Signal Processing 128 (2016) 88–97

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Multiple Gaussian graphical estimation with jointly sparse penalty ☆

Qinghua Tao [a], Xiaolin Huang [b,c], Shuning Wang [a], Xiangming Xi [a], Li Li [a,*]

[a] TNList, Department of Automation, Tsinghua University, Beijing 100084, PR China
[b] Pattern Recognition Lab of the Friedrich-Alexander-Universität, Erlangen-Nürnberg 91058, Erlangen, Germany
[c] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200400, PR China

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form
14 February 2016
Accepted 14 March 2016
Available online 31 March 2016

Keywords:
Gaussian graphical models
Non-convex penalty
Block coordinate descent
Majorization and minimization
Re-weighted algorithm

ABSTRACT

In this paper, we consider estimating multiple Gaussian graphs with a similar sparsity structure. Most related solving methods, such as GGL (Group graphical lasso) and FMGL (Fused multiple graphical lasso), focus on the information of the edge values, and pay few attention to the estimation based on structure information. We construct a jointly sparse penalty to encourage graphs to share a similar sparsity structure by utilizing information of the common structure across the graphs. The new objective function is neither convex nor differentiable. Combining block coordinate descent and majorization–minimization strategies, we derive a new re-weighed algorithm to solve the problem by transforming the subproblems in every iteration into convex ones. Experimental results show that the proposed algorithm outperforms FMGL and GGL when the sparsity structure is similar but the edge values are not.

1. Introduction

The relationship among multiple variables can be revealed by an undirected graph, where each of the nodes equals a variable (feature), and the dependence between any two edges of the graph can be described by the elements of a matrix. Due to this character, the estimation of undirected graphical models can be applied in many fields, such as computer vision, finance, social networks and bio-informatics [1–4]. One typical instance is the analysis of gene expression data. Genes tend to work in groups based on their biological functions, and there are some regulatory relations among genes [5]. Such biological knowledge can be expressed as a graph, where nodes are the genes, and edges represent the regulatory relationships. Graphical models provide a useful tool for modeling these relations and exploring gene activities [6].

In this paper, we focus on the undirected Gaussian graphical models (GGMs), which is the most popular undirected graphs. In GGMs, all the samples (observations) follow Gaussian distributions. Assume that a Gaussian graph has p nodes, each of which represents a distinctive variable. Denote the variable vector as Y and the inverse of the covariance matrix (precision matrix) as $\mathbf{\Theta}$, where $Y \in \mathbb{R}^p$ and $\mathbf{\Theta} \in \mathbb{R}^{p\times p}$. With all other variables given, any two variables y_i and y_j are conditionally independent, if and only if the corresponding element $[\mathbf{\Theta}]_{ij}$ is zero. Since the precision matrix can describe the relationship among the variables, the problem of learning GGMs is equivalent to estimating the precision matrix [7–9].

Many methods have been proposed to estimate the precision matrices of GGMs. Minimizing the negative log-likelihood function is most popular, since it can lead to an unconstrained convex problem with respect to the precision matrices. Denote the data set as $\mathbf{X} \in \mathbb{R}^{n\times p}$, where n is the number of the observations. Thus, each row in the $\mathbf{X}$ represents an observation. Denote $\mathbf{S}$ as the empirical covariance matrix of $\mathbf{X}$, where $\mathbf{S} = \mathbf{X}^T\mathbf{X}/n$. The problem of minimizing the negative log-likelihood function is shown as follows:

$$\min_{\mathbf{\Theta}} -\log\det\mathbf{\Theta} + \mathrm{trace}(\mathbf{S\Theta}). \tag{1}$$

Minimizing the negative log-likelihood function equals the maximum likelihood estimate (MLE), which always yields to a dense solution [6,10,11]. However, the precision matrix is naturally sparse in practical applications, since there are many conditionally independent pairs of nodes in a graph. Therefore, MLE is far from satisfactory [6].

The key problem in this field comes down to identifying the sparsity structure of the precision matrix $\mathbf{\Theta}$. As a natural choice, the l_1 regularization is employed by numerous researchers to induce sparsity. In 2006, Meinshausen proposed a penalized regression approach to achieve sparsity [12], which is extended by Peng [13]. Yuan and Lin [14], Friedman et al. [15], Banerjee et al. [7] and Rothman et al. [16] discussed penalized log-likelihood

☆This project is jointly supported by the National Natural Science Foundation of China (61473165, 61134012), the National Basic Research Program of China (2012CB720505), and the Alexander von Humboldt Foundation.

* Corresponding author.
E-mail addresses: taoqh14@mails.tsinghua.edu.cn (Q. Tao), huangxl06@mails.tsinghua.edu.cn (X. Huang), swang@mail.tsinghua.edu.cn (S. Wang), xxm10@mails.tsinghua.edu.cn (X. Xi), li-li@mail.tsinghua.edu.cn (L. Li).

http://dx.doi.org/10.1016/j.sigpro.2016.03.009
0165-1684/

approaches instead, which aim to solve the following problem:

$$\min_{\mathbf{\Theta}} -\log\det(\mathbf{\Theta}) + \mathrm{trace}(\mathbf{S\Theta}) + \lambda \| \mathbf{\Theta} \|_1, \tag{2}$$

where λ is a non-negative penalty parameter, and $\| \mathbf{\Theta} \|_1$ denotes the sum of the absolute values of the off-diagonal elements of $\mathbf{\Theta}$ [17]. In general, a larger penalty parameter λ results in more zero elements in the precision matrix, which leads to sparsity to the solution [18].

However, the methods mentioned above assume that the observations are independently drawn from a single Gaussian distribution, which is unreasonable. In many practical data sets, observations come from different distributions [19]. For example, suppose that we have a collection of gene expression measurement samples from cancer patients and healthy people. In order to estimate the graphical models for the cancer samples and the normal samples, one would expect the two graphical models to be similar in structure to each other, but may also have differences stemming from the dysregulation of gene expression in cancer [17]. Estimating the two graphs separately is unable to utilize the information across the graphs to exploit the similarities between the graphs, while just estimating a single graph for both of the classes contradicts with the fact that the true graphs are not expected to be identical. For multiple GGMs, graphs share a similar sparsity structure, but maybe to have distinctive edges stemming from individual differences. In such cases, the joint estimation can borrow strength across different graphs to reveal the common structure shared by the graphs and the differences among them. It can also reduce the variance of the estimation.

From the perspective of multi-view learning, multiple GGMs can be regarded as a special case of it. Each Gaussian graph equals a view. In multi-view graph learning, views may be obtained from multiple sources or different features subsets [20]. Similarly, in multiple GGMs, each graph represents the activities of one group. Both multi-view graph learning and multiple GGMs borrow information across each graph to better solve the optimization problem [21–24].

To achieve joint estimation, various penalty functions are proposed to estimate the common structure shared by multiple Gaussian graphs [8,6,25]. The details of the existing penalties will be discussed in Section 2.1. These penalties hold desirable estimation accuracy only for the cases where the corresponding elements in each precision matrix are close enough. However, in general cases, the values of the corresponding edges in each Gaussian graphs come from different distributions, and even possibly deviate greatly from each other. The existing penalties focus on the edge values, and pay few attention to utilizing the information of the structure. Thus, the structure is difficult to be accurately estimated merely with the information of the edge values in the general cases. From this motivation, we construct a jointly sparse penalty, which utilizes the structure information across the graphs to deal with the general cases.

In this paper, we address the general cases where all Gaussian graphs share a similar structure, but the values of the same edges in each graph can be quite different. The corresponding nonzero elements in each precision matrix are allowed to have different signs and even from different distributions. In general cases, we consider GGMs with a similar sparsity structure, and we also allow small differences among the precision matrices. Since GGL and FMGL have limitations for the general cases, we consider the problem of simultaneously estimating GGMs via a new penalized log-likelihood approach. In the new model, we construct a jointly sparse penalty function to enhance sparsity as well as to encourage the graphs to share a similar structure. One remarkable point is that we can maintain the differences of corresponding edge values among the graphs when estimating the similarities in structure. Our model yields to a non-convex and non-differential problem. To solve the problem, we combine the block coordinate descent algorithm with the majorization-minimization algorithm to derive a new re-weighted algorithm, where subproblems in each iteration are convex. This kind of method was first proposed in [26], and it was shown to be effective for sparse multiple measurement problems.

The rest of the paper is organized as follows. In Section 2, we present the basic problem formulation, and introduce two popular method FMGL and GGL. In Section 3, we construct the jointly sparse penalty and derive a new re-weighted algorithm to solve the new model. In Section 4, a series of numerical experiments are conducted to present the overall performance of the proposed algorithm. Section 5 concludes the whole paper.

2. Problem formulation

2.1. Penalized MLE approach

Suppose that there are K distinctive Gaussian graphs, from which we are given K data sets $\mathbf{X}^i \in \mathbb{R}^{n_i \times p}$, $i = 1, \ldots, K$, where n_i is the number of observations and p is the number of variables. The p variables are the same for all K data sets. Moreover, observations in each data set $\mathbf{X}^i$ are independent and identically distributed with Gaussian distribution $N(\mu^i, \mathbf{\Sigma}^i)$, where $\mu^i \in \mathbb{R}^p$, and $\mathbf{\Sigma}^i$ is a symmetric positive definite $p \times p$ matrix. The l_1 penalized MLE only achieves sparsity in solutions, but we also need to learn the common structure shared by all the Gaussian graphs. Therefore, researchers proposed many penalty functions to urge all the graphs to match each other in sparsity structure [27]. In the penalized MLE approach, one usually considers the following problem:

$$\min_{\mathbf{\Theta}} \sum_{k=1}^{K} \left(-\log\det(\mathbf{\Theta}^k) + \mathrm{trace}(S^k\mathbf{\Theta}^k) \right) + P(\mathbf{\Theta}), \tag{3}$$

where $P(\mathbf{\Theta}) = \lambda_1 \sum_k \sum_{i \neq j} |\theta_{ij}^k| + \lambda_2 \tilde{P}(\mathbf{\Theta})$, $\mathbf{\Theta} = [\mathbf{\Theta}^1, \ldots, \mathbf{\Theta}^K]$, and $\theta_{ij}^k = [\mathbf{\Theta}^k]_{ij}$. λ_i is the non-negative penalty parameter.

The $\tilde{P}(\mathbf{\Theta})$ penalty is employed to encourage a similar sparsity structure across the K precision matrices. Yuan and Lin applied group graphical lasso (GGL) penalty to induce a similar pattern of non-zero elements to each $\mathbf{\Theta}^k$ in [8]. Hoefling proposed the fused graphical lasso (FGL) penalty to encourage all graphs to share similar sparsity structure in [25], and Yang et al. extended it to the fused multiple graphical lasso (FMGL) in [6].

2.2. Group graphical lasso and fused multiple graphical lasso

In GGL and FMGL, the penalty functions are considered as follows:

- GGL: $P_G(\mathbf{\Theta}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^k| + \lambda_2 \sum_{i \neq j} \left(\sum_{k=1}^{K} \theta_{ij}^{(k)^2} \right)^{1/2}$,
- FMGL: $P_F(\mathbf{\Theta}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^k| + \lambda_2 \sum_{k=1}^{K-1} \sum_{i \neq j} |\theta_{ij}^k - \theta_{ij}^{k+1}|$.

$P_G(\mathbf{\Theta})$ and $P_F(\mathbf{\Theta})$ share an identical first item, which results in a sparse solution when λ_1 is large enough.

The second item in P_G encourages a similar sparsity pattern across all the graphs, i.e. there will be a tendency for the nonzero elements in the estimated precision matrices to occur in the same places [17]. Since the first item in GGL is l_1 regularization, the proportion of the two items in GGL may be sensitive, where small changes of λ_2/λ_2 may results in huge differences to the results. The second item in P_G penalizes less for those elements whose position are different across the precision matrices. When the sparsity