Multi-task Attributed Graphical Lasso

Anonymous Author(s)

No Institute Given

Abstract. Sparse inverse covariance estimation, *i.e.*, Graphical Lasso, can estimate the connections among a set of random variables basing on their observations. Recent research on Graphical Lasso has been extended to multi-task settings, where multiple graphs sharing the same set of variables are estimated collectively to reduce variances. However, different tasks usually involve different variables. For example, when we want to estimate gene networks w.r.t different diseases simultaneously, the related gene sets vary. In this paper, we study the problem of multi-task Graphical Lasso where the tasks may involve different variable sets. To share information across tasks, we consider the attributes of variables and assume that the structures of graphs are not only determined by observations, but influenced by attributes. We formulate the problem of learning multiple graphs jointly with observations and attributes, *i.e.*, Multi-task Attributed Graphical Lasso (MAGL), and propose an effective algorithm to solve it. We introduce the LogDet divergence to explore latent relations between attributes of the variables and linkage structures among the variables. Multiple precision matrices and a projection matrix are optimized such that the ℓ_1-penalized negative log-likelihood and the divergence are minimized.

1 Introduction

Gaussian Graphical Models (GGMs) [25] provide a powerful framework for describing the dependencies among a set of variables and have been attracting much attention in the fields of finance, social networks and bio-informatics, *etc.* [16,28]. In these applications, some of the edges between the nodes are usually unknown and must be inferred from observations of the node activities. It has been shown that the non-zero elements of the precision matrix, *i.e.*, the inverse of the covariance matrix, correspond to the edges in the underlying graph [25]. Thus structure learning of a GGM is equivalent to estimating its precision matrix, which can be solved via Graphical Lasso (GLasso) [7].

In some cases, multiple GLasso tasks are involved and each contains several observations. Observations in different tasks may come from different distributions, but they are all on the same set of variables. For example, researchers may want to estimate gene regulatory networks for cancer patients and healthy subjects separately using their gene expression levels. Since the gene networks in multiple tasks are highly related, we often estimate multiple precision matrices collectively. The multi-task Graphical Lasso could borrow strength across tasks and reduce the variance of the estimates [19]. There have been some recent work

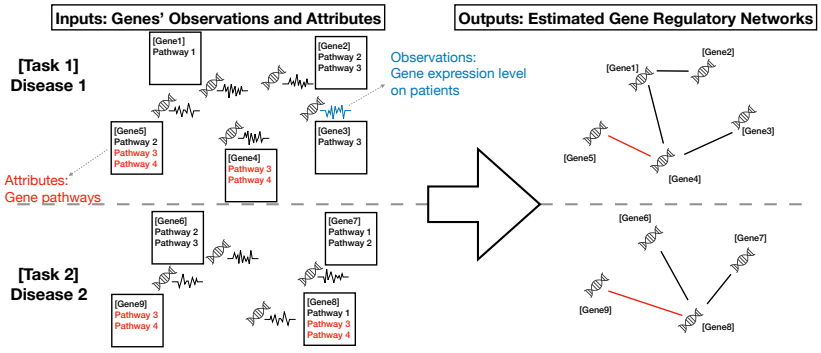


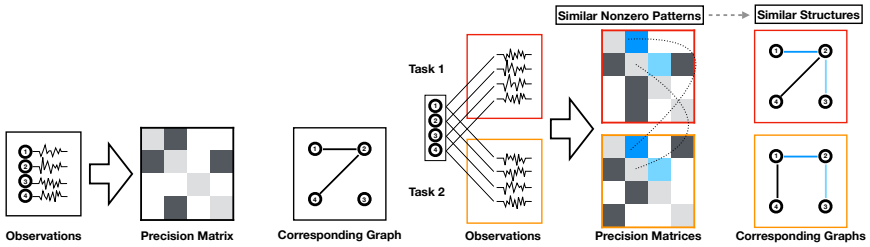
Fig. 1: An illustration of multi-task attributed Graphical Lasso. The tasks are to estimate gene regulatory networks for two diseases collectively to reduce the variance of the estimates. Each gene is accompanied by its expression level on different patients as observations, and pathways it belongs to as attributes. Similar attributes drive certain pairs of variables to be connected (shown in red).

on the multi-task Graphical Lasso [3,12,24,27], but they assume that the sets of variables across tasks are identical and the nonzero patterns in precision matrices are similar across multiple graphs. This is not always the case in the real world where each task could have its own associated variable set. For instance, our tasks are to estimate gene regulatory networks for multiple diseases collectively, but sets of genes involved may not be identical across these diseases. We try to consider multiple sets of variables in multiple tasks.

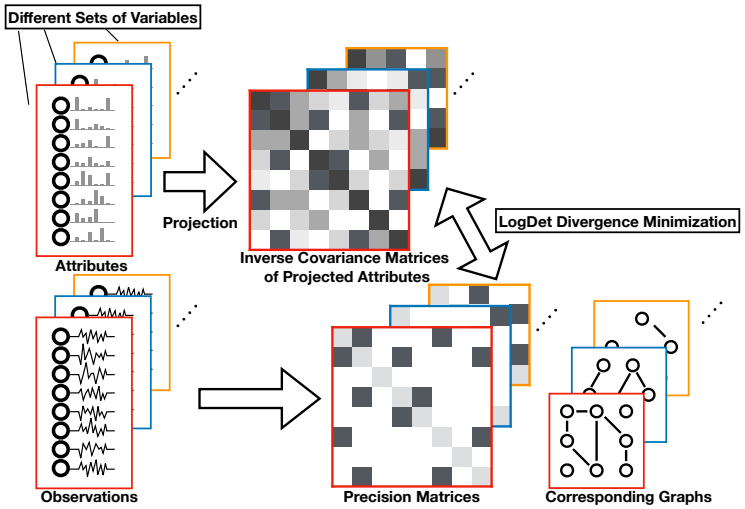
It is not clear how to jointly solve tasks with different variable sets, but there is a key observation that variables are often accompanied by attributes that might help. For example, each gene is associated with attributes, such as gene families, pathways and related-diseases. In this paper, we study the problem of multi-task attributed Graphical Lasso, where the goal is to simultaneously estimate multiple graphs by exploiting the relationship between attributes and graph structures as illustrated in Figure 1.

Despite the significance, the multi-task attributed Graphical Lasso is highly challenging due to:

- **Heterogeneity of Variables**: Since the sets of variables are not necessarily identical across tasks, the existing methods based on the assumption that the similar nonzero patterns in precision matrices are no longer applicable. It is challenging to share information across tasks with heterogeneous sets of variables to improve the quality of estimates.
- **Relations between Attributes and Graphs**: Previous methods do not utilize the attributes and infer the graphs by only using observations of variables. But in attributed graphs, connectivities between variables are also influenced by their attributes. It is challenging to define the relationship between attributes and graphs. Besides, how to inject attributes into the multi-task framework is unclear.



(a) Graphical Lasso [7] (b) Multi-task Graphical Lasso [3,12,24,27]



(c) Multi-task Attributed Graphical Lasso (this paper)

Fig. 2: (a) Graphical Lasso estimates a single precision matrix and the corresponding graph from observations of variables. (b) Multi-task Graphical Lasso estimates graphs jointly from multiple sets of observations under assumption that the nonzero patterns in precision matrices should be similar. (c) Multi-task attributed Graphical Lasso accepts attributes as side information and supports different sets of variables across tasks. It assumes that the structures of graphs are related to the attributes of variables.

To address these issues, we present a novel method called MAGL (Multi-task Attributed Graphical Lasso), which uses the LogDet divergence [18] to build a connection between structures and attributes. Multiple precision matrices and a projection matrix are simultaneously optimized so that the ℓ_1-penalized negative log-likelihood is minimized, meanwhile the LogDet divergence between the precision matrix of graphs and the inverse covariance matrix of projected attributes in each task is also minimized. Since the information is shared indirectly through the projection matrix, our formulation supports heterogeneous sets of variables.

We illustrate the differences between our proposal and existing related problem settings in Figure 2. The main contributions of our paper are as follows: (1)

We study the problem of multi-task attributed Graphical Lasso, and incorporate attributes into the framework of multi-task Graphical Lasso by using the LogDet divergence. (2) We propose an efficient algorithm to solve MAGL using block coordinate descent and augmented Lagrangian method. (3) The conducted experiments illustrate the effectiveness of the proposal.

2 Problem Formulation

In this section, we briefly review some related concepts and notions. We then formulation the problem of multi-task attributes Graphical Lasso.

• **Notations:** In this paper, $\Re$ stands for the set of all real numbers. The space of symmetric matrices is denoted by $\mathcal{S}^n$. The cone of positive semi-definite matrices is denoted by $\mathcal{S}^n_+$, and its interior is $\mathcal{S}^n_{++}$. $\|X\|_1 = \sum_{i,j} |X_{ij}|$ is the element-wise ℓ_1 norm. $\|X\|_F^2 = \sum_{i,j} X_{ij}^2$ is the squared Frobenius norm. $\mathrm{Tr}(\cdot)$ and $\det(\cdot)$ denote the trace and the determinant of a matrix respectively. $\sigma(X)$ returns all singular values of X. $\mathbb{1}\{\text{condition}\}$ is the indicator function.

2.1 Preliminaries

Graphical Lasso: Assume we have a set of samples $X \in \Re^{p\times n}$ drawn i.i.d. from a p-variate Gaussian distribution: $x_j \sim \mathcal{N}_p(\mathbf{0}, \Sigma),\ j = 1, \ldots, n$, where $\Sigma \in \mathcal{S}^p_{++}$, and x_j is the j-th column of X. A natural way to estimate the precision matrix $\Theta = \Sigma^{-1}$ is via maximum log-likelihood estimation (MLE). The log-likelihood function takes the form (up to a constant) $l(S, \Theta) = \log\det\Theta - \mathrm{Tr}(S\Theta)$, where $S = \frac{1}{n}XX^T \in \mathcal{S}^p_+$ is the sample covariance matrix. However, the MLE fails when $p > n$ because S becomes singular. Even if $p \leqslant n$ and S is not singular, S^{-1} is usually dense. To obtain a meaningful estimate, the ℓ_1 regularization has been employed to induce sparsity. This leads to the sparse inverse covariance matrix estimation problem, also known as Graphical Lasso (GLasso) [7]: $\min_\Theta\ -l(S, \Theta) + \lambda\|\Theta\|_1$, where $\lambda > 0$ is an ℓ_1 regularization parameter.

LogDet Divergence: The LogDet divergence [18] is proposed to measure the "closeness" between two matrices $X, Y \in \mathcal{S}^p_{++}$. It is defined as

$$D_{ld}(X, Y) = \mathrm{Tr}(XY^{-1}) - \log\det(XY^{-1}) - p.$$

The LogDet divergence is non-negative, and $D_{ld}(X, Y) = 0$ if and only if $X = Y$. It is convex in the first argument. It has been shown [4] that the KL divergence between two multivariate Gaussian distributions with the same mean vector, $\mathcal{N}(\mu, \Theta^{-1})$ and $\mathcal{N}(\mu, \Omega^{-1})$, is proportional to the LogDet divergence between the corresponding precision matrices:

$$KL\left(\mathcal{N}(\mu, \Theta^{-1}), \mathcal{N}(\mu, \Omega^{-1})\right) = \frac{1}{2} D_{ld}(\Theta, \Omega).$$

2.2 Multi-task Attributed Graphical Lasso

Consider that we are given $K \geqslant 2$ tasks, each consisting of not only variables' activities $X^k \in \Re^{p_k \times n_k}$, but also attributes $A^k \in \Re^{p_k \times m}$, where the i-th row of A^k is the i-th variable's attributes in the k-th task. The samples within each task X^k are identically distributed with a p_k-variate Gaussian distribution with zero mean and covariance matrix $\left(\Theta^k\right)^{-1} \in \mathcal{S}_{++}^{p_k}$. Further we assume that the structures of graphs are influenced by the variables' attributes. We wish to borrow information across the K tasks to estimate the K precision matrices jointly.

For notational simplicity, we assume that $p_i = p$ and $n_i = n$ $\forall i$, but our formulation and algorithm can be easily adapted to the general setting. We formulate the problem of multi-task attributed Graphical Lasso (MAGL) as

$$\min_{\substack{\Theta^k, U \\ 1 \leqslant k \leqslant K}} \sum_{k=1}^{K} \left[-l\left(S^k, \Theta^k\right) + \lambda_1 \|\Theta^k\|_1 \right] + \lambda_2 \sum_{k=1}^{K} D_{ld}\left(\Theta^k, \Omega^k\right) + \frac{\lambda_3}{2} \|U\|_F^2, \quad (2.1)$$

where $\Omega^k = \left(\epsilon I + A^k U U^T (A^k)^T\right)^{-1}$, and $U \in \Re^{m \times d}$ is a projection matrix from a m-dimensional input space to a d-dimensional output space. $\lambda_1, \lambda_2, \lambda_3, \epsilon > 0$ are the model parameters. The first part in the objective function is the sum of K GLasso problems. We view the projected attributes $A^k U \in \Re^{p \times d}$ as d samples drawn from the Gaussian distribution $\mathcal{N}_p\left(0, (\Omega^k)^{-1}\right)$. $\Omega^k = \left(\epsilon I + A^k U U^T (A^k)^T\right)^{-1}$ is the estimate of the inverse of the precision matrix, where ϵ is used to make it non-singular. Now, $D_{ld}(\Theta^k, \Omega^k)$ is the KL divergence between the two Gaussian distributions. By this means, we build a connection between the structures of graphs and the variables' attributes. We also use the squared Frobenius norm of U to prevent overfitting. As illustrated in Figure 2(c), the Problem 2.1 finds K precision matrices and a projection matrix that minimize the negative log-likelihood of data, and meanwhile minimize the divergence between the precision matrices of data and projected attributes in each task.

3 Methodology

We propose an algorithm based on block coordinate descent to alternatively update $\{\Theta^k\}_{k=1}^K$ and U until convergence. Subproblems then are solved by the Augmented Lagrangian Method (ALM).

Update Θ^k: To update $\{\Theta^k\}_{k=1}^K$, with U fixed, we can decompose the Problem 2.1 into K independent parts (suppressing superscript k for simplicity):

$$\text{argmin}_\Theta -l(S, \Theta) + \lambda_1 \|\Theta\|_1 + \lambda_2 D_{ld}(\Theta, \Omega) = \text{argmin}_\Theta -l(\tilde{S}, \Theta) + \frac{\lambda_1}{1+\lambda_2} \|\Theta\|_1, \quad (3.1)$$

which is a Graphical Lasso problem with a scaled and shifted covariance matrix

$$\tilde{S} = \frac{1}{1+\lambda_2} \left[S + \lambda_2 \left(\epsilon I + A U U^T A \right) \right].$$

Algorithm 1 Multi-task Attributed Graphical Lasso (Problem 2.1)

Require: $\{(S^k, A^k)\}_{k=1}^K, d, \lambda_1, \lambda_2, \lambda_3 > 0, \epsilon = 0.01, \rho_0 = 2, \gamma = 1.05$
Randomly initialize U
repeat
  Solve Problem 3.1 for $\{\Theta^k\}_{k=1}^K$
  Initialize $Y^k = \mathbf{0}, \rho = \rho_0$
  repeat
    Solve Problems 3.3 and 3.4 for $\{Z^k\}_{k=1}^K$
    Solve the linear system 3.5 for U
    Update $Y^k := Y^k + \rho(Z^k - \tilde{A}^k U)$
    Update $\rho := \gamma \cdot \rho$
  until convergence
until convergence
return $\{\Theta^k\}_{k=1}^K, U$

This problem can be seen as a "supervised" Graphical Lasso since the LogDet term hopes two distributions to be similar. Since $\tilde{S}$ is positive semi-definite, Problem 3.1 can be solved by most classical Graphical Lasso solvers efficiently [2, 15, 21, 29].

Update U: The Problem 2.1 with $\{\Theta^k\}_{k=1}^K$ fixed can be re-organized into

$$\begin{aligned}\min_{U} \quad & \sum_{k=1}^{K}[-\log\det(I + \frac{A^k}{\sqrt{\epsilon}}UU^T\left(\frac{A^k}{\sqrt{\epsilon}}\right)^T) + \mathrm{Tr}\left((A^k)^T\Theta^k A^k UU^T\right)] \\ & + \frac{\lambda_3}{2\lambda_2}\|U\|_F^2.\end{aligned} \tag{3.2}$$

Though Problem 3.2 is not convex, we could use the Augmented Lagrangian Method (ALM) to solve it effectively. It can then be rewritten as

$$\begin{aligned}\min_{\substack{Z^k, U \\ 1\leqslant k\leqslant K}} \quad & \sum_{k=1}^{K}\left[-\log\det\left(I + Z^k(Z^k)^T\right) + \mathrm{Tr}\left(H^k UU^T\right)\right] + \frac{\lambda_3}{2\lambda_2}\|U\|_F^2, \\ \text{s.t.} \quad & Z^k = \tilde{A}^k U,\end{aligned}$$

where $Z^k \in \Re^{p\times d}$ are auxiliary variables, and $H^k = (A^k)^T\Theta^k A^k$, $\tilde{A}^k = \frac{A^k}{\sqrt{\epsilon}}$. The augmented Lagrangian function is given by

$$\begin{aligned}&\mathcal{L}_\rho\left(U, \{Z^k\}, \{Y^k\}\right) \\ =&\sum_{k=1}^{K}\left[-\log\det\left(I + Z^k(Z^k)^T\right) + \mathrm{Tr}\left(H^k UU^T\right)\right. \\ &\left. + \mathrm{Tr}\left((Y^k)^T(Z^k - \tilde{A}^k U)\right) + \frac{\rho}{2}\|Z^k - \tilde{A}^k U\|_F^2\right] + \frac{\lambda_3}{2\lambda_2}\|U\|_F^2,\end{aligned}$$

where $\rho > 0$ is a penalty parameter and $Y^k \in \Re^{p\times d}$ are dual variables. Solving Problem 3.2 is equivalent to minimizing $\mathcal{L}_\rho\left(U, \{Z^k\}, \{Y^k\}\right)$ with a sufficiently large ρ. In practice, we minimize $\{\mathcal{L}_{\rho_t}\}_{t=0}^{\infty}$ iteratively with a monotonic increasing sequence $\{\rho_t\}_{t=0}^{\infty}$ satisfying $\lim_{t\to\infty} \rho_t \to \infty$.

Given the initial $U_0, Z_0^k, Y_0^k, \rho_0$, we do the following block coordinate updates:

Step 1: Compute optimal $\{Z_{t+1}^k\}$ with U_t and $\{Y_t^k\}$ fixed. The $\mathcal{L}_{\rho_t}$ is separable w.r.t Z^k, so minimizing $\mathcal{L}_{\rho_t}$ over Z^k takes the form (suppressing k) of

$$\underset{Z}{\operatorname{argmin}} - \log\det(I + ZZ^T) + \frac{\rho_t}{2}\|Z - (\tilde{A}U_t - \frac{1}{\rho_t}Y_t)\|_F^2. \tag{3.3}$$

The above problem can be converted to a set of scalar minimization problems using the following theorem [17]:

Theorem 1. *For unitarily invariant function $F(Z) = f \circ \sigma(Z)$, assuming the singular value decomposition of $R \in \Re^{p\times d}$ is $R = U\Sigma_R V^T$, $\Sigma_R = diag(\{\sigma_{R,i}\}_{i=1}^{\min(p,d)})$, the optimal solution to the problem*

$$\min_Z \; F(Z) + \frac{\rho}{2}\|Z - R\|_F^2$$

is $Z^\star = U\Sigma_Z^\star V^T$, with $\Sigma_Z^\star = diag(\{\sigma_i^\star\}_{i=1}^{\min(p,d)})$ obtained by solving scalar minimization problems

$$\sigma_i^\star = \underset{x}{\operatorname{argmin}}\, f(x) + \frac{\rho}{2}(x - \sigma_{R,i})^2, \;\; i = 1, \ldots, \min(p,d). \tag{3.4}$$

Since $F(Z) = -\log\det(I + ZZ^T) = -\sum_{i=1}^{\min(p,d)} \log(1 + \sigma_{Z,i}^2)$, $F(Z)$ is a unitarily invariant function with $f(\sigma_{Z,i}) = -\log(1 + \sigma_{Z,i}^2)$, where $\sigma_{Z,i}$ is the i-th singular value of Z. By checking the gradient equation of Problem 3.4, we can find that the optimal $\sigma_i^\star$ is the non-negative root of the cubic equation:

$$g(x) = x^3 - \sigma_i x^2 + (1 - \frac{2}{\rho_t})x - \sigma_i = 0,$$

where $\sigma_i \geq 0$ is the i-th singular value of $\tilde{A}U_t - \frac{1}{\rho_t}Y_t$. Observe that there exists at least one non-negative root. Besides, by checking the discriminant of the cubic equation, we can find that the equation $g(x) = 0$ only has one real root if $\rho_t \geq 2$ or a triple root 0 if $\rho_t = 2$ and $\sigma_i = 0$. Therefore, Problem 3.4 has a unique optimum if $\rho_t \geq 2$, so does Problem 3.3.

Step 2: Compute optimal U_{t+1} with $\{Z_{t+1}^k\}, \{Y_t^k\}$ fixed. The gradient equation is

$$\left[\sum_k \left(2H^k + \rho_t(\tilde{A}^k)^T\tilde{A}^k\right) + \frac{\lambda_3}{\lambda_2}I\right] U_{t+1} = \sum_k \left[(\tilde{A}^k)^T(Y_t^k + \rho_t Z_{t+1}^k)\right]. \tag{3.5}$$

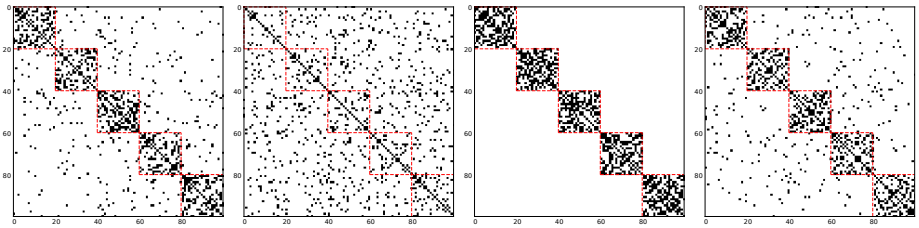
Thus the optimal U_{t+1} can be solved from this linear system.

Step 3: Update the dual variables:

$$Y_{t+1}^k := Y_t^k + \rho_t(Z_{t+1}^k - \tilde{A}^k U_{t+1}), \;\; \forall k.$$

Step 4: Update the penalty parameter $\rho_{t+1} = \gamma \cdot \rho_t$, where $\gamma > 1$.

The algorithm for MAGL is summarized in Algorithm 1.



(a) GroundTruth (b) FMGL (c) PathGLasso (d) MAGL

Fig. 3: The precision matrices learned by three comparing methods.

4 Experiments

4.1 Data Collection

We evaluate the proposed method on real-world datasets and synthetic datasets:
• **DBLP** is a subset of a bibliographical network. Following settings in [23], we extracted 20 conferences and top-5000 authors among 4 areas from 2006 to 2015. After removing stop words in paper titles, we get 679 frequent terms as the vocabulary to generate bag-of-words representations as authors' activities. Here we assume the life cycle of each author is 5 years, *i.e.*, the length of the PhD program. Given a year, each author is accompanied by a one-hot attribute vector of length 5, which indicates the stage he was in. The tasks are to estimate connections among authors in each year.
• **AML** contains two groups of gene expression levels of AML (acute myeloid leukemia) studies [8, 11] used in [9]. Each gene is categorized into at least one pathway, which is used as its attributes. Specifically, the j-th attribute of the i-th gene $A_{ij} = 1$ if the gene is in the j-th pathway, otherwise $A_{ij} = 0$.

The generative method of synthetic data is as follows: given the number of tasks K, the number of variables p, the number of observations n, and the number of classes m, first we generate variables' classes in two ways:
• **Dataset-1 (Ordered)**: We assign a random integer $c_i^1 \in \{c \in \mathbb{N} | -\lfloor m/2 \rfloor \leqslant c \leqslant m\}$ to each variable as its class in the first task. For the k-th ($k > 1$) task, the i-th variable's class is randomly picked in the set $c_i^k \in \{c_i^{k-1}, c_i^{k-1} + 1\}$.
• **Dataset-2 (Unordered)**: $c_i^k \in \{c \in \mathbb{N} | 1 \leqslant c \leqslant m\}$ is always randomly picked for all tasks. The i-th variable's attribute vector in the k-th task a_i^k is a vector of all zeros, except that the c_i^k-th element is 1 if $1 \leqslant c_i^k \leqslant m$. The element of a precision matrix $\left(\Sigma^k\right)_{ij}^{-1}$ is nonzero with the probability $\frac{(4-t)p}{\sum_{u,v} \mathbb{1}\{\delta_{uv}^k = t\}}$ if $\delta_{ij}^k = t \in \{0, 1, 2\}$, otherwise $\frac{p}{\sum_{u,v} \mathbb{1}\{\delta_{uv}^k \geq 3\}}$, where $\delta_{ij}^k = |c_i^k - c_j^k|$. By this means, the number of nonzero off-diagonal elements in each precision matrix is about $10p$. We calculate the sample covariance matrix S^k using n samples.

Dataset-1 simulates the case that there is a natural order among multiple tasks, and tasks share a common set of variables, while Dataset-2 does not assume the identical variable sets nor orderliness among tasks.

4.2 Compared Methods

To validate the effectiveness of our proposal, we test the following methods: (1) **GLasso** [7] is the vanilla Graphical Lasso. We fit a GLasso model for each task separately. (2) **PathGLasso** [9] takes a sample covariance matrix and a set of pathways as input. It assumes that a pair of variables will not be connected if they do not participate together in any pathways. We fit a PathGLasso model independently for each task. (3) **FMGL** [27] jointly estimates multiple tasks of Graphical Lasso using a sequential fused ℓ_1 penalty for adjacent precision matrices. It requires that the tasks have a natural order. (4) **JGL** [3] jointly estimates multiple tasks of Graphical Lasso under the assumption that all graphs have similar non-zero patterns by using fused penalty or group lasso penalty. (5) **MAGL** is our proposal, which makes use of attributes and jointly estimates multiple tasks. All comparing methods have a parameter λ_1 for the ℓ_1 penalty. FMGL and JGL have an extra parameter λ_2 to weight the penalty terms. MAGL uses λ_2 to weight the LogDet divergence term and λ_3 for regularization.

4.3 Experiment Settings

To test whether these methods can correctly recover the nonzero patterns and fit the data distributions, we use F1 score and Relative Log-likelihood as the evaluation metrics. The larger the value, the better the performance.

To ensure a fair comparison, the parameter λ_1 is searched using the bisection technique to make the number of edges in the estimated graphs approximately equal to the number of edges in the true graphs. The λ_2 for FMGL, JGL and MAGL is determined by cross validation. Besides, for MAGL, we simply let the dimension of the output space of projection $d = 100$, and the regularization parameter $\lambda_3 = 1$ throughout the experiments. Other default values for algorithm parameters are: $\epsilon = 0.01, \rho_0 = 2, \gamma = 1.05$.

4.4 Experiment Results

Following the settings in literatures [3, 27, 28], we only show numerical results on synthetic datasets, since ground truth in real datasets is hard to obtain. For example, the network structure in DBLP does not correspond to its Gaussian graphical model. Case studies on real-world datasets are conducted instead.

We first summarize our findings on synthetic data. We show the averaged result of 5 runs with different random seeds for each experiment.

Before we show the quantitative results, we manually generate a toy example and show the learned Θ^1 in Figure 3. Because FMGL, GLasso and JGL show the similar patterns, we only show the result of FMGL. We can see that in the ground truth, most non-zero elements appear on the diagonal blocks. FMGL cannot capture the block structures, and thus performs poorly. For PathGLasso, pathway constraints are employed so that non-zeros elements on off-diagonal blocks are not allowed. Our proposal, MAGL, learns a precision matrix that is closest to the ground truth because block structures are revealed by finding a

Table 1: Results on Dataset-1.

			F1↑					Log-likelihood (%)↑				
K	m	p	GLasso	PathGLasso	FMGL	JGL	MAGL	GLasso	PathGLasso	FMGL	JGL	MAGL
5	3	500	0.3375	0.3375	0.3408	0.3419	**0.3423**	1.2412	1.2412	1.2502	1.2426	**1.2546**
		1000	0.4256	0.4321	0.4302	0.4297	**0.4372**	1.1843	1.2264	1.1977	1.1913	**1.2268**
	5	500	0.3325	0.3477	0.3365	0.3321	**0.3524**	1.2116	1.2234	1.2222	1.2086	**1.2416**
		1000	0.4222	0.4396	0.4286	0.4286	**0.4410**	1.1848	1.2020	1.1977	1.2004	**1.2235**
	10	500	0.3248	**0.3375**	0.3280	0.3273	0.3345	1.2421	1.2647	1.2554	1.2517	**1.2863**
		1000	0.4124	0.4301	0.4167	0.4138	**0.4352**	1.2243	1.2783	1.2348	1.2274	**1.2875**
10	3	500	0.3434	0.3435	0.3499	0.3499	**0.3543**	1.1182	1.1181	1.1430	1.1427	**1.1441**
		1000	0.4192	0.4282	0.4264	0.4236	**0.4425**	1.1201	1.1465	1.1442	1.1305	**1.1523**
	5	500	0.3381	**0.3526**	0.3447	0.3426	0.3521	1.1191	1.1450	1.1401	1.1330	**1.1487**
		1000	0.4012	0.4105	0.4100	0.4081	**0.4237**	1.1457	1.1683	1.1732	1.1611	**1.1795**
	10	500	0.3401	0.3532	0.3448	0.3430	**0.3629**	1.2343	1.2394	**1.2481**	1.2398	1.2450
		1000	0.3877	0.3970	0.3923	0.3891	**0.4075**	1.2342	1.2653	1.2514	1.2410	**1.2828**

Table 2: Results on Dataset-2.

			F1↑			Log-likelihood (%)↑		
K	m	p	GLasso	PathGLasso	MAGL	GLasso	PathGLasso	MAGL
5	3	500	0.3318	0.3318	**0.3576**	1.2449	1.3013	**1.3143**
		1000	0.4170	0.4341	**0.4346**	1.1842	1.2182	**1.2216**
	5	500	0.3300	0.3378	**0.3401**	1.2247	1.2348	**1.2553**
		1000	0.4136	0.4372	**0.4448**	1.1704	1.2361	**1.2642**
	10	500	0.3237	**0.3306**	0.3294	1.2334	1.2456	**1.2877**
		1000	0.4073	0.4249	**0.4293**	1.2260	1.2774	**1.2831**
10	3	500	0.3404	0.3404	**0.3541**	1.1304	1.1414	**1.1568**
		1000	0.4117	0.4188	**0.4332**	1.1217	1.1451	**1.1532**
	5	500	0.3321	0.3391	**0.3448**	1.1079	1.1101	**1.1200**
		1000	0.3986	0.4027	**0.4163**	1.1478	1.1698	**1.1741**
	10	500	0.3350	0.3516	**0.3606**	1.2275	1.2797	**1.2859**
		1000	0.3785	0.3999	**0.4003**	1.2280	1.2331	**1.2453**

projection matrix across tasks. Since MAGL does not constrain non-zero patterns, elements on off-diagonal blocks are also successfully recovered.

Our first set of experiments are conducted on the synthetic Dataset-1. The results are shown in Table 1. As we can see, MAGL performs well in most cases. This is because our proposal considers the relations between attributes of variables and linkage structures among variables, and shares information across tasks to improve the quality of estimates. FMGL does not perform well due to the fact that the sequential fused ℓ_1 penalty only considers the values in the adjacent precision matrices and may hardly capture the global property. Another multi-task method, JGL, performs worse than GLasso in some cases, due to the inappropriate assumption, *i.e.*, similar non-zeros patterns across tasks. We can

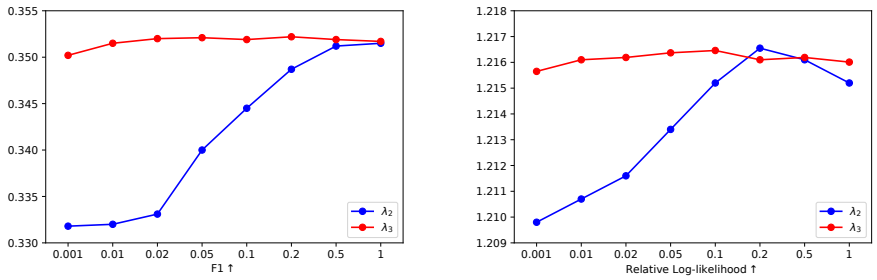
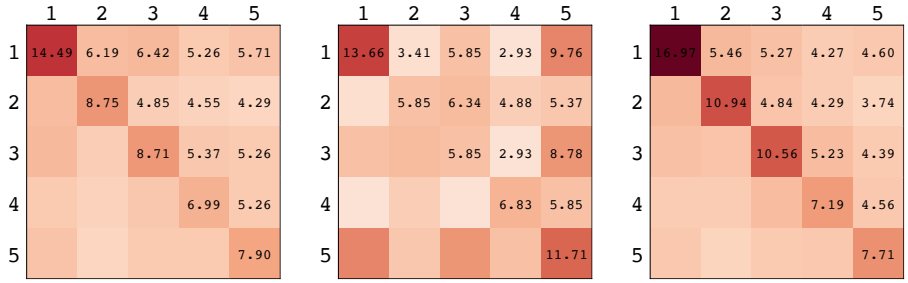


Fig. 4: The influence of $\lambda_{2,3}$.



(a) DBLP (b) GLasso (c) MAGL

Fig. 5: Co-author patterns on DBLP. The number in cell (i, j) indicates how often co-author activities happened between authors in stage-i and stage-j.

notice that due to the generative methods of datasets and pathways, by excluding a large number of impossible edges, PathGLasso gains a huge advantage when m is large w.r.t F1. Nevertheless, as illustrated in Figure 3, missing elements on off-diagonal blocks lower the log-likelihood scores.

The experimental results on Dataset-2 are shown in Table 2, which reveals the similar patterns. FMGL and JGL are not tested here because the sets of variables are not the same in different tasks and there is no oder among them.

•**Parameter Study:** In this subsection, we test the performance of MAGL under different λ_2 and λ_3. The results are shown in Figure 4. We can see that MAGL is robust w.r.t. λ_3. The performance is also stable w.r.t. λ_2 in a wide range. Specifically, as λ_2 grows, the F1 score increases as well but after some point, the log-likelihood decreases slightly. Recall that Problem 3.1 uses a scaled and shifted covariance matrix, and hence a large λ_2 may skew the data distribution and harm the likelihood.

4.5 Case Study

We also apply MAGL to the DBLP and AML datasets. Because of lack of the ground truth, we only show the results qualitatively.

For DBLP dataset, we count the number of co-author activities happened in different stages and show them in Figure 5(a). For example, about 14.49% co-authors activities are between authors who are both in the Stage-1, *i.e.*, the 1st year of PhD. We apply MAGL and GLasso on the dataset and count the number of edges in the learned graphs. From Figures 5(b) and 5(c) we can see that, with the help of authors' attributes (*i.e.*, life stages), MAGL reveals co-author patterns better.

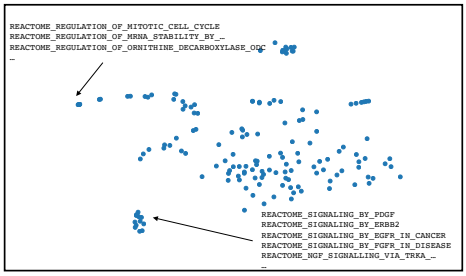


Fig. 6: Visualization of projection matrix on AML dataset.

In AML dataset, since the attributes of variables (genes) are pathways, the i-th row of U can be viewed as the latent features of the i-th pathway. We map U into a 2-dimensional plane using t-SNE [20] and show it in Figure 6. As we can see, points are clustered. Take a closer look and we find that the points in bottom-left corner correspond to pathways in which genes are involved in Signaling, and top-left points are pathways involved in Regulation, which means MAGL could make use of attributes properly to help it improve the performance.

5 Related Work

To obtain a sparse and meaningful estimate of the precision matrix, numerous researchers have considered the ℓ_1 penalized minimum negative log-likelihood estimation problem [1,5–7], *i.e.*, Graphical Lasso. A bunch of algorithms [2,15,21,29] have also been developed. However, most of these methods suffer from intensive computation.To make Graphical Lasso applicable in large problems, [26] and [21] derived a necessary and sufficient condition that a GLasso problem can be decomposed into several smaller sized and independent problems. Further, pathway Graphical Lasso [9] provides an efficient framework dealing with overlapping blocks.Based on pathway Graphical Lasso, [30] uses a related heterogeneous information network to provide different types of "pathways" and learn a graph with multiple types of edges.

Recently, there are some prior works on multi-task Graphical Lasso that learn multiple precision matrices simultaneously for related tasks. These methods differ in the choice of penalty functions: [14] suggested to estimate multiple Graphical Lasso by replacing the ℓ_1 norm with an $\ell_{1,\infty}$ norm. [10] proposed a non-convex hierarchical penalty. [12, 13, 19] assumed that there are common (sub)structures among multiple graphs. [3] estimated multiple precision matrices jointly using a pairwise fused penalty or grouping penalty. [27] considered the case that multiple tasks have a natural order and proposed a sequential fused penalty. A necessary and sufficient condition for the graphs to be decomposable is also given. [22] proposed a method on the assumption that the network differences are introduced from node perturbations. Different from the aforementioned methods that inspected the values in precision matrices, [24] utilized the structure information directly. However, these methods all require that the sets of variables are the same among tasks. Besides, they only focus on the variables' observations and cannot deal with attributed graphs.

6 Conclusion

In this paper, we incorporate variables' attributes into the framework of multi-task Graphical Lasso, and propose Multi-task Attributed Graphical Lasso (MAGL). We introduce the LogDet divergence to bridge graphs structures and attributes so that information could be shared across multiple tasks. The experiments on synthetic datasets show the effectiveness of MAGL, and the case studies demonstrated that our method can produce a meaningful result. As for future work, we could try other ways to connect variables' observations and attributes. Besides, we will consider applying our proposal to more real world applications.

References

1. Banerjee, O., Ghaoui, L.E., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. JMLR **9**(Mar), 485–516 (2008)
2. Cai, T., Liu, W., Luo, X.: A constrained l1 minimization approach to sparse precision matrix estimation. JASA **106**(494), 594–607 (2011)
3. Danaher, P., Wang, P., Witten, D.M.: The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Series B Stat Methodol **76**(2), 373–397 (2014)
4. Davis, J.V., Dhillon, I.S.: Differential entropic clustering of multivariate gaussians. In: NeurIPS. pp. 337–344 (2007)
5. Duchi, J.C., Gould, S., Koller, D.: Projected subgradient methods for learning sparse gaussians. In: UAI (2008)
6. Fan, J., Liao, Y., Liu, H.: An overview of the estimation of large covariance and precision matrices. ECONOMET J **19**(1) (2016)
7. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

8. Gentles, A.J., Plevritis, S.K., Majeti, R., Alizadeh, A.A.: Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. JAMA **304**(24), 2706–2715 (2010)
9. Grechkin, M., Fazel, M., Witten, D., Lee, S.: Pathway graphical lasso. In: AAAI. pp. 2617–2623 (2015)
10. Guo, J., Levina, E., Michailidis, G., Zhu, J.: Joint estimation of multiple graphical models. Biometrika **98**(1), 1–15 (2011)
11. Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Béné, M., De V, J., Hernández, J.M., Hofmann, W., Mills, K.I., et al.: Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. Int. J. Clin. Oncol. **28**(15), 2529–2537 (2010)
12. Hara, S., Washio, T.: Common substructure learning of multiple graphical gaussian models. In: ECMLPKDD. pp. 1–16 (2011)
13. Hara, S., Washio, T.: Learning a common substructure of multiple graphical Gaussian models. Neural Networks **38**, 23–38 (2013)
14. Honorio, J., Samaras, D.: Multi-task learning of gaussian graphical models. In: ICML. pp. 447–454 (2010)
15. Hsieh, C., Sustik, M.A., Dhillon, I.S., Ravikumar, P.D.: QUIC: Quadratic approximation for sparse inverse covariance estimation. JMLR **15**(1), 2911–2947 (2014)
16. Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., Reiman, E., Initiative, A.D.N., et al.: Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. NeuroImage **50**(3), 935–949 (2010)
17. Kang, Z., Peng, C., Cheng, J., Cheng, Q.: Logdet rank minimization with application to subspace clustering. COMPUT INTEL NEUROSC **2015**, 68 (2015)
18. Kulis, B., Sustik, M., Dhillon, I.: Learning low-rank kernel matrices. In: ICML. pp. 505–512 (2006)
19. Lee, W., Liu, Y.: Joint estimation of multiple precision matrices with common structures. JMLR **16**(1), 1035–1062 (2015)
20. Maaten, L.V.D., Hinton, G.: Visualizing data using t-sne. JMLR **9**(Nov), 2579–2605 (2008)
21. Mazumder, R., Hastie, T.: The graphical lasso: New insights and alternatives. EJS **6**, 2125 (2012)
22. Mohan, K., London, P., Fazel, M., Witten, D., Lee, S.: Node-based learning of multiple gaussian graphical models. JMLR **15**(1), 445–488 (2014)
23. Sun, Y., Han, J., Gao, J., Yu, Y.: itopicmodel: Information network-integrated topic modeling. In: ICDM. pp. 493–502 (2009)
24. Tao, Q., Huang, X., Wang, S., Xi, X., Li, L.: Multiple Gaussian graphical estimation with jointly sparse penalty. Signal Processing **128**, 88–97 (2016)
25. Whittaker, J.: Graphical models in applied multivariate statistics. Wiley Publishing (2009)
26. Witten, D.M., Friedman, J.H., Simon, N.: New insights and faster computations for the graphical lasso. J COMPUT GRAPH STAT **20**(4), 892–900 (2011)
27. Yang, S., Lu, Z., Shen, X., Wonka, P., Ye, J.: Fused multiple graphical lasso. SIOPT **25**(2), 916–943 (2015)
28. Yin, H., Liu, X., Kong, X.: Coherent graphical lasso for brain network discovery. In: ICDM (2018)
29. Yuan, X.: Alternating direction methods for sparse covariance selection. Optimization Online (2009)
30. Zhang, Y., Xiong, Y., Liu, X., Kong, X., Zhu, Y.: Meta-path graphical lasso for learning heterogeneous connectivities. In: SDM. pp. 642–650 (2017)