Journal of Statistical Planning and Inference 190 (2017) 1–14

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Multilevel Gaussian graphical model for multilevel networks

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

[a] Regulatory Statistics Technology Center, Monsanto Company, St. Louis, MO, USA
[b] Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

ARTICLE INFO

Article history:
Received 30 March 2016
Received in revised form 28 October 2016
Accepted 6 May 2017
Available online 20 May 2017

Keywords:
Gaussian graphical model
Graphical LASSO
Pathway
Penalized likelihood
Sparse

ABSTRACT

Gaussian graphical models have become a popular tool to represent networks among variables such as genes. They use the conditional correlations from the joint distribution to describe the dependencies between gene pairs, and employ the precision matrix of the genes. Because of the sparse nature of the gene networks and small sample sizes in high dimensional genetic data, regularization approaches attracted much attention in aim at obtaining the shrinkage estimates of the precision matrix. However, existing methods have been focused on the Gaussian graphical model among genes; that is, they are only applicable to a single level Gaussian graphical model. It is known that pathways are not independent of each other because of shared genes and interactions among pathways. Developing multipathway analysis has been a challenging problem because of the complex dependence structure among pathways. By considering the dependency among pathways as well as the genes within each pathway, we propose a multilevel Gaussian graphical model (MGGM) in which one level describes the networks for genes and the other for pathways. We have developed a multilevel L1 penalized likelihood approach to achieve the sparseness on both levels. In addition, we have developed an iterative weighted graphical LASSO algorithm for MGGM. Our simulation results supported the advantages of our approach; our method estimated the network more accurately on the pathway level and sparser on the gene level. We also demonstrated the usefulness of our approach using a canine genes-pathways data set.

1. Introduction

Gaussian graphical models (Dempster, 1972), known as "covariance selection" or "concentration graph" models, have recently become a popular tool to learn gene association networks. It assumes the nodes (i.e., gene expression data observed in our study) are randomly sampled observational or experimental data from a multivariate Gaussian distribution. That is, let $V = \{v_1, \ldots, v_p\}$ be the set of nodes (genes), and $X_1, \ldots, X_p$ denote the expression data for the p genes; we assume that $(X_1, \ldots, X_p) \sim N(0, \Sigma)$ with positive definite variance–covariance matrix $\Sigma = (\sigma_{ij})$ and precision matrix $\Omega = \Sigma^{-1} = (\omega_{ij})$. Then, the Gaussian graphical model uses the precision matrix Ω as the adjacent matrix (i.e. $\omega_{ij} \neq 0$ implies an association between the gene pair and $\omega_{ij} = 0$ implies no association between the gene pair). A related but completely different concept are the so-called gene "relevance networks", which are based on the covariance matrix Σ. The simple reason why Gaussian graphical models should be preferred over relevance networks for the identification of gene networks is that the off-diagonal elements of Ω are proportional to partial correlations, while the off-diagonal elements of Σ are proportional to marginal correlations. In the latter, interactions are defined through standard correlation coefficients so that missing edges denote

* Corresponding author. Fax: +1 540 231 3863.
E-mail address: inyoungk@vt.edu (I. Kim).

http://dx.doi.org/10.1016/j.jspi.2017.05.003
0378-3758/

marginal independence only. The correlation coefficient is a weak criterion for measuring dependence, because marginally, i.e. directly and indirectly, more or less all genes will be correlated. This implies that zero marginal correlation is in fact a strong indicator for independence. On the other hand, partial correlation coefficients do provide a strong measure of dependence and, correspondingly, offer only a weak criterion of independence as most partial correlation coefficients usually vanish. And more often, with high dimension of genetic data, one would prefer concentrating the network size rather than trapping in a large amount of relevances resulting from relevance networks.

A number of studies have worked on estimating Ω. A popular way to estimate the precision matrix for Gaussian graphical models with small sample modeling is to introduce a penalty to the off-diagonal elements in Ω, which is feasible in computing when $n < p$ and which allows us to estimate the off-diagonal element simultaneously. The sparsity of the obtained precision matrix would be able to take the nature of the genetic networks into account. Due to the small sample size in gene expression data, researchers usually take a penalized log-likelihood approach and solve the following objective function,

$$\max_{\Omega}\big[\log\{\det(\Omega)\} - tr(S\Omega) - \lambda P(\Omega)\big],$$

where λ is a non-negative penalty parameter and $P(\cdot)$ is a penalty function on the precision matrix elements. A popular penalty function is to use the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996; Friedman et al., 2008; Yuan and Lin, 2007; Levina et al., 2008), which can be applied to shrink the off-diagonal elements in the precision matrix exactly to zero. Friedman et al. (2008), which is based on a coordinate descent procedure, is fast and can be adopted easily by many extensions of LASSO. For example, to remedy the bias issue in LASSO, Zou (2006) proposed the adaptive LASSO penalty and used the reciprocal of the absolute value of a consistent estimator raised to some power as the weight for each component. The solution can be obtained iteratively using weighted GLASSO. Another example is for the joint estimation of multiple graphical models (Guo et al., 2011). They proposed a factor across data categories for each off-diagonal element to represent the homogeneity network structure and put LASSO penalty on both the elements and factors. Their solution could also be obtained from an iterative weighted GLASSO algorithm.

However, these recent studies only work on association among genes. That is, these methods can describe the association between single genes only. It is known that pathways are sets of genes that serve a particular cellular or physiological function. Hence pathways are not independent of each other because of shared genes and interactions among them. Multi-pathway analysis has been a challenging problem because of the complex dependence structure among pathways. On the other hand, subtle connections between genes in two pathways may indicate strong connection between two pathways but can be ignored by individual gene network analysis. The main goal of our study is to develop a Gaussian graphical model for the gene and pathway network. Thus, by considering the dependency among pathways as well as genes within each pathway, we have proposed a multilevel Gaussian graphical model: one level is for pathway network structure and the second level is for gene network structure. We will propose a hierarchically structured graphical model for this in Section 2.

This paper is organized as follows. In Section 2, we propose a multilevel Gaussian graphical model for the gene and pathway network. Section 3 contains the penalized log-likelihood approach and the development of the algorithm for the solution. In Section 4 we compare our method with GLASSO method for individual gene networks based on several criteria. We introduce a definition of the degree of pathway-level connection. We also give a real data analysis in Section 5. Section 6 contains the conclusion and discussion.

2. Multilevel Gaussian graphical model

In this Section, we describe how to build a multilevel Gaussian graphical model for gene and pathway networks. First, we will provide the precision matrix for the Gaussian graphical model, then we will explain how to extract the pathway network information, and finally we will give a graphical illustration of the multilevel network model. Suppose we have p genes, the expression data for each were denoted by $X_1, \ldots, X_p$; the whole gene network can be represented by the precision matrix Ω,

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1p} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{p1} & \omega_{p2} & \cdots & \omega_{pp} \end{pmatrix}.$$

In this setting, if the off-diagonal element $\omega_{ij} = 0$, it means the ith and jth genes are conditionally independent. Furthermore, suppose these genes are in k predefined pathways, denoted by $P_1, \ldots, P_k$. Without loss of generality, we can re-denote the genes as: $X_{11}, \ldots, X_{1p_1}, X_{21}, \ldots, X_{2p_2}, \ldots, X_{k1}, \ldots, X_{kp_k}$, where $p_1, p_2, \ldots, p_k$ are the number of genes in each pathway. The conditional correlations among genes in the k and k'th pathways can be rewritten as a p_k-by-$p_{k'}$ sub-block precision matrix $\Omega_{kk'}$,

$$\Omega_{kk'} = \begin{pmatrix} \omega_{11}^{kk'} & \omega_{12}^{kk'} & \cdots & \omega_{1p_{k'}}^{kk'} \\ \omega_{21}^{kk'} & \omega_{22}^{kk'} & \cdots & \omega_{2p_{k'}}^{kk'} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{p_k1}^{kk'} & \omega_{p_k2}^{kk'} & \cdots & \omega_{p_kp_{k'}}^{kk'} \end{pmatrix}.$$