estimator to exhibit model selection consistency.

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

Assumption (A2) is needed to control the convergence rate of the de-biased estimator via controlling the magnitude of the elements of $\mathbf{\Theta}$.

Without loss of generality, suppose that $\mathbb{E}(\dot{\mathbf{X}}) = \mathbf{0}$. Consider now an i.i.d. sample of size n from $\dot{\mathbf{X}}$ and suppose that it is divided randomly into K non-overlapping sub-samples each with size n_k for the k-th sub-sample, $k = 1, \ldots, K$. Consider the i-th row of the k-th sub-sample as $\dot{\mathbf{X}}_{i,k} = (X_{i,k}^1, \ldots, X_{i,k}^p)^\top$, $i = 1, \ldots, n_k$, and denote the k-th sub-sample in matrix form as $\mathbf{X}_k = \left[\dot{\mathbf{X}}_{1,k}, \dot{\mathbf{X}}_{2,k}, \ldots, \dot{\mathbf{X}}_{n_k,k}\right]^\top$, which is of dimension $n_k \times p$ and obtained by appending column vectors $\dot{\mathbf{X}}_{1,k}, \ldots, \dot{\mathbf{X}}_{n_k,k}$ 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 $\mathbf{\Theta}$, by combining K estimators, each obtained on 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 log-likelihood function. This kind of regularization effectively forces some of the elements of $\mathbf{\Theta}$ 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 $\mathbf{X}_k$ $(k = 1, \ldots, K)$, the graphical Lasso estimator $\hat{\mathbf{\Theta}}_k$, defined by Friedman et al. [2008], is the solution to the optimization problem

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

where $\lambda_k > 0$ is the penalty term, $\hat{\mathbf{\Sigma}}_k = \mathbf{X}_k^\top \mathbf{X}_k / n_k$ is the sample covariance in the k-th sub-sample, $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal penalty of the matrix defined as $\|\mathbf{\Theta}\|_{1,\text{off}} = \sum_{a \neq b} |\mathbf{\Theta}_{ab}|$, the quantity $\mathbf{\Theta}_{ab}$ is the (a, b)-th element of $\mathbf{\Theta}$ and S_{++}^p is the space of positive definite matrices of dimension $p \times p$. The graphical Lasso estimator (1) is biased due to the ℓ_1 penalty which is added to the loss function. This estimator satisfies the Karush-Kuhn-Tucker (KKT) condition (see for example, Ravikumar et al. [2011]) as $\hat{\mathbf{\Sigma}}_k - \hat{\mathbf{\Theta}}_k^{-1} + \lambda_k \hat{\mathbf{D}}_k = \mathbf{0}$,

where the matrix $\hat{\mathbf{D}}_k$ belongs to the sub-differential of the off-diagonal norm $\|\cdot\|_{1,\text{off}}$ evaluated at $\hat{\mathbf{\Theta}}_k$. Jankova and van de Geer [2015] proposed the de-biased graphical Lasso estimator to correct the bias, which for our setting can be constructed using the k-th sub-sample as

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

This de-biased 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 $(a, b) \in \mathcal{V} \times \mathcal{V}$,

$$\sqrt{n_k}(\hat{\mathbf{\Theta}}_{ab,k}^d - \mathbf{\Theta}_{ab})/\sigma_{ab} = \sqrt{n_k}\{\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}\}_{ab}/\sigma_{ab} + \sqrt{n_k}\mathbf{R}_{ab,k}/\sigma_{ab}, \tag{3}$$

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 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 $d^{3/2} = o\big(\sqrt{n_k}/(C\log p)\big)$ with

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

the random sequence $\sqrt{n_k}\{\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}\}_{ab}/\sigma_{ab}$ converges weakly to $\mathcal{N}(0, 1)$, where $\kappa_{\mathbf{\Sigma}} := |||\mathbf{\Sigma}|||_\infty$, $\kappa_{\mathbf{\Gamma}} := |||\big(\mathbf{\Gamma}_{S_1S_1}\big)^{-1}|||_\infty$ and α is defined in (A1). Moreover, the elementwise ℓ_∞ norm of $\mathbf{R}_k$ is of order

$$\|\mathbf{R}_k\|_\infty = \mathcal{O}_p\left(\frac{1}{\alpha^2}\kappa_{\mathbf{\Gamma}}\max\left\{d^{3/2}(\log p)/n_k, \frac{1}{\alpha}\kappa_{\mathbf{\Gamma}}d^2((\log p)/n_k)^{3/2}\right\}\right), \tag{4}$$

and by letting $1/\alpha = \mathcal{O}(1)$, $\kappa_{\mathbf{\Sigma}} = \mathcal{O}(1)$, $\kappa_{\mathbf{\Gamma}} = \mathcal{O}(1)$, 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 de-biased estimator $\hat{\mathbf{\Theta}}_k^d$ is of order

$$\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty = \mathcal{O}_p\Big(\max\big\{d\sqrt{(\log p)/n_k}, 1/\alpha^2\kappa_{\mathbf{\Gamma}}d^{3/2}(\log p)/n_k, 1/\alpha^3\kappa_{\mathbf{\Gamma}}^2d^2((\log p)/n_k)^{3/2}\big\}\Big), \tag{5}$$

where under the above bounds on $\kappa_{\mathbf{\Sigma}}$, $\kappa_{\mathbf{\Gamma}}$ and $1/\alpha$, it is simplified to

$$\|\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}\|_\infty = \mathcal{O}_p\Big(\max\left\{ d\sqrt{(\log p)/n_k}, d^{3/2}(\log p)/n_k, d^2((\log p)/n_k)^{3/2} \right\} \Big). \tag{6}$$

A possible 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 4, where we leverage the asymptotic distribution of each $\hat{\mathbf{\Theta}}_k^d$ $(k = 1, \ldots, K)$ from (3) to construct the aggregated estimator using K estimators from the sub-samples.

4 Combined estimator across the sub-samples

The estimator in (2) is defined at the level of the k-th sub-sample and as such one creates a sequence of estimators $\hat{\mathbf{\Theta}}_k^d$ $(k = 1, \ldots, K)$. One can envision multiple ways to combine estimators $\hat{\mathbf{\Theta}}_k^d$ and obtain a pooled, final estimator. One simple, naive method consists in averaging over the sub-samples, i.e., $\hat{\mathbf{\Theta}}_{\text{naive}}^I = (1/K)\sum_{k=1}^K \hat{\mathbf{\Theta}}_k^d$. However, this estimator tends to underperform severely in unbalanced settings. A slightly more adapted naive estimator would account for the unbalanced setting by weighting according to the sizes of the samples on each machine, i.e., $\hat{\mathbf{\Theta}}_{\text{naive}}^{II} = \sum_{k=1}^K (n_k/n)\hat{\mathbf{\Theta}}_k^d$. In practice, as Section 5 shows, this correction might not be enough to guarantee a good performance, hence we introduce here an aggregated estimator based on the pseudo log-likelihood, which is better equipped for unbalanced distributed settings.

Consider $\hat{f}_{ab,k}(\cdot \mid \mathbf{\Theta}_{ab}, \hat{\sigma}_{ab,k})$ as the asymptotic normal density of $\hat{\mathbf{\Theta}}_{ab,k}^d$ obtained from (3) by substituting the consistent estimator $\hat{\sigma}_{ab,k}$ for σ_{ab}. By similar techniques as Tang et al. [2020] for generalized linear models, we propose to obtain a combined estimator by maximizing the pseudo log-likelihood function obtained from the asymptotic densities as

$$\hat{\mathbf{\Delta}}_{ab} = \arg\max_{\mathbf{\Theta}_{ab}} \left\{ \log\Big(\prod_{k=1}^K \hat{f}_{ab,k}(\cdot \mid \mathbf{\Theta}_{ab}, \hat{\sigma}_{ab,k}) \Big) \right\}. \tag{7}$$

Whenever this estimator is derived, the general theory of maximum likelihood estimation can be leveraged to produce valid standard statistical tests and other useful results for statistical inference. By maximizing the pseudo log-likelihood function (7) with respect to $\mathbf{\Theta}_{ab}$, we obtain the final estimator as

$$\hat{\mathbf{\Delta}}_{ab} = \frac{1}{\sum_{k=1}^K \frac{n_k}{\hat{\sigma}_{ab,k}^2}} \times \sum_{k=1}^K \frac{n_k}{\hat{\sigma}_{ab,k}^2} \hat{\mathbf{\Theta}}_{ab,k}^d. \tag{8}$$

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], in each sub-sample. Using (3) and (8), it can be shown that

$$\sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \left(\hat{\mathbf{\Delta}}_{ab} - \mathbf{\Theta}_{ab} \right) = \mathbf{W}_{ab,\dagger} + \mathbf{R}_{ab,\dagger}, \tag{9}$$

where

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

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 $\mathbf{\Theta}_a$ is the a-th column of $\mathbf{\Theta}$.

In the following, it is assumed that the precision matrix $\mathbf{\Theta}$ satisfies the irrepresentability condition (A1) with a constant $\alpha \in (0, 1]$, where $1/\alpha = \mathcal{O}(1)$, and the bounded eigenvalues (A2). The quantities $\kappa_{\mathbf{\Sigma}}$, $\kappa_{\mathbf{\Gamma}}$ are also considered bounded as $\kappa_{\mathbf{\Sigma}} = \mathcal{O}(1)$ and $\kappa_{\mathbf{\Gamma}} = \mathcal{O}(1)$, respectively. Lemma 1 shows that by considering an upper bound on K, the remainder term $\mathbf{R}_{ab,\dagger}$ is negligible. Moreover, Theorem 1 states that the term $\mathbf{W}_{ab,\dagger}$ converges weakly to $\mathcal{N}(0, 1)$ as K and $n_\dagger = \min_{1 \leqslant k \leqslant K} n_k$ grow.

Lemma 1 Suppose that $\hat{\mathbf{\Theta}}_k$ $(k = 1, \ldots, K)$ is the solution to the optimization problem (1) with tuning parameter $\lambda_k \asymp \sqrt{(\log p)/n_k}$ and $\hat{\mathbf{\Delta}}_{ab}$ is the pooled estimator as expressed in (8). Let $n_k/n \to c_k \in (0, 1)$ as $n_k \to \infty$ such that $\lim_{K \to \infty} \sum_{k=1}^{K} c_k = 1$. Then it follows

$$|\mathbf{R}_{ab,\dagger}| = \mathcal{O}_p\big(K/\sqrt{n} \max\{d^{3/2} \log p, d^2 (\log p)^{3/2}/\sqrt{n_\dagger}\}\big), \tag{10}$$

and under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$ and $K = \mathcal{O}\big(n^{1/2-\epsilon}/(d \log p)\big)$, $\epsilon \in [1/6, 1/2)$, it holds that $|\mathbf{R}_{ab,\dagger}| = o_p(1)$.

Proof First note that using Lemma 2 of Jankova and van de Geer [2015], under the multivariate Gaussian distribution, $1/\sigma_{ab} = \mathcal{O}(1)$, and under $1/\alpha = \mathcal{O}(1)$ and $\kappa_{\mathbf{\Gamma}} = \mathcal{O}(1)$, we get $|\hat{\sigma}^2_{ab,k} - \sigma^2_{ab}| = \mathcal{O}_p(\sqrt{(\log p)/n_k})$ which is $o_p(1)$ as n_k grows faster than

$\log p$. Using the continuous map $g(x) = 1/x$ for the consistent estimator $\hat{\sigma}^2_{ab,k}$, we get $\sigma^2_{ab}/\hat{\sigma}^2_{ab,k} \xrightarrow{p} 1$ and since $\frac{n_k}{n} \xrightarrow[n_k \to \infty]{} c_k$, then $\frac{n_k}{n} \times \frac{\sigma^2_{ab}}{\hat{\sigma}^2_{ab,k}} \xrightarrow{p} c_k$, as $n_k \to \infty$. Using the dominated convergence theorem and the assumption $\lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1$, we get

$$\frac{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}}{\sum_{k=1}^{K} \frac{n_k}{\sigma^2_{ab}}} = \sum_{k=1}^{K} \left\{ \frac{\frac{n_k}{\hat{\sigma}^2_{ab,k}}}{\sum_{k=1}^{K} \frac{n_k}{\sigma^2_{ab}}} \right\} = \sum_{k=1}^{K} \left\{ \frac{n_k}{n} \times \frac{\sigma^2_{ab}}{\hat{\sigma}^2_{ab,k}} \right\} \xrightarrow{p} 1, \tag{11}$$

as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$. Again, by considering the continuous map $g(x) = 1/\sqrt{x}$, the sequence $\sqrt{\frac{\sum_{k=1}^{K} n_k/\sigma^2_{ab}}{\sum_{k=1}^{K} n_k/\hat{\sigma}^2_{ab,k}}}$ 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 (10) for the term $\sum_{k=1}^{K} \frac{n_k \sigma_{ab}}{\sqrt{n}\hat{\sigma}^2_{ab,k}} \mathbf{R}_{ab,k}$. Since the graphical Lasso estimator $\hat{\mathbf{\Theta}}_k$ is positive definite, there exists a positive constant L_k such that with high probability $\hat{\sigma}^2_{ab,k} = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}^2_{ab,k} \geqslant \Lambda^2_{\min}(\hat{\mathbf{\Theta}}_k) > L_k$, where $\Lambda_{\min}(\hat{\mathbf{\Theta}}_k)$ is the minimum eigenvalue of $\hat{\mathbf{\Theta}}_k$, and then $1/\hat{\sigma}^2_{ab,k} = \mathcal{O}_p(1)$. Thus, using (4), with high probability,

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

and the claimed result in (10) follows. By considering $K = \mathcal{O}\big(n^{1/2-\epsilon}/(d \log p)\big)$ and the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$, it can be shown that

$$|\mathbf{R}_{ab,\dagger}| = o_p\Big(\max\big\{ \frac{n_\dagger^{1/6}}{n^\epsilon (\log p)^{1/3}}, \frac{1}{(n_\dagger \log p)^{1/6} n^\epsilon} \big\} \Big),$$

which concludes $|\mathbf{R}_{ab,\dagger}| = o_p(1)$. □

Theorem 1 below proves the asymptotic unbiasedness and normality of $\hat{\mathbf{\Delta}}_{ab}$. As such, one can easily construct confidence intervals and statistical tests based on it.

Theorem 1 *Under the assumptions of* Lemma 1, $\mathbf{W}_{ab,\dagger}$ *in* (9) *converges weakly to* $\mathcal{N}(0, 1)$.

Proof Define

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

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

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

we first 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 convergence in distribution of $\xi_{ab,\dagger}$. Since $1/\hat{\sigma}_{ab,k}^2 = \mathcal{O}_p(1)$, with high probability,

$$\begin{aligned} |\xi_{ab,\dagger} - \xi_{ab,\dagger}^{'}| &= \Big| \sum_{k=1}^{K} \big\{ (\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2} - 1) \times \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top \dot{\mathbf{X}}_{l,k} \dot{\mathbf{X}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}) \big\} \Big| \\ &\leqslant \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \Big\{ \big|\sigma_{ab}^2 - \hat{\sigma}_{ab,k}^2\big| \times \big| \frac{1}{\sigma_{ab}} \sum_{l=1}^{n_k} (\boldsymbol{\Theta}_a^\top \dot{\mathbf{X}}_{l,k} \dot{\mathbf{X}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}) \big| \Big\} \quad (12) \end{aligned}$$

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

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

Under assumption (A2) and using Lemma 8 of Jankova and van de Geer [2015], it holds that $\|\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\|_\infty = \mathcal{O}_p(d\sqrt{(\log p)/n_k})$. Moreover, by Lemma 2 of Jankova and van de Geer [2015], $|\hat{\sigma}_{ab,k}^2 - \sigma_{ab}^2| = \mathcal{O}_p(\sqrt{(\log p)/n_k})$. Substituting $K = \mathcal{O}\big(n^{1/2-\epsilon}/(d \log p)\big)$ and the mentioned bounds in (12), we get

$$|\xi_{ab,\dagger} - \xi_{ab,\dagger}^{'}| = \mathcal{O}_p(Kd(\log p)/\sqrt{n}) = \mathcal{O}_p(1/n^\epsilon),$$

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

Then to show the asymptotic normal distribution of $\xi_{ab,\dagger}^{'}$, we have $\mathbb{E}(\mathbf{Z}_{ab,l,k}/(\sqrt{n}\sigma_{ab})) = 0$, where $\mathbf{Z}_{ab,l,k} := \boldsymbol{\Theta}_a^\top \dot{\mathbf{X}}_{l,k} \dot{\mathbf{X}}_{l,k}^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$. Moreover,

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

where the first equality is deduced based on the definition of $\mathbb{V}\mathrm{ar}(\mathbf{Z}_{ab,l,k})$ which is equal to σ_{ab}^2 under the multivariate Gaussian distribution of $\dot{\mathbf{X}}_{l,k}$. Since n_k/n is bounded and $\lim_{n_k\to\infty} n_k/n = c_k$, using the dominated convergence theorem, the second equality also follows. Since $\mathbf{Z}_{ab,l,k}$ are identical for $l = 1, \ldots, n_k$ and $k = 1, \ldots, K$, we can write

$$\begin{aligned} &\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \frac{1}{n\sigma_{ab}^2} \mathbb{E}\big(\mathbf{Z}_{ab,l,k}^2 \mathrm{I}(|\mathbf{Z}_{ab,l,k}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \\ &= \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{n_k}{n\sigma_{ab}^2} \mathbb{E}\big(\mathbf{Z}_{ab,1,1}^2 \mathrm{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \\ &= \Big(\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \frac{1}{\sigma_{ab}^2} \mathbb{E}\big(\mathbf{Z}_{ab,1,1}^2 \mathrm{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \Big) \end{aligned}$$

$$\times \left(\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{n_k}{n} \right), \tag{13}$$

where $\mathrm{I}(\cdot)$ is the indicator function. Under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$, Jankova and van de Geer [2015] showed that $\lim_{n\to\infty} \frac{1}{\sigma_{ab}^2}\mathbb{E}\big(\mathbf{Z}_{ab,1,1}^2 \mathrm{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) = 0$. Due to the fact that $n = \sum_{k=1}^{K} n_k$, the first limit in (13) is zero and using the dominated convergence theorem, the second limit is equal to 1. As such, the requirements of the Lindeberg-Feller condition (see for example, Theorem 4.12 of Kallenberg [1997]) hold and $\xi'_{ab,\dagger}$ converges in distribution to $\mathcal{N}(0,1)$. □

Remark 1 Based on the asymptotic distribution of $\hat{\mathbf{\Delta}}_{ab}$ from Theorem 1, one can construct inferential procedures. Accordingly, the $(1-\alpha)100\%$ asymptotic confidence interval for $\mathbf{\Theta}_{ab}$ based on the distributed estimator $\hat{\mathbf{\Delta}}_{ab}$ is constructed as $\mathrm{CI}_{ab} = \hat{\mathbf{\Delta}}_{ab} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}_{ab,k}^2}$, 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 $H_0 : \mathbf{\Theta}_{ab} = 0$ (which indicates that there is no edge between nodes corresponding to X^a and X^b) vs $H_1 : \mathbf{\Theta}_{ab} \neq 0$ can be constructed as $|\hat{\mathbf{\Delta}}_{ab}| > \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}_{ab,k}^2}$.

The coverage probabilities and the length of the confidence intervals are investigated via a simulation study in Section 5. To compare the performance of the proposed estimator with the naive estimators, we also derived in Appendix A of the Supplementary materials similar expressions for the asymptotic distribution of the two naive estimators.

Consider $\hat{\mathbf{\Delta}}$ as the matrix form of the pooled final estimator with the elements $\hat{\mathbf{\Delta}}_{ab}$, $(a,b) \in \mathcal{V}\times\mathcal{V}$, derived in (8). Theorem 2 and then Remark 2 provide the convergence rate of the distributed estimator $\hat{\mathbf{\Delta}}$ with respect to the elementwise ℓ_∞ and Frobenius norms.

Theorem 2 *Under assumptions* (A1) *and* (A2) *with* $1/\alpha = \mathcal{O}(1)$, $\kappa_{\mathbf{\Sigma}} = \mathcal{O}(1)$, $\kappa_{\mathbf{\Gamma}} = \mathcal{O}(1)$, *the maximal distance of the distributed estimator* $\hat{\mathbf{\Delta}}$ *from the true precision matrix* $\mathbf{\Theta}$ *is of order*

$$\|\hat{\mathbf{\Delta}} - \mathbf{\Theta}\|_\infty = \mathcal{O}_p\Big(\max\big\{ d\sqrt{(\log p)/n_\dagger}, d^{3/2}(\log p)/n_\dagger, d^2((\log p)/n_\dagger)^{3/2} \big\} \Big). \tag{14}$$

Moreover, under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$, *the estimator* $\hat{\mathbf{\Delta}}$ *is consistent for* $\mathbf{\Theta}$.

Proof Based on the definition of the elementwise ℓ_∞ norm and using (6), for every $(a,b) \in \mathcal{V}\times\mathcal{V}$ we have

$$|\hat{\mathbf{\Delta}}_{ab} - \mathbf{\Theta}_{ab}| \leqslant \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_{ab,k}^2}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_{ab,k}^2} |\hat{\mathbf{\Theta}}_{ab,k}^d - \mathbf{\Theta}_{ab}|$$

$$\leqslant \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \mathcal{O}_p\Big(\max\big\{ d\sqrt{(\log p)/n_k}, d^{3/2}(\log p)/n_k, d^2((\log p)/n_k)^{3/2} \big\} \Big)$$

$$\leqslant \mathcal{O}_p\Big(\max\big\{ d\sqrt{(\log p)/n_\dagger}, d^{3/2}(\log p)/n_\dagger, d^2((\log p)/n_\dagger)^{3/2} \big\} \Big),$$

and assuming the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$, the consistency of $\hat{\mathbf{\Delta}}$ follows. □

Remark 2 Using the relation between the Frobenius and the elementwise ℓ_∞ norms, under the sparsity condition $d^{3/2} = o(\sqrt{n_k}/\log p)$, we obtain

$$\|\hat{\mathbf{\Delta}} - \mathbf{\Theta}\|_F = \mathcal{O}_p\Big(p \max\big\{ d\sqrt{(\log p)/n_\dagger}, d^{3/2}(\log p)/n_\dagger, d^2((\log p)/n_\dagger)^{3/2} \big\} \Big).$$

Remark 3 When the true precision matrix is block diagonal with M blocks, the quantities $\kappa_{\mathbf{\Sigma}}$ and $\kappa_{\mathbf{\Gamma}}$ 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 $d_m \leqslant d$,

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

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 $\kappa_{\mathbf{\Sigma}} = \mathcal{O}(\sqrt{d})$. The same argument can be made for $\kappa_{\mathbf{\Gamma}}$ and it holds that $\kappa_{\mathbf{\Gamma}} = \mathcal{O}(d)$. Then, using (5), the convergence rate of $\hat{\mathbf{\Delta}}$ simplifies to

$$\|\hat{\mathbf{\Delta}} - \mathbf{\Theta}\|_\infty = \mathcal{O}_p\Big(\max\big\{ d\sqrt{(\log p)/n_\dagger}, d^{5/2}(\log p)/n_\dagger, d^4((\log p)/n_\dagger)^{3/2} \big\} \Big), \quad (15)$$

and considering the sparsity condition $d^{5/2} = o(\sqrt{n_\dagger}/\log p)$, the consistency of $\hat{\mathbf{\Delta}}$ follows.

Note that the convergence rate (15) 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 5 corresponding to the block diagonal data generating process, the errors and confidence intervals are computed for all elements of the matrix.

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 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 de-biased estimator based on the non-distributed, full data given by the de-biased graphical Lasso (the estimator of Jankova and van de Geer [2015]), denoted by $\hat{\mathbf{\Theta}}^F$.
2) (Naive I) An estimator based on splitting data and averaging directly the de-biased graphical Lasso estimators from each machine i.e., $\hat{\mathbf{\Theta}}^I_{\text{naive}}$ (see Section 4).
3) (Naive II) An estimator based on splitting data and taking the weighted average of the de-biased graphical Lasso estimators from each machine, i.e., $\hat{\mathbf{\Theta}}^{II}_{\text{naive}}$ (see Section 4).
4) (Top1) The most powerful machine which took $(55 - K)\%$ of the 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.
5) (Thresholded I) The thresholded distributed estimator introduced by Arroyo and Hou [2016], which is a sparse estimator built by thresholding de-biased estimators. To obtain this estimator, a hard threshold ρ is applied on the de-biased estimator (2) in the k-th sub-sample, such that for every $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$,

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

where $\mathrm{I}(\cdot)$ 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] in 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 $\hat{\mathbf{\Theta}}^{D,\tau}_{ab} = \hat{\mathbf{\Theta}}^D_{ab}\mathrm{I}(|\hat{\mathbf{\Theta}}^D_{ab}| > \tau\hat{\sigma}_{ab})$, 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. This estimator is constructed by obtaining the nonparametric Kendall's τ statistic 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 de-biased estimators, the aggregated estimator is constructed. A hard threshold is applied on this estimator to get the final aggregated sparse estimator.

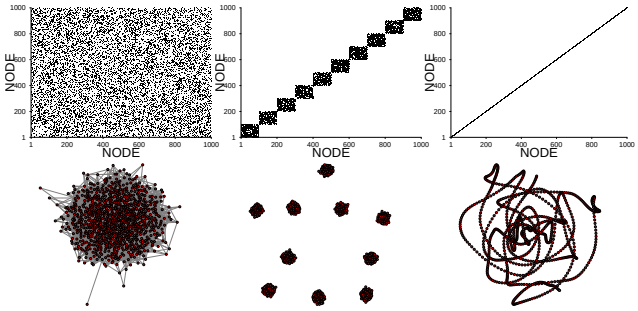

Fig. 1 Visual representation of three $\boldsymbol{\Theta}$ matrices and their graph structures for $p = 1000$ nodes. From left to right: random, block diagonal with 10 blocks and tridiagonal.

A comparison with the full estimator reveals how much the performance deteriorates due to splitting the data, while a comparison with the naive 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 Fig. 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 they proposed this value for the theoretical guarantees. Moreover, as they showed a communication bandwidth of size $B \asymp p^{2-c}$, where $0 < c \leqslant 1$ 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 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 of the Supplementary materials. 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 Naive I 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 Naive II estimator is better performing than Naive I. However, its performance is not as satisfactory as that of the distributed estimator and it gets far from the full estimator with increasing K. As observed from Table 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 II and the Thresholded I with bandwidth $B = 50p$ are also reported in this table. The Frobenius norm of 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 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 $\boldsymbol{\Theta}$ 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 Fig. 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

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

is obtained using the technique proposed in Jankova and van de Geer [2015] as $\sqrt{n}\big(\hat{\mathbf{\Theta}}^F_{ab} - \mathbf{\Theta}_{ab}\big)/\hat{\sigma}_{ab}$, 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 $\sqrt{\sum_{k=1}^K n_k/\hat{\sigma}^2_{ab,k}}\big(\hat{\mathbf{\Delta}}_{ab} - \mathbf{\Theta}_{ab}\big)$. 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 $\alpha = .05$ except for the thresholded estimators which are not asymptotically normally distributed. To this end, we estimated the coverage probability of the

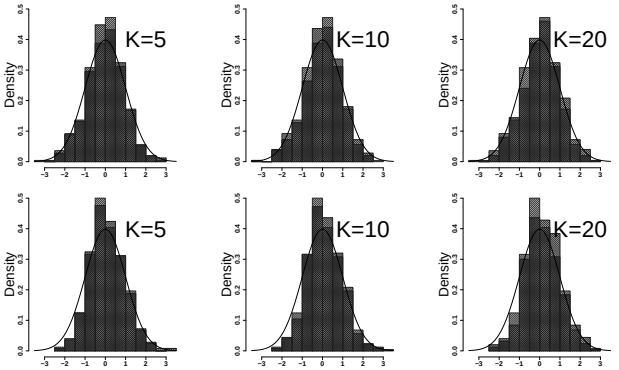

Fig. 2 Histograms of normalized full and distributed estimators corresponding to pair $(a, b) = (1, 3)$ on the top row and $(a, b) = (1, 10)$ on the bottom row for random graph structure when $n = 50000$, $p = 1000$ and the number of machines is $K = 5,\ 10$ or 20.

confidence intervals by computing their empirical version. The empirical probability that the true value $\mathbf{\Theta}_{ab}$ is included in the confidence interval is defined as $\hat{\mathbb{P}}_{ab} = \frac{\#\{\mathbf{\Theta}_{ab} \in \mathrm{CI}_{ab,r}\}}{R}$, where R is the number of simulation repetitions, $\mathrm{CI}_{ab,r}$ is the confidence interval for $\mathbf{\Theta}_{ab}$ at the r-th repetition, and # denotes the number of times that $\mathbf{\Theta}_{ab}$ 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 for the random and block diagonal structures. The results of tridiagonal were close to the random one and they are reported in Appendix B of the Supplementary materials, due to the space constraints. The proposed estimator performs similarly to the 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 of Naive I 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 Naive II is generally better than that of Naive I, 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 paper, 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 Fig. 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.

Table 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			
		Distributed		.91	.93	.90		.03	.03	.03
		Top1		.93	.93	.94		.04	.04	.05
		Naive I		.93	.89	.62		.04	.05	.05
		Naive II		.92	.95	.88		.03	.03	.03
	Non-active set	Full	.97				.03			
		Distributed		.97	.97	.97		.03	.03	.03
		Top1		.97	.97	.97		.04	.04	.05
		Naive I		.97	.96	.95		.04	.05	.05
		Naive II		.97	.98	.98		.03	.03	.03
Block diag.	Active set	Full	.85				.03			
		Distributed		.86	.89	.72		.03	.03	.03
		Top1		.90	.90	.91		.04	.04	.05
		Naive I		.91	.72	.24		.04	.05	.05
		Naive II		.87	.90	.60		.03	.03	.03
	Non-active set	Full	.97				.03			
		Distributed		.97	.97	.96		.03	.03	.03
		Top1		.97	.97	.97		.04	.04	.04
		Naive I		.96	.96	.95		.04	.04	.05
		Naive II		.97	.97	.98		.03	.03	.03

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 were not considered in the six main categories. In this study, we considered the four largest categories which are

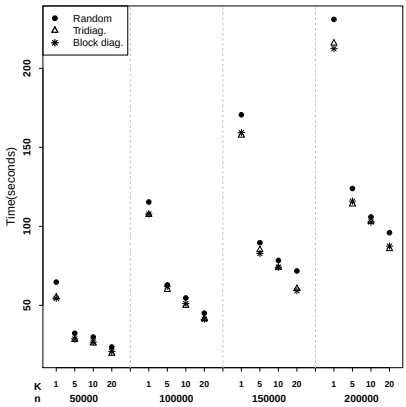

Fig. 3 Running time in seconds for the full and proposed estimator for the random graph structure, when $p = 1000$. The regularization parameter for the distributed estimator is considered as $\lambda_k = \sqrt{(\log p)/n_k}$.

Table 3 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						
Distributed		855	835	818	811	789	751
Top1		569	561	569	482	477	478
Naive I		720	547	418	646	478	360
Naive II		821	671	515	783	657	512

student, faculty, course and project containing 1641, 1124, 930 and 504 web-pages, respectively. To calculate the term-document matrix $\mathbf{X}$, 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, Naive I and II, the proposed distributed estimator with $K = 3, 4, 5$ and the de-biased 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 naive estimators is provided in Appendix A of the Supplementary materials and for the non-distributed estimator (that uses the full data) of Jankova and van de Geer [2015] it 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 3. 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 Naive I. By increasing K,

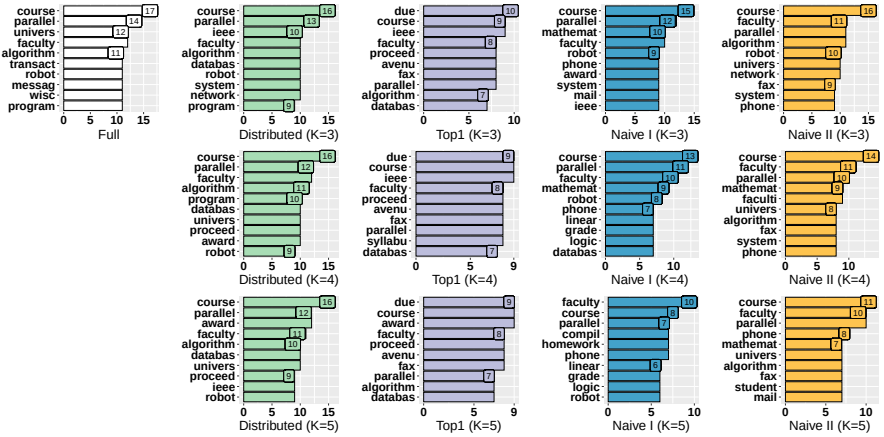

Fig. 4 The first 10 terms with the highest node degrees, identified by different estimators.

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 Fig. 4. The term "course" is identified as the term with the highest node degree by all estimators except Top1 and Naive I 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 matrix should be estimated as zero. By implementing separately the de-biased estimator using the full data, the Top1, the Naive II 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, Naive I identified wrongly between 2 and 19 false positive edges, which confirms the weak performance of this estimator.

7 Conclusion

In this paper, 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. Finally, all estimators were pooled together into a composite estimator using a pseudo log-likelihood function. Then, statistical guarantees for the distributed estimator were provided. Its tractable limit distribution was derived and it was shown that under the sparsity condition $d^{3/2} = o(\sqrt{n_\dagger}/\log p)$ and the upper bound $K = O(n^{1/2-\epsilon}/(d \log p)), \epsilon \in [1/6, 1/2)$, 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 of the proposed estimator 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.

References

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.

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.

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.

Cardoso-Cachopo, A. (2007). Improving methods for single-label text categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers 23*(2), 229–236.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics 9*(3), 432–441.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

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.

Jankova, J. and S. van de Geer (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics 9*(1), 1205–1229.

Kallenberg, O. (1997). *Foundations of modern probability*, Volume 2. Springer.

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.

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.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282.

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

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.

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.

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, 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.

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.

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.