Meta Learning for Support Recovery in High-dimensional Precision Matrix Estimation

Qian Zhang [1] **Yilin Zheng** [2] **Jean Honorio** [2]

Abstract

In this paper, we study meta learning for support (i.e., the set of non-zero entries) recovery in high-dimensional precision matrix estimation where we reduce the sufficient sample complexity in a novel task with the information learned from other auxiliary tasks. In our setup, each task has a different random true precision matrix, each with a possibly different support. We assume that the union of the supports of all the true precision matrices (i.e., the true support union) is small in size. We propose to pool all the samples from different tasks, and *improperly* estimate a single precision matrix by minimizing the ℓ_1-regularized log-determinant Bregman divergence. We show that with high probability, the support of the *improperly* estimated single precision matrix is equal to the true support union, provided a sufficient number of samples per task $n \in O((\log N)/K)$, for N-dimensional vectors and K tasks. That is, one requires less samples per task when more tasks are available. We prove a matching information-theoretic lower bound for the necessary number of samples, which is $n \in \Omega((\log N)/K)$, and thus, our algorithm is minimax optimal. Then for the novel task, we prove that the minimization of the ℓ_1-regularized log-determinant Bregman divergence with the additional constraint that the support is a subset of the estimated support union could reduce the sufficient sample complexity of successful support recovery to $O(\log(|S_{\text{off}}|))$ where $|S_{\text{off}}|$ is the number of off-diagonal elements in the support union and is much less than N for sparse matrices. We also prove a matching information-theoretic lower bound of $\Omega(\log(|S_{\text{off}}|))$ for the necessary number of samples.

[1]Department of Statistics, Purdue University, West Lafayette, USA [2]Department of Computer Science, Purdue University, West Lafayette, USA. Correspondence to: Qian Zhang <zhan3761@purdue.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

Precision (or inverse covariance) matrix estimation is an important problem in high-dimensional statistical learning (Wang et al., 2016) with great application in time series (Chen et al., 2013), principal component analysis (Fan et al., 2016), probabilistic graphical models (Meinshausen et al., 2006), etc. For example, in Gaussian graphical models where we model the variables in a graph as a zero-mean multivariate Gaussian random vector, the set of off-diagonal non-zero entries of the precision matrix corresponds exactly to the set of edges of the graph (Ravikumar et al., 2011). For this reason, estimating the precision matrix to recover its support set, which is the set of non-zero entries, is the common strategy of structure learning in Gaussian graphical models. An estimate of the precision matrix is called sign-consistent if it has the same support and sign of entries with respect to the true matrix.

However, the learner faces several challenges in precision matrix estimation. The first challenge is the high-dimensionality of the data. The dimension of the data, N, could be much higher than the sample size n, and thus the empirical sample covariance and its inverse will behave badly (Johnstone, 2001). Secondly, unlike in Gaussian graphical models, the data may not follow multivariate Gaussian distribution. The third challenge is the heterogeneity of the data. There could be limited samples from the distribution of interest but a large amount of samples from multiple multivariate distributions with different precision matrices.

For the first two challenges, we assume the precision matrices are sparse and consider a general class of distributions, i.e., multivariate sub-Gaussian distributions later described in Definition 1. The class of sub-Gaussian variates (Buldygin & Kozachenko, 1980) includes for instance Gaussian variables, any bounded random variable (e.g. Bernoulli, multinomial, uniform), any random variable with strictly log-concave density, and any finite mixture of sub-Gaussian variables. Then we address the high-dimension challenge by using ℓ_1-regularized log-determinant Bregman divergence minimization (Ravikumar et al., 2011), which is also the ℓ_1-regularized maximum likelihood estimator for multivariate Gaussian distributions (Yuan & Lin, 2007).

For the challenge of heterogeneity, prior works have considered a multi-task learning problem where the learner treats each different distribution as a task with a related precision matrix and solves each and every task simultaneously. Suppose there are K tasks and n samples with dimension N per task. When there is only one task ($K = 1$), Ravikumar et al. (Ravikumar et al., 2011) proved that $n \in O(\log N)$ is sufficient for the sign-consistency of ℓ_1-regularized log-determinant Bregman divergence minimization with multivariate sub-Gaussian data. When $K > 1$, Honorio et al. (Honorio et al., 2012) proposed the $\ell_{1,p}$-regularized log-determinant Bregman divergence minimization to estimate the precision matrices of all tasks and proved that $n \in O(\log K + \log N)$ is sufficient for the correct support union recovery with high probability. Guo et al. (Guo et al., 2011) introduced a different regularized maximum likelihood estimation to learn all precision matrices and proved $n \in O((N \log N)/K)$ is sufficient for the correct support recovery of the precision matrix in each task with high probability. Ma and Michailidis (Ma & Michailidis, 2016) proposed a joint estimation method consisting of a group Lasso regularized neighborhood selection step and a maximum likelihood step. They proved that their method recovers the support of the precision matrix in each task with high probability if $n \in O(K + \log N)$. There are also several algorithms for the multi-task problem but without theoretical guarantees for the consistency of their estimates (Mohan et al., 2014; Chiquet et al., 2011).

In this paper, we solve the heterogeneity challenge with meta learning where we recover the support of the precision matrix in a novel task with the information learned from other auxiliary tasks. Unlike previous methods, we also use improper estimation in our meta learning method to have better theoretical guarantees for support recovery. Specifically, instead of estimating each and every precision matrix in the auxiliary tasks, we pool all the samples from the auxiliary tasks together to estimate a single "common precision matrix" (see Definition 3) in order to recover the "support union" (see Definition 3) of the precision matrices in those tasks. Then we estimate the precision matrix of the novel task with the constraint that its support is a subset of the estimated support union and its diagonal entries are equal to the diagonal entries of the estimated common precision matrix. We prove that for the sign-consistency of our estimates, the sufficient and necessary sample size per auxiliary task is $n \in \Theta((\log N)/K)$ which is much better than the results of the aforementioned multi-task learning methods and enables the learner to gather more tasks (instead of more samples per task) to get a more accurate estimate since the sample complexity is inversely proportional to K. The sufficient and necessary sample complexity of the novel task is $\Theta(\log(|S_{\text{off}}|))$ where $|S_{\text{off}}|$ is the number of off-diagonal elements in the support union S and $|S_{\text{off}}| \ll N$ for sparse graphs, which is better than the result in (Ravikumar et al., 2011).

Moreover, to the best of our knowledge, we are the first to introduce randomness in the precision matrices of different tasks while previous methods assume the precision matrix in each task to be deterministic. Our theoretical results hold for a wide class of distributions of the precision matrices under some conditions, which broadens the application scenarios of our method. The use of improper estimation in our method is innovative for the problem of support recovery of high-dimensional precision matrices. Our work also fills in the blank of the theory and methodology of meta learning in high-dimensional precision matrix estimation. Generally, meta learning aims to develop learning approaches that could have good performance on an extensive range of learning tasks and generalize to solve new tasks easily and efficiently with only a few training examples (Vanschoren, 2019). Thus it is also referred to as *learning to learn* (Lake et al., 2015). Current research mainly focuses on designing practical meta learning algorithms, for instance, (Koch et al., 2015; Vinyals et al., 2016; Sung et al., 2018; Santoro et al., 2016; Munkhdalai & Yu, 2017; Finn et al., 2017). We believe our work could provide some insights for the theoretical understanding of meta learning.

This paper has the following four contributions. Firstly, we propose a meta learning approach by introducing multiple auxiliary learning tasks for support recovery of high-dimensional precision matrices with improper estimation. Secondly, we add randomness to the precision matrices in different learning tasks, which is a significant innovation compared to previous methods. Thirdly, we prove that for N-dimensional multivariate sub-Gaussian random vectors and K auxiliary tasks with support union S, the sufficient sample complexity of our method is $O((\log N)/K)$ per auxiliary task for support union recovery and $O(\log(|S_{\text{off}}|))$ for support recovery of the novel task, which provides the theoretical basis for introducing more tasks for meta learning in support recovery of precision matrices. Fourthly, we prove information-theoretic lower bounds for the failure of support union recovery in the auxiliary tasks and the failure of support recovery in the novel task. We show that $\Omega((\log N)/K)$ samples per auxiliary task and $\Omega(\log(|S_{\text{off}}|))$ samples for the novel task are necessary for the recovery success, which proves that our meta learning method is minimax optimal. Lastly, we conduct synthetic and real-world experiments to validate our theory. We calculate the support union recovery rates of our meta learning approach and multi-task learning approaches for different sizes of samples and tasks. For a fixed task size K, our approach achieves high support union recovery rates when the sample size per task has the order $O((\log N)/K)$. For a fixed sample size per task, our method performs the best when the task size K is large. Our meta learning approach also achieves the minimum

log-determinant Bregman divergence in the estimation of the precision matrices of the novel tasks in two real-world datasets compared to multi-task learning approaches and the graphical lasso method.

2. Preliminaries

This section introduces our mathematical models and the meta learning problem. The important notations used in the paper are illustrated in Table 1.

2.1. Multivariate Sub-Gaussian Distributions with Random Precision Matrices

We first define a general class of multivariate distributions, the multivariate sub-Gaussian distribution.

Definition 1. *We say a random vector $X \in \mathbb{R}^N$ follows a multivariate sub-Gaussian distribution with precision $\Omega \in \mathbb{R}^{N\times N}$ and parameter σ if*

(i) $\mathbb{E}\left[X_t^{(k)}\right] = 0$, $Cov\,(X) = \Sigma = (\Omega)^{-1}$, *and*

(ii) $\frac{X_i}{\sqrt{\Sigma_{ii}}}$ *is a sub-Gaussian random variable with parameter σ for $1 \le i \le N$.*

The definition of sub-Gaussian random variable is as follows (Buldygin & Kozachenko, 2000):

Definition 2. *A random variable $X \in \mathbb{R}$ is called sub-Gaussian with parameter $\sigma \ge 0$ if*

$$\mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \quad \forall\, \lambda \in \mathbb{R} \tag{1}$$

Obviously, Gaussian variables are sub-Gaussian and the Gaussian graphical model is a special case of the multivariate sub-Gaussian distribution.

In this paper, we consider multiple multivariate sub-Gaussian distributions whose precision matrices are randomly generated, which makes our model more reasonable and universal compared to the deterministic setting in all the previous works. Formally, we define the following family of multivariate sub-Gaussian distributions with random precision matrices:

Definition 3. *Let $X_1^{(k)}\, X_2^{(k)}, ..., X_{n^{(k)}}^{(k)} \in \mathbb{R}^N$ be i.i.d. random vectors for $1 \le k \le K$. Let $X_{t,i}^{(k)}$ be the i-th entry of $X_t^{(k)}$ for $1 \le i \le N$. We say $\left\{X_t^{(k)}\right\}_{1\le t\le n^{(k)},\ 1\le k\le K}$ follows a family of random N-dimensional multivariate sub-Gaussian distributions of size K with parameter σ if*

(i) $\bar{\Omega}^{(k)} = \bar{\Omega} + \Delta^{(k)}$ with $\bar{\Omega}, \Delta^{(k)} \in \mathbb{R}^{N\times N}$, $\bar{\Omega} \succ 0$ deterministic, and $\Delta^{(k)}, 1 \le k \le K$, are i.i.d. random matrices drawn from distribution P;

(ii) For some $\gamma > 0, c_{\max} \in \left(0, \lambda_{\min}(\bar{\Omega})/2\right]$, we have

$$\begin{aligned}\mathbb{P}_{\Delta\sim P}[&\bar{\Omega} + \Delta \succ 0, supp(\Delta) \subseteq supp(\bar{\Omega}),\\ &\|(\bar{\Omega}+\Delta)^{-1}\|_\infty \le \gamma, |||\Delta|||_2 \le c_{\max}] = 1\end{aligned} \tag{2}$$

and $\beta := \|(\bar{\Omega})^{-1} - \mathbb{E}_{\Delta\sim P}[(\bar{\Omega}+\Delta)^{-1}]\|_\infty < \infty$;

(iii) $\mathbb{E}\left[X_t^{(k)}\middle|\bar{\Sigma}^{(k)}\right] = 0$, $Cov\left(X_t^{(k)}\middle|\bar{\Sigma}^{(k)}\right) = \bar{\Sigma}^{(k)}$ for $\bar{\Sigma}^{(k)} := \left(\bar{\Omega}^{(k)}\right)^{-1}$, $1 \le t \le n^{(k)}$, $1 \le k \le K$;

(iv) $\left\{X_t^{(k)}\right\}_{1\le t\le n^{(k)},\ 1\le k\le K}$ are conditionally independent given $\{\bar{\Omega}^{(k)}\}_{k=1}^K$;

(v) $\frac{X_{t,i}^{(k)}}{\sqrt{\bar{\Sigma}_{ii}^{(k)}}}$ conditioned on $\bar{\Omega}^{(k)}$ is sub-Gaussian with parameter σ for $1 \le i \le N$, $1 \le t \le n^{(k)}$, $1 \le k \le K$.

We refer to $\bar{\Omega}$ as the true common precision matrix and $S := supp(\bar{\Omega})$ as the support union of the above family of distributions.

Notice that we define the support union as $S = \text{supp}(\bar{\Omega})$ instead of $\cup_{k=1}^K \text{supp}(\bar{\Omega}^{(k)})$ which is a random subset of the deterministic set S because we are interested on a novel task where the support of its precision matrix is a subset of the support of Ω, i.e., S.

2.2. Problem Setting

In this paper, we focus on the problem of estimating the support of the precision matrix of a multivariate sub-Gaussian distribution. Following the principles of meta learning, we solve a novel task by first estimating a superset of the support of the precision matrix in the novel task from K auxiliary tasks.

Specifically, suppose there are $n^{(K+1)}$ samples from a multivariate sub-Gaussian distribution with precision matrix $\bar{\Omega}^{(K+1)}$ for the novel task. We introduce $n^{(k)}$ samples for each auxiliary task $k \in \{1, ..., K\}$ and assume all samples in the K auxiliary tasks follow a family of random multivariate sub-Gaussian distributions with common precision matrix $\bar{\Omega}$ specified in Definition 3. Our meta learning method aims to recover the support union $S = \text{supp}(\bar{\Omega})$ with the K auxiliary tasks and use S to assist in recovering $S^{(K+1)} := \text{supp}(\bar{\Omega}^{(K+1)})$ with the assumption that $S^{(K+1)} \subseteq S$.

3. Our Novel Improper Estimation Method

As illustrated in Section 2.2, in the first step of our method, we recover the support union S of the K auxiliary tasks by estimating the true common precision matrix $\bar{\Omega}$. To be specific, we pool all samples from the K tasks together and estimate $\bar{\Omega}$ by minimizing the ℓ_1-regularized log-determinant Bregman divergence between the estimate and $\bar{\Omega}$; i.e., we

Table 1. Notations used in the paper

Notation	Description
$\text{sign}(x)$	The sign of $x \in \mathbb{R}$, i.e., $\text{sign}(x) = x/\|x\|$ if $x \neq 0$; $\text{sign}(x) = 0$ if $x = 0$
$\|a\|_\infty$	The ℓ_∞-norm of vector $a \in \mathbb{R}^n$, i.e., $\max_{i=1}^n \|a_i\|$
$\|a\|_1$	The ℓ_1-norm of vector $a \in \mathbb{R}^n$, i.e., $\sum_{i=1}^n \|a_i\|$
$\|A\|_\infty$	The ℓ_∞-norm of matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\max_{1\le i\le m, 1\le j\le n} \|A_{ij}\|$
$\|A\|_1$	The ℓ_1-norm of matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\sum_{1\le i\le m, 1\le j\le n} \|A_{ij}\|$
$\|\|\|A\|\|\|_\infty$	The ℓ_∞-operator-norm of matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\max_{1\le i\le m} \sum_{j=1}^n \|A_{ij}\|$
$\lambda_{\min}(A)$	The minimum eigenvalue of matrix $A \in \mathbb{R}^{m\times m}$
$\lambda_{\max}(A)$	The maximum eigenvalue of matrix $A \in \mathbb{R}^{m\times m}$
$\|\|\|A\|\|\|_2$	The ℓ_2-operator-norm of matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$
$A \succ 0$	The matrix A is symmetric and positive-definite.
$\det(A)$	The determinant of matrix $A \in \mathbb{R}^{m\times n}$
$\text{supp}(A)$	The support set of matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\{(i,j)\|A_{ij} \neq 0\}$
$\text{diag}(A)$	The vector consisting of the diagonal entries of matrix $A \in \mathbb{R}^{n\times n}$, i.e., $[A_{11}, A_{22}, \ldots, A_{nn}]^{\mathrm{T}}$
$\|S\|$	The number of elements in the set S
S_{off}	The set of off-diagonal elements in the set S, i.e., $\{(i,j) : (i,j) \in S, i \neq j\}$
A_S	The sub-matrix composed by the entries according to the set S of $A \in \mathbb{R}^{m\times n}$, i.e., $\left(A_{(i,j)}\right)_{(i,j)\in S}$
$\langle A, B\rangle \in \mathbb{R}$	The Frobenius inner product of $A, B \in \mathbb{R}^{m\times n}$, i.e., $\sum_{1\le i\le m, 1\le j\le n} A_{ij}B_{ij}$
$A \odot B \in \mathbb{R}^{m\times n}$	The Hadamard product of $A, B \in \mathbb{R}^{m\times n}$, i.e., $[A \odot B]_{ij} = A_{ij}B_{ij}$
$A \otimes B \in \mathbb{R}^{mp\times nq}$	The Kronecker product of $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$, i.e., $[A \otimes B]_{(i,j),(k,l)} = [A \otimes B]_{p(i-1)+k,q(j-1)+l} = A_{ij}B_{kl}$
$[A \otimes B]_{S_1S_2}$	The sub-matrix composed by the entries according to the set $S_1 \times S_2$ of the matrix $A \otimes B \in \mathbb{R}^{mp\times nq}$ for $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$, i.e., $\left([A \otimes B]_{(i,j),(k,l)}\right)_{(i,j)\in S_1,(k,l)\in S_2}$

solve the following optimization problem with regularization constant $\lambda > 0$:

$$\hat{\Omega} = \arg\min_{\Omega \succ 0} \sum_{k=1}^{K} T^{(k)} \left(\log\det(\Omega) - \langle \hat{\Sigma}^{(k)}, \Omega\rangle\right) - \lambda\|\Omega\|_1 \tag{3}$$

where $\hat{\Sigma}^{(k)} := \frac{1}{n^{(k)}} \sum_{t=1}^{n^{(k)}} X_t^{(k)} \left(X_t^{(k)}\right)^{\mathrm{T}}$ is the empirical sample covariance and $T^{(k)}$ is proportional to the number of samples $n^{(k)}$ for task k. Define the following loss function:

$$\ell(\Omega) = \sum_{k=1}^{K} T^{(k)} \left(\langle \hat{\Sigma}^{(k)}, \Omega\rangle - \log\det(\Omega)\right) \tag{4}$$

Then we can rewrite (3) as

$$\hat{\Omega} = \arg\min_{\Omega \succ 0} \left(\ell(\Omega) + \lambda\|\Omega\|_1\right) \tag{5}$$

For clarity of exposition, we assume the number of samples per auxiliary task is the same, i.e., $n^{(k)} = n$, $T^{(k)} = 1/K$ for $1 \le k \le K$ in our analysis. In addition, we do not assume $n^{(K+1)} = n$. Notice that (5) is an improper estimation because we estimate a single precision matrix with data from different distributions. This will enable us to recover the support union with the most efficient sample size per task (see Section 4.2.1).

For the second step, suppose that we have successfully recovered the true support union S in the first step. Then for a novel task, i.e., the $(K+1)$-th task, since we have assumed the support $S^{(K+1)}$ of its precision matrix $\hat{\Omega}^{(K+1)}$ is also a subset of the support union S, we propose the following constrained ℓ_1-regularized log-determinant Bregman divergence minimization for $\hat{\Omega}^{(K+1)}$:

$$\begin{aligned}\hat{\Omega}^{(K+1)} = &\arg\min_{\Omega \succ 0} \ell^{(K+1)}(\Omega) + \lambda\|\Omega\|_1 \\ &\text{s.t. } \text{supp}(\Omega) \subseteq \text{supp}(\hat{\Omega}),\ \ \text{diag}(\Omega) = \text{diag}(\hat{\Omega}).\end{aligned} \tag{6}$$

where $\ell^{(K+1)}(\Omega) := \langle \hat{\Sigma}^{(K+1)}, \Omega\rangle - \log\det(\Omega)$, $\hat{\Sigma}^{(K+1)} := \frac{1}{n^{(K+1)}} \sum_{t=1}^{n^{(K+1)}} X_t^{(K+1)} \left(X_t^{(K+1)}\right)^{\mathrm{T}}$ is the empirical sample covariance and $\hat{\Omega}$ is obtained in (5). Note that (6) is also an improper estimation because of the constraint $\text{diag}(\Omega) = \text{diag}(\hat{\Omega})$. For our target of support recovery and sign-consistency, there is no need to estimate the diagonal entries of the precision matrix since they are always positive. Hence, we introduce this constraint to reduce the sample complexity by only focusing on estimating the off-diagonal entries (see Section 4.2.2).

4. Theoretical Results

In this section, we formally state our assumptions and theoretical results.

4.1. Assumptions

Our theoretical results require an assumption on the true common precision matrix $\bar{\Omega}$ which is called mutual incoherence or irrepresentability condition in (Ravikumar et al., 2011). The Hessian of the loss function (4) when $\Omega = \bar{\Omega}$ is

$$\nabla^2 \ell\left(\bar{\Omega}\right) = T\bar{\Gamma} \tag{7}$$

where $T := \sum_{k=1}^{K} T^{(k)}$ and $\bar{\Gamma} := \nabla^2 \log\det(\bar{\Omega}) = \bar{\Omega}^{-1} \otimes \bar{\Omega}^{-1} \in \mathbb{R}^{N^2 \times N^2}$. The mutual incoherence assumption is as follows:

Assumption 1. *There exists some $\alpha \in (0, 1]$ such that $\left|\left|\left|\bar{\Gamma}_{S^cS}(\bar{\Gamma}_{SS})^{-1}\right|\right|\right|_\infty \leq 1 - \alpha$*

We should notice that $\left|\left|\left|\bar{\Gamma}_{S^cS}(\bar{\Gamma}_{SS})^{-1}\right|\right|\right|_\infty = \max_{u\in S^c} \|\bar{\Gamma}_{uS}(\bar{\Gamma}_{SS})^{-1}\|_1$. Thus this assumption in fact places restrictions on the influence of non-support terms indexed by S^c, on the support-based terms indexed by S (Ravikumar et al., 2011).

We also require the mutual incoherence assumption for the precision matrix $\bar{\Omega}^{(K+1)}$ in the novel task:

Assumption 2. *There exists $\alpha^{(K+1)} \in (0, 1]$ such that*

$$\left|\left|\left|\bar{\Gamma}^{(K+1)}_{(S^{(K+1)})^cS^{(K+1)}}(\bar{\Gamma}^{(K+1)}_{S^{(K+1)}S^{(K+1)}})^{-1}\right|\right|\right|_\infty \leq 1 - \alpha^{(K+1)} \tag{8}$$

where $\bar{\Gamma}^{(K+1)} := (\bar{\Omega}^{(K+1)})^{-1} \otimes (\bar{\Omega}^{(K+1)})^{-1}$.

For $\bar{\Gamma}$ and $\bar{\Gamma}^{(K+1)}$, our analysis keeps explicit track of the quantities $\kappa_{\bar{\Gamma}} := \left|\left|\left|(\bar{\Gamma}_{SS})^{-1}\right|\right|\right|_\infty$ and $\kappa_{\bar{\Gamma}^{(K+1)}} := \left|\left|\left|(\bar{\Gamma}^{(K+1)}_{S^{(K+1)}S^{(K+1)}})^{-1}\right|\right|\right|_\infty$.

To relate the two norms $\|\cdot\|_\infty$ and $|||\cdot|||_\infty$, we define the degree of a matrix as the maximal size of the supports of its row vectors. The degree of $\bar{\Omega}$ is $d := \max_{1\leq i\leq N}\left|\left\{j : 1 \leq j \leq N, \bar{\Omega}_{ij} \neq 0\right\}\right|$ and the degree of $\bar{\Omega}^{(K+1)}$ is $d^{(K+1)} := \max_{1\leq i\leq N}\left|\left\{j : 1 \leq j \leq N, \bar{\Omega}^{(K+1)}_{ij} \neq 0\right\}\right|$.

We call $\bar{\Sigma} := \bar{\Omega}^{-1}$ the true common covariance matrix and denote its ℓ_∞-operator-norm by $\kappa_{\bar{\Sigma}} := \left|\left|\left|\bar{\Sigma}\right|\right|\right|_\infty$. Similarly for the covariance matrix $\bar{\Sigma}^{(K+1)} = (\bar{\Omega}^{(K+1)})^{-1}$ in the novel task, we define $\kappa_{\bar{\Sigma}^{(K+1)}} := \left|\left|\left|\bar{\Sigma}^{(K+1)}\right|\right|\right|_\infty$.

In order to bound $\left|\left|\left|\bar{\Sigma}\right|\right|\right|_2$ in our proof, we define $\lambda_{\min} := \lambda_{\min}(\bar{\Omega})$.

To show the sign-consistency of our estimators, we also need to consider the minimal magnitude of non-zero entries in $\bar{\Omega}$ and $\bar{\Omega}^{(K+1)}$, i.e., $\omega_{\min} := \min_{(i,j)\in S} |\bar{\Omega}_{ij}|$, $\omega^{(K+1)}_{\min} := \min_{(i,j)\in S} |\bar{\Omega}^{(K+1)}_{ij}|$,

4.2. Main Theorems

For our meta learning method, we have

Lemma 1. *For $\lambda > 0$, the problem in (5) and (6) are strictly convex and have unique solutions $\hat{\Omega}$ and $\hat{\Omega}^{(K+1)}$ respectively.*

The detailed proofs of all the lemmas, theorems and corollaries in the paper are in the supplementary material. We then study the theoretical behaviors of $\hat{\Omega}$ in (5) and $\hat{\Omega}^{(K+1)}$ in (6).

4.2.1. Support Union Recovery

Our first theorem specifies a probability lower bound of recovering a subset of the true support union by our estimator in (5) for multiple random multivariate sub-Gaussian distributions.

Theorem 1. *For a family of N-dimensional random multivariate sub-Gaussian distributions of size K with parameter σ described in Definition 3 with $n^{(k)} = n$, $1 \leq k \leq K$ and satisfying Assumption 1, consider the estimator $\hat{\Omega}$ obtained in (5) with $T^{(k)} = 1/K$ and $\lambda = (8\delta + 4\delta^*)/\alpha$ for $\delta \in (0, \delta^*/2]$ where $\delta^* := \frac{\alpha^2}{2\kappa_{\bar{\Gamma}}(\alpha+8)^2}\min\left\{\frac{1}{3\kappa_{\bar{\Sigma}}d}, \frac{1}{3\kappa^3_{\bar{\Sigma}}\kappa_{\bar{\Gamma}}d}\right\}$. If $\beta \leq \delta^*/2$, then with probability at least*

$$1 - 2N(N+1)\exp\left(-\frac{nK}{2}\min\left\{\frac{\delta^2}{64(1+4\sigma^2)^2\gamma^2}, 1\right\}\right) - 2N\exp\left(-\frac{K\lambda^4_{\min}}{128c^2_{\max}}\left(\frac{\delta^*}{2} - \beta\right)^2\right) \tag{9}$$

we have:

(i) supp$(\hat{\Omega}) \subseteq$ supp $\left(\bar{\Omega}\right)$

(ii) $\|\hat{\Omega} - \bar{\Omega}\|_\infty \leq \kappa_{\bar{\Gamma}}\left(\frac{8}{\alpha} + 1\right)(2\delta + \delta^*)$

Proof sketch for Theorem 1. We use the primal-dual witness approach (Ravikumar et al., 2011) to prove Theorem 1. The key step is to verify that the strict dual feasibility condition holds. Using some norm inequalities and Brouwer's fixed point theorem (see e.g. (Ortega & Rheinboldt, 2000)), we show that it suffices to bound the random term $\|\sum_{k=1}^{K}\frac{1}{K}W^{(k)}\|_\infty$ with $W^{(k)} = \hat{\Sigma}^{(k)} - \bar{\Sigma}$ for $1 \leq k \leq K$ after some careful and involved derivation. Then we decompose the random term into two parts as

follows

$$\begin{aligned}
&\|\sum_{k=1}^{K}\frac{1}{K}W^{(k)}\|_\infty \\
=&\|\frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}^{(k)}-\bar{\Sigma}^{(k)}+\bar{\Sigma}^{(k)}-\bar{\Sigma}\|_\infty \\
\leq&\underbrace{\|\frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}^{(k)}-\bar{\Sigma}^{(k)}\|_\infty}_{Y_1}+\underbrace{\|\frac{1}{K}\sum_{k=1}^{K}\bar{\Sigma}^{(k)}-\bar{\Sigma}\|_\infty}_{Y_2}
\end{aligned}$$

Conditioning on $\{\bar{\Sigma}^{(k)}\}_{k=1}^K$, Y_1 can be bounded by the sub-Gaussianity of the samples after some careful derivation. Then by the law of total expectation we can get the term $2N(N+1)\exp\left(-\frac{nK}{2}\min\left\{\frac{\delta^2}{64(1+4\sigma^2)^2\gamma^2},1\right\}\right)$ in (9).

Define $H:=\frac{1}{K}\sum_{k=1}^{K}\bar{\Sigma}^{(k)}$. We bound Y_2 with the following two terms

$$\begin{aligned}
Y_2&=\|H-\mathbb{E}[H]+\mathbb{E}[H]-\bar{\Sigma}\|_\infty \\
&\leq\|\mathbb{E}[H]-\bar{\Sigma}\|_\infty+|||H-\mathbb{E}[H]|||_2 \qquad (10)\\
&=\beta+|||H-\mathbb{E}[H]|||_2
\end{aligned}$$

since $\|\mathbb{E}[H]-\bar{\Sigma}\|_\infty=\|(\bar{\Omega})^{-1}-\mathbb{E}_{\Delta\sim P}[(\bar{\Omega}+\Delta)^{-1}]\|_\infty=\beta$. Then we bound $|||H-\mathbb{E}[H]|||_2$ with Corollary 7.5 in (Tropp, 2011) to get the term $2N\exp\left(-\frac{K\lambda_{\min}^4}{128c_{\max}^2}\left(\frac{\delta^*}{2}-\beta\right)^2\right)$ in (9). The detailed proof is in the supplementary material. □

Our proof follows the primal-dual witness approach (Ravikumar et al., 2011). From Theorem 1, we can see that for our method, a sample complexity of $O((\log N)/K)$ per task is sufficient for the recovery of a subset of the true support union.

The next theorem addresses the sign-consistency of the estimate (5). We say the estimator $\hat{\Omega}$ is sign-consistent if

$$\text{sign}(\hat{\Omega}_{ij})=\text{sign}(\bar{\Omega}_{ij})\ \text{ for }\ \forall i,j\in\{1,2,...,N\} \qquad (11)$$

It is obvious that sign-consistency immediately implies the success of support recovery.

Theorem 2. *For a family of N-dimensional random multivariate sub-Gaussian distributions of size K with parameter σ described in Definition 3 with $n^{(k)}=n$, $1\leq k\leq K$ and satisfying Assumption 1, consider the estimator $\hat{\Omega}$ obtained in (5) with $T^{(k)}=1/K$ and $\lambda=8\delta^\dagger/\alpha$ where*

$$\delta^\dagger:=\begin{cases}\frac{\alpha^2}{2\kappa_{\bar{\Gamma}}(\alpha+8)^2}\min\left\{\frac{1}{3\kappa_{\bar{\Sigma}}d},\frac{1}{3\kappa_{\bar{\Sigma}}^3\kappa_{\bar{\Gamma}}d}\right\},\\ \quad \textit{if } \omega_{\min}\geq\frac{2\alpha}{8+\alpha}\min\left\{\frac{1}{3\kappa_{\bar{\Sigma}}d},\frac{1}{3\kappa_{\bar{\Sigma}}^3\kappa_{\bar{\Gamma}}d}\right\};\\ \frac{\alpha\omega_{\min}}{4(8+\alpha)\kappa_{\bar{\Gamma}}},\quad \textit{otherwise,}\end{cases}$$

If $\beta\leq\delta^\dagger/2$, then with probability at least

$$\begin{aligned}
&1-2N(N+1)\exp\left(-\frac{nK}{2}\min\left\{\frac{(\delta^\dagger)^2}{256(1+4\sigma^2)^2\gamma^2},1\right\}\right)\\
&-2N\exp\left(-\frac{K\lambda_{\min}^4}{128c_{\max}^2}\left(\frac{\delta^\dagger}{2}-\beta\right)^2\right)
\end{aligned} \qquad (12)$$

the estimator $\hat{\Omega}$ is sign-consistent and thus supp$(\hat{\Omega})=$ supp $(\bar{\Omega})$.

According to Theorem 2, a sample complexity of $O((\log N)/K)$ per task is sufficient for the recovery of the true support union by our estimator in (5).

We also prove the following information-theoretic lower bound on the failure of support union recovery for some family of random multivariate sub-Gaussian distributions.

Theorem 3. *For some family of N-dimensional random multivariate sub-Gaussian distributions of size K with parameter σ and covariance matrices $\{\bar{\Sigma}^{(k)}\}_{k=1}^K$, suppose $N\geq 5$, $\bar{\Sigma}^{(k)}=(I+H\odot Q^{(k)})^{-1}$ for $1\leq k\leq K$ with $Q^{(k)}\in[-1/(2d),1/(2d)]^{N\times N}$ symmetric, degree $d\in\mathbb{Z}^+$ even and $H\in\{0,1\}^{N\times N}$ such that H is symmetric and $H_{ij}=1$ iff $(i,j)\in E$. Thus $S:=E\cup\{(i,i)\}_{i=1}^N$ is the support union of all precision matrices. Assume E is randomly generated in the following way:*

(i) Obtain a permutation $\pi=(\pi_1,\pi_2,...,\pi_N)$ of $V=\{1,2,...,N\}$ uniformly at random.

(ii) Let $\pi_{N+j}:=\pi_j$ for $1\leq j\leq d/2$

(iii) For $i=1,...,N$, add (π_i,π_{i+j}) to E for $1\leq j\leq d/2$.

Thus d is the degree of the precision matrices in all tasks. Suppose that for each of the K distributions, we have n samples randomly drawn from them. Then for any estimate $\hat{S}$ of S, we have

$$\mathbb{P}\{\hat{S}\neq S\}\geq 1-\frac{nNK+\log 2}{N\log N-N-\log 2N} \qquad (13)$$

Proof sketch for Theorem 3. For the random set S, random samples $\mathbf{X}=\{X_t^{(k)}\}_{1\leq t\leq n,1\leq k\leq K}$, and $\mathbf{Q}:=\{Q^{(k)}\}_{k=1}^K$, we prove that the conditional entropy $H(S|\mathbf{Q})=\log((N-1)!/2)$ and the conditional mutual information $\mathbb{I}(\mathbf{X};S|\mathbf{Q})\leq nNK$.

By the Fano's inequality extension in (Ghoshal & Honorio, 2017), we have

$$\mathbb{P}\{\hat{S} \neq S\} \geq 1 - \frac{\mathbb{I}(\mathbf{X}; S|\mathbf{Q}) + \log 2}{H(S|\mathbf{Q})} \geq 1 - \frac{nNK + \log 2}{\log[(N-1)!/2]}$$

which leads to (13). The detailed proof is in the supplementary material. □

According to Theorem 3, if the sample size per distribution is $n \leq (\log N)/(2K) - 1/(2K) - (\log(8N))/(2NK)$, then with probability larger than $1/2$, any method will fail to recover the support union of the multiple random multivariate sub-Gaussian distributions specified in Theorem 3. Thus a sample complexity of $\Omega((\log N)/K)$ per task is necessary for the support union recovery of the N-dimensional multivariate sub-Gaussian distributions in K tasks, which, combined with Theorem 2, indicates that our estimate (5) is minimax optimal with a necessary and sufficient sample complexity of $\Theta((\log N)/K)$ per task.

4.2.2. Support Recovery for Novel Task

For the novel task, the next theorem proves a probability lower bound for the sign-consistency of the estimate (6).

Theorem 4. *Suppose we have recovered the true support union S of a family of N-dimensional random multivariate sub-Gaussian distributions of size K with parameter σ described in Definition 3 with $n^{(k)} = n$ for $k = 1, .., K$. For a novel task of multivariate sub-Gaussian distribution with precision matrix $\bar{\Omega}^{(K+1)}$ such that supp$(\bar{\Omega}^{(K+1)}) \subseteq S$ and satisfying Assumption 2, consider the estimator $\hat{\Omega}^{(K+1)}$ obtained in (6) with $\lambda = \frac{8\delta^{(K+1),\dagger}}{\alpha^{(K+1)}}$ where*

$$\delta^{(K+1),\dagger} := \begin{cases} \frac{(\alpha^{(K+1)})^2}{2\kappa_{\bar{\Gamma}}(\alpha^{(K+1)}+8)^2 d^{(K+1)}} \min\left\{\frac{1}{3\kappa_{\bar{\Sigma}^{(K+1)}}}, \frac{1}{3\kappa^3_{\bar{\Sigma}^{(K+1)}}\kappa_{\bar{\Gamma}^{(K+1)}}}\right\}, \\ \quad \text{if } \omega^{(K+1)}_{\min} \geq \frac{2\alpha^{(K+1)}}{(8+\alpha^{(K+1)})d^{(K+1)}} \cdot \\ \quad \min\left\{\frac{1}{3\kappa_{\bar{\Sigma}^{(K+1)}}}, \frac{1}{3\kappa^3_{\bar{\Sigma}^{(K+1)}}\kappa_{\bar{\Gamma}^{(K+1)}}}\right\}; \\ \frac{\alpha^{(K+1)}\omega^{(K+1)}_{\min}}{4(8+\alpha^{(K+1)})\kappa_{\bar{\Gamma}^{(K+1)}}}, \quad \text{otherwise.} \end{cases}$$

If $\|\bar{\Sigma}^{(K+1)}\|_\infty \leq \gamma^{(K+1)}$, then with probability at least,

$$1 - 2|S_{off}| \exp\left(-\frac{n^{(K+1)}}{2} \cdot \min\left\{\frac{(\delta^{(K+1),\dagger})^2}{64(1+4\sigma^2)^2(\gamma^{(K+1)})^2}, 1\right\}\right) \tag{14}$$

the estimator $\hat{\Omega}^{(K+1)}$ is sign-consistent and thus supp$(\hat{\Omega}^{(K+1)})$ = supp $\left(\bar{\Omega}^{(K+1)}\right)$.

Proof sketch for Theorem 4. We use the primal-dual witness approach. Since we have two constraints in (6), we can consider the Lagrangian

$$L(\Omega, \mu, \nu) = \ell^{(K+1)}(\Omega) + \lambda\|\Omega\|_1 + \langle\mu, \Omega\rangle + \langle\nu, \text{diag}(\Omega - \hat{\Omega})\rangle \tag{15}$$

where $\mu \in \mathbb{R}^{N\times N}, \nu \in \mathbb{R}^N$ are the Lagrange multipliers satisfying $\mu_S = 0$. Here we set $\mu = (\bar{\Sigma}^{(K+1)}_{S^c}, 0)$ (i.e., entries of μ with index in S equal 0 and entries of μ with index in S^c equal corresponding entries of $\bar{\Sigma}$) and $\nu = \text{diag}(\bar{\Sigma}^{(K+1)} - \hat{\Sigma}^{(K+1)})$ in (15). Then we show that it suffices to bound $W^{(K+1)} := [\hat{\Sigma}^{(K+1)} - \bar{\Sigma}^{(K+1)}]_{S_{\text{off}}}$ for the strict dual feasibility condition to hold. $W^{(K+1)}$ can be bounded by the sub-Gaussianity of the samples. The detailed proof is in the supplementary material. □

This theorem shows that $n^{(K+1)} \in O(\log(|S_{\text{off}}|))$ is sufficient for recovering the true support of the novel task with our estimate (6). Therefore, the overall sufficient sample complexity for the sign-consistency of the estimators in the two steps of our meta learning approach is $O(\log(N)/K)$ for each auxiliary task and $O(\log(|S_{\text{off}}|))$ for the novel task, which is much better than the results of (Ravikumar et al., 2011), (Honorio et al., 2012), (Guo et al., 2011), and (Ma & Michailidis, 2016), especially for large number of auxiliary tasks K and high dimension N, as discussed in Section 1.

We also prove the following information-theoretic lower bound for the failure of support recovery for some random multivariate sub-Gaussian distribution where the support set is a subset of a known set S_{off}.

Theorem 5. *For n samples generated from some N-dimensional multivariate sub-Gaussian distribution with $N \geq 4$, suppose the true covariance matrix is $\bar{\Sigma} = (I + H \odot Q)^{-1}$ with $Q \in [-\frac{1}{N\log s}, \frac{1}{N\log s}]^{N\times N}$ symmetric and $H \in \{0,1\}^{N\times N}$ such that H is symmetric and $H_{ij} = 1$ iff $(i,j) \in E^{(K+1)}$. Thus $S^{(K+1)} := E^{(K+1)} \cup \{(i,i)\}_{i=1}^N$ is the support set of the precision matrix of this distribution. Assume $E^{(K+1)}$ is chosen uniformly at random from the edge set family $\mathcal{E} := \{E \subseteq S_{off} : (i,j) \in E \implies (j,i) \in E\}$ for a known edge set S_{off}. Define $s := |S_{off}|$. Assume $4 \leq s \leq N$. Then for any estimate $\hat{S}^{(K+1)}$ of $S^{(K+1)}$, we have*

$$\mathbb{P}\{\hat{S}^{(K+1)} \neq S^{(K+1)}\} \geq 1 - \frac{4n}{(\log 2)(\log s)} - \frac{2}{s} \tag{16}$$

Proof sketch for Theorem 5. For the random set $S^{(K+1)}$, random vectors $\mathbf{X} = \{X_t\}_{t=1}^n$, and Q, we prove that the conditional entropy $H(S^{(K+1)}|Q) = \log|\mathcal{E}| \geq \frac{s}{2}\log 2$ and the conditional mutual information $\mathbb{I}(\mathbf{X}; S^{(K+1)}|Q) \leq \frac{2ns}{\log s}$.

By the Fano's inequality extension in (Ghoshal & Honorio,

2017), we have

$$\mathbb{P}\{\hat{S}^{(K+1)} \neq S^{(K+1)}\} \geq 1 - \frac{\mathbb{I}(\mathbf{X}; S^{(K+1)}|Q) + \log 2}{H(S^{(K+1)}|Q)} \geq 1 - \frac{4n}{(\log 2)(\log s)} - \frac{2}{s}$$

The detailed proof is in the supplementary material. □

According to Theorem 5, if $n \leq \frac{\log 2}{8} \log s - \frac{\log 2}{2s} \log s$, then $\mathbb{P}\{S^{(K+1)} \neq \hat{S}^{(K+1)}\} \geq \frac{1}{2}$, which indicates that the necessary sample complexity for the support recovery of the novel task is $\Omega(\log s) = \Omega(\log |S_{\text{off}}|)$ and our estimate (6) is minimax optimal. Therefore, our two-step meta learning method is minimax optimal.

4.3. Computational Complexity

Several algorithms have been developed to solve the ℓ_1-regularized log-determinant Bregman divergence minimization (Hsieh et al., 2012; 2013; Johnson et al., 2012; Cai et al., 2011). We have proved in Lemma 1 that the problems in (5) and (6) are convex, which therefore can be solved in polynomial time with respect to the dimension of the random vector N by using interior point methods (Boyd et al., 2004). Further, state-of-the-art methods for inverse covariance estimation can potentially scale to a million variables (Hsieh et al., 2013).

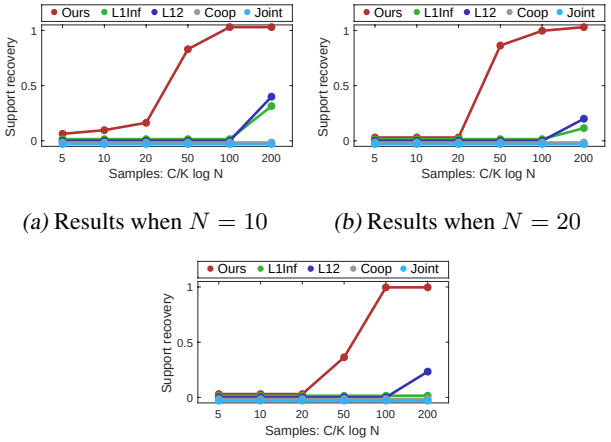
5. Validation Experiments

5.1. Synthetic Experiments

We validate our theories with synthetic experiments by reporting the success rate for the recovery of the support union. We simulate Erdos-Renyi random graphs in this experiment and compare the results of our estimator in (5) with four multi-task learning methods. We generate Erdos-Renyi random graphs as follows. We first generate $\bar{\Omega}$ by assigning an edge with probability $d/(N-1)$ for each pair of nodes (i, j). Then for each edge (i, j), we set $\bar{\Omega}_{ij} = \bar{\Omega}_{ji}$ to 1 with probability 0.5 and to -1 otherwise. For $1 \leq k \leq K$ and $(i, j) \in S$, $\bar{\Omega}_{ij}^{(k)}$ is set to $\bar{\Omega}_{ij} X_{ij}$ with $X_{ij} \sim$Bernoulli(0.9). Then we add some constant to the diagonal elements of all the precision matrices to ensure that their minimum eigenvalue is at least 0.1.

For Figure 1, we fix the number of auxiliary tasks $K = 10$ and run experiments with sample size per auxiliary task $n = (C \log N)/K$ for C ranging from 5 to 200. We can see that our method sucessfully recovers the true support union with probability close to 1 when the sample size per auxiliary task is in the order of $O((\log N)/K)$ while the four multi-task learning methods fail. This result provides experimental evidence for Theorem 2.

For Figure 2, we run experiments for different number of auxiliary tasks K that ranges from 2 to 100 with the sample size per auxiliary task $n = 200(\log N)/K$. According to Figure 2, for our method, the support union recovery probability increases with K and converges to 1 for K large enough. For the four multi-task learning methods, however, the probability decreases to 0 as K grows. The results indicate that even with a small number of samples per auxiliary task, we can get a sufficiently accurate estimate using our meta learning method by introducing more auxiliary tasks.



(a) Results when $N = 10$ *(b)* Results when $N = 20$

(c) Results when $N = 50$

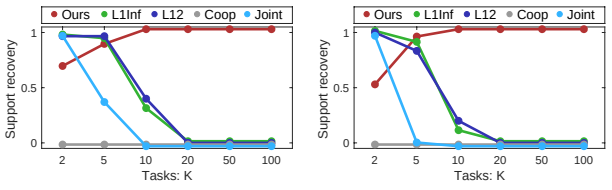
Figure 1. The success rate of support union recovery for different sample size $n = (C \log N)/K$ and task size $K = 10$. Y-axis shows the success probability and X-axis shows the values of C. "Ours" is our meta learning method, which we compare against several multitask methods. "L1Inf" is the $\ell_{1,\infty}$-regularized method (Honorio & Samaras, 2010). "L12" is the $\ell_{1,2}$-regularized method (Varoquaux et al., 2010). "Coop" is the Cooperative-LASSO method in (Chiquet et al., 2011). "Joint" is the joint estimation method in (Guo et al., 2011).

5.2. Real-World Data Experiments

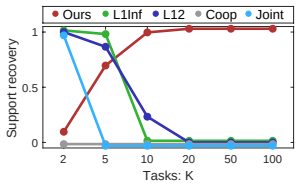
We also use our two-step meta learning method to conduct experiments with two real-world datasets, the single-cell gene expression dataset from (Kouno et al., 2013) and the cancer genome atlas dataset from `http://tcga-data.nci.nih.gov/tcga/`. There are multiple sub datasets in the two datasets and we treat the estimation of the precision matrix of each sub dataset as a learning task. The single-cell gene expression dataset contains 8 tasks. Each task contains 120 samples and corresponds to a different time point (0 h, 1 h, 6 h, 12 h, 24 h, 48 h, 72 h, and 96 h). Each sample has 45 features. In order to simulate a challenging scenario similar to the ones encountered on meta-learning, we use 10 samples of each task 1 to 7 to recover the support union and then use 10 samples of task 8 (the novel task) to recover its precision matrix. The cancer genome atlas dataset contains 5 tasks. Each task corresponds to a

Table 2. Negative log-determinant Bregman divergence of the estimated precision matrices of the novel tasks in the two real-world datasets using different methods.

Method	Negative log-determinant Bregman divergence	
	Single-cell gene expression dataset	Cancer genome atlas dataset
Our meta learning method	**-47**	**-109**
The $\ell_{1,\infty}$-regularized method (Honorio & Samaras, 2010)	-179	-123
The $\ell_{1,2}$-regularized method (Varoquaux et al., 2010)	-100	-117
The Cooperative-LASSO method (Chiquet et al., 2011)	-85	-181
The joint estimation method (Guo et al., 2011)	-534	-150
The graphical lasso method (applied only on the novel task) (Friedman et al., 2008)	-324	-270



(a) Results when $N = 10$ *(b)* Results when $N = 20$



(c) Results when $N = 50$

Figure 2. The success rate of support union recovery for different task size K with the sample size per task $n = (200 \log N)/K$. Y-axis shows the success probability and X-axis shows the value of K. "Ours" is our meta learning method, which we compare against several multitask methods. "L1Inf" is the $\ell_{1,\infty}$-regularized method (Honorio & Samaras, 2010). "L12" is the $\ell_{1,2}$-regularized method (Varoquaux et al., 2010). "Coop" is the Cooperative-LASSO method in (Chiquet et al., 2011). "Joint" is the joint estimation method in (Guo et al., 2011).

different type of cancer (breast invasive carcinoma, colon adenocarcinoma, glioblastoma multiforme, lung squamous cell carcinoma, and ovarian serous cystadenocarcinoma) and contains 590, 174, 595, 155, and 590 samples respectively. Each sample consists of 187 genes commonly regulated in cancer that were identified on independent data sets by (Lu et al., 2007). In order to simulate a challenging scenario similar to the ones encountered on meta-learning, we use 15 samples of each task 1 to 4 to recover the support union and then use 15 samples of task 5 (the novel task) to recover its precision matrix. In Table 2, we report the negative log-determinant Bregman divergence (i.e., the log-likelihood of a multivariate Gaussian distribution) of our meta-learning method for the novel tasks and compare it with the results of four multi-task methods and the graphical lasso method.

According to Table 2, our method generalizes better than the comparison methods for the two datasets since it obtains the minimum log-determinant Bregman divergence.

6. Conclusion

We develop a meta learning approach for support recovery in precision matrix estimation. Specifically, we pool all the samples from K auxiliary tasks with K random precision matrices, and estimate a single precision matrix by ℓ_1-regularized log-determinant Bregman divergence minimization to recover the support union of the auxiliary tasks. Then we estimate the precision matrix of the novel task with the constraint that its support set is a subset of the support union to reduce the sufficient sample complexity. We prove that the sample complexities of $O((\log N)/K)$ per auxiliary task and $O(\log(|S_{\text{off}}|))$ for the novel task are sufficient for our estimators to recover the support union and the support of the precision matrix of the novel task. We also prove that our meta learning method is minimax optimal. Synthetic experiments are conducted and validate our theoretical results.

Finally, we believe that the idea of improper estimation developed on this paper will be useful for other machine learning problems beyond sparse precision matrix estimation analyzed in this paper and sparse regression analyzed in (Wang & Honorio, 2021).

References

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Buldygin, V. V. and Kozachenko, I. V. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc., 2000.

Buldygin, V. V. and Kozachenko, Y. V. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6): 483–489, 1980.

Cai, T., Liu, W., and Luo, X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Chen, X., Xu, M., Wu, W. B., et al. Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics*, 41(6):2994–3021, 2013.

Chiquet, J., Grandvalet, Y., and Ambroise, C. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

El Ghaoui, L. Inversion error, condition number, and approximate inverses of uncertain matrices. *Linear Algebra and its Applications*, 343-344:171–193, 2002. ISSN 0024-3795. doi: https://doi.org/10.1016/S0024-3795(01)00273-7. URL https://www.sciencedirect.com/science/article/pii/S0024379501002737. Special Issue on Structured and Infinite Systems of Linear equations.

Fan, J., Liao, Y., and Liu, H. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

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

Ghoshal, A. and Honorio, J. Information-theoretic limits of bayesian network structure learning. In *Artificial Intelligence and Statistics*, pp. 767–775. PMLR, 2017.

Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU press, 2012.

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

Honorio, J. and Samaras, D. Multi-task learning of Gaussian graphical models. *International Conference on Machine Learning*, pp. 447–454, 2010.

Honorio, J., Jaakkola, T., and Samaras, D. On the statistical efficiency of $\ell_{1,p}$ multi-task learning of gaussian graphical models. *arXiv preprint arXiv:1207.4255*, 2012.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Horn, R. A., Horn, R. A., and Johnson, C. R. *Topics in matrix analysis*. Cambridge university press, 1994.

Hsieh, C.-J., Banerjee, A., Dhillon, I. S., and Ravikumar, P. K. A divide-and-conquer method for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pp. 2330–2338, 2012.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., and Poldrack, R. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pp. 3165–3173, 2013.

Johnson, C., Jalali, A., and Ravikumar, P. High-dimensional sparse inverse covariance estimation using greedy methods. In *Artificial Intelligence and Statistics*, pp. 574–582, 2012.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pp. 295–327, 2001.

Koch, G., Zemel, R., and Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

Kouno, T., de Hoon, M., Mar, J. C., Tomaru, Y., Kawano, M., Carninci, P., Suzuki, H., Hayashizaki, Y., and Shin, J. W. Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome biology*, 14(10):1–12, 2013.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Lu, Y., Yi, Y., Liu, P., Wen, W., James, M., Wang, D., and You, M. Common human cancer genes discovered by integrated gene-expression analysis. *PloS one*, 2(11): e1149, 2007.

Ma, J. and Michailidis, G. Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, 17(1):5777–5824, 2016.

Marshall, A. W., Olkin, I., and Arnold, B. C. Matrix theory. In *Inequalities: Theory of Majorization and Its Applications*, pp. 297–365. Springer, 2010.

Meinshausen, N., Bühlmann, P., et al. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

Mohan, K., Palma London, M. F., Witten, D., and Lee, S.-I. Node-based learning of multiple gaussian graphical models. *Journal of machine learning research: JMLR*, 15(1):445, 2014.

Munkhdalai, T. and Yu, H. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2554–2563. JMLR. org, 2017.

Ortega, J. M. and Rheinboldt, W. C. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850, 2016.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

Vanschoren, J. Meta-learning: A survey. *The Springer Series on Challenges in Machine Learning: Automated Machine Learning*, pp. 35–61—, 2019.

Varoquaux, G., Gramfort, A., Poline, J., and Thirion, B. Brain covariance selection: Better individual functional connectivity models using population prior. *Neural Information Processing Systems*, 23:2334–2342, 2010.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Wang, L., Ren, X., and Gu, Q. Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In *Artificial Intelligence and Statistics*, pp. 177–185. PMLR, 2016.

Wang, Z. and Honorio, J. The sample complexity of meta sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 2323–2331. PMLR, 2021.

Weiss, N., Holmes, P., and Hardy, M. *A Course in Probability*. Pearson Addison Wesley, 2005. ISBN 9780321189547. URL https://books.google.com/books?id=p-rwJAAACAAJ.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.