Computational Statistics and Data Analysis 121 (2018) 89–103

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Joint estimation of multiple Gaussian graphical models across unbalanced classes

Liang Shan [a], Inyoung Kim [b,*]

[a] Department of Biomedical Affairs and Research, Edward Via College of Osteopathic Medicine, USA
[b] Department of Statistics, Virginia Polytechnic Institute and State University, USA

ARTICLE INFO

Article history:
Received 5 October 2016
Received in revised form 8 October 2017
Accepted 30 November 2017
Available online 21 December 2017

Keywords:
Gene network exploration
Joint adaptive graphical lasso
Precision matrix estimation
Unbalanced multi-class

ABSTRACT

The problem of jointly estimating unbalanced multi-class Gaussian graphical models is considered. Most existing methods require equal or similar sample sizes among classes. However, many real applications do not have similar sample sizes. Hence, the joint adaptive graphical lasso, a weighted l_1 penalized approach is proposed for unbalanced multi-class problems. The joint adaptive graphical lasso approach combines information across classes so that their common characteristics can be shared during the estimation process. Regularization is also introduced into the adaptive term. Simulation studies show that the new approach performs better than existing methods in terms of false positive rate, accuracy, Mathews correlation coefficient, and false discovery rate. The advantages of the new approach are also demonstrated using a liver cancer data set.

1. Introduction

In mathematics, a graph is composed of nodes and edges between nodes; the edges can be directed, undirected, or bi-directed. In recent years, graphical models have become popular in investigating networks. For instance, a gene network that is composed of genes and connections among genes can be illustrated by a graph in which genes are represented by nodes and connections are represented by edges. Under the multivariate Gaussian distribution assumption, a graphical model is called a Gaussian graphical model and edges are undirected. The main idea behind inferring a graph from a set of variables of certain samples is to identify an inverse covariance matrix (or precision matrix), elements of which indicate conditional dependency between pairs of variables. Specifically, if the (i, j)th element in a precision matrix is 0, then variables i and j are conditionally independent; otherwise, they are dependent, given all other variables. To illustrate using a gene network again, if the (i, j)th element in a precision matrix is 0, then genes i and j are unconnected; otherwise, they are connected.

One natural way to estimate the precision matrix is to obtain a maximum likelihood estimator (MLE). However, an MLE can hardly generate exact zeros in the estimated precision matrix, which gives us no clues about the conditional dependency among variables. Moreover, under high-dimensional settings in which the number of variables is larger than or equal to the number of samples, the MLE is ill defined. A number of studies have been proposed to obtain a sparse estimate of a precision matrix. Related ideas date back to Dempster (1972), who suggested the idea of setting elements of a precision matrix to zero and provided rules and algorithms, which were illustrated by a simple sample data set. However, Dempster's (1972) approach is computationally expensive except for in very low-dimensional settings. Meinshausen and Bühlmann (2006) proposed neighborhood selection to estimate sparse precision matrices for high-dimensional settings. They firstly fit a regression model using the least absolute shrinkage and selection operator (LASSO), proposed by Tibshirani (1996), for

* Correspondence to: Department of Statistics, Virginia Tech., Blacksburg, VA 24061, USA.
E-mail address: inyoungk@vt.edu (I. Kim).

https://doi.org/10.1016/j.csda.2017.11.009
0167-9473/

each variable, while treating all other variables as predictors. Then, if either the regression coefficient of variables i on j or that of variables j on i is nonzero, the (i, j)th element in the precision matrix is estimated to be nonzero. Yuan and Lin (2007), Friedman et al. (2008) and Rothman et al. (2008) studied penalized likelihood approaches with the l_1 penalty and estimated penalized MLEs using different algorithms. By doing this, model selection and parameter estimation were simultaneously achieved. Yuan and Lin (2007) used the determinant maximization (MAXDET) algorithm. Friedman et al. (2008) took advantage of the clockwise coordinate descent approach and developed the graphical lasso (Glasso) algorithm, which is remarkably fast. Rothman et al. (2008) derived the optimization algorithm using Cholesky decomposition and the local quadratic approximation to produce the sparse estimator which is permutation invariant. Nonetheless, it has been shown that the LASSO method produces biases in regression. To correct biases, the smoothly clipped absolute deviation (SCAD) penalty and the adaptive lasso were proposed by Fan and Li (2001) and Zou (2006), respectively. Fan et al. (2009) employed the two penalties above in precision matrix estimation and solved the bias problem.

However, all of these approaches ignore the fact that observations may come from different classes. Since true precision matrices may have some differences among classes, assuming that observations all come from the same multivariate normal distribution is inappropriate. On the other hand, classes are related to each other in certain ways, so the network structures of different classes may have something in common. For instance, patients with different types of diabetes (Type 1 diabetes, Type 2 diabetes, and gestational diabetes) may have different gene network structures, but parts of the structures may be exactly the same because they are all from diabetes patients. In this situation, it is inappropriate to estimate the precision matrix by viewing all the observations as one group, since doing so ignores the distinctions among classes. Separately estimating precision matrices with respect to each class fails to take advantage of the common structure among classes. Therefore, jointly estimating precision matrices across multiple classes will take advantage of using information across classes, so that the common structure can be estimated more precisely than when using separate estimation, and unique structures can be found as well. Guo et al. (2011) proposed jointly estimating precision matrices for different classes by re-parameterizing their off-diagonal elements to be multiples of a common factor across categories and using a unique factor for each category. Their method could be solved by iterative weighted Glasso (Friedman et al., 2008). In addition, Danaher et al. (2014) used generalized fused lasso or group lasso as the penalty and employed the alternating directions method of multipliers (ADMM) algorithm to solve the optimization problem. Nevertheless, neither of them considered the problem of unbalanced data, which is common in many real applications. For instance, certain types of cancer are rarely found, so the numbers of samples for those types of cancer and for the normal population are very unbalanced. In those scenarios, the majority class could easily dominate the estimation results when precision matrices are estimated jointly.

Therefore, the goal of this paper is to propose a joint estimation method for multiple Gaussian graphical models across unbalanced classes, with a weighted l_1 penalized approach, so that a common structure is estimated more precisely than by using separate estimations, and unique structures can be discovered.

This article is organized as follows. In Section 2, we propose the weighted penalized likelihood approach. In Section 3, we conduct simulation studies to compare our method with the existing methods. In Section 4, we apply our approaches to the liver cancer data set analyzed by Chen et al. (2002) and de Souto et al. (2008). Section 5 contains our concluding remarks.

2. Joint adaptive graphical lasso approach

In this section, we first explain our model in Section 2.1 and then describe our joint adaptive graphical lasso (JAGL) approach in Section 2.2.

2.1. Unbalanced multi-class Gaussian graphical models

Suppose we have M heterogeneous classes with p variables, where $M \geq 2$. The mth class is expressed as a $n_m \times p$ matrix, which is denoted as X^m where $m = 1, \ldots, M$. Each row of X^m corresponds to an observation, and each column corresponds to a variable. Let $\underline{x}_i^m = (x_{i,1}^m, \ldots, x_{i,p}^m)$ be the ith row of X^m, $i = 1, \ldots, n_m$. With this notation, we write X^m as follows:

$$X^m = \begin{bmatrix} x_{1,1}^m & \cdots & x_{1,p}^m \\ \vdots & \ddots & \vdots \\ x_{n_m,1}^m & \cdots & x_{n_m,p}^m \end{bmatrix} = \begin{bmatrix} \underline{x}_1^m \\ \vdots \\ \underline{x}_{n_m}^m \end{bmatrix} \quad m = 1, \ldots, M.$$

Multiple Gaussian graphical models across unbalanced classes have the following two assumptions:

(i) Within each class m, $\underline{x}_1^m, \ldots, \underline{x}_{n_m}^m \in \mathbb{R}^p$ are i.i.d MN $[\underline{\mathbf{0}}, (\Omega^m)^{-1}]$, where the precision matrix,

$$\Omega^m = \begin{bmatrix} \omega_{1,1}^m & \cdots & \omega_{1,p}^m \\ \vdots & \ddots & \vdots \\ \omega_{p,1}^m & \cdots & \omega_{p,p}^m \end{bmatrix},$$