Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Estimating large covariance matrix with network topology for high-dimensional biomedical data

Shuo Chen [a,b,*], Jian Kang [c], Yishi Xing [d], Yunpeng Zhao [e], Donald Milton [f]

[a] Division of Biostatistics and Bioinformatics, School of Medicine, University of Maryland, Baltimore, MD, USA
[b] Maryland Psychiatric Research Center, School of Medicine, University of Maryland, Baltimore, MD, USA
[c] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[d] Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
[e] Department of Statistics, George Mason University, Fairfax, VA, USA
[f] Maryland Institute for Applied Environmental Health, University of Maryland, College Park, MD, USA

ARTICLE INFO

*Article history:*
Received 4 December 2017
Received in revised form 26 April 2018
Accepted 12 May 2018
Available online xxxx

*Keywords:*
Correlation matrix
Graph
Parsimony
Shrinkage
Thresholding

ABSTRACT

Interactions between features of high-dimensional biomedical data often exhibit complex and organized, yet latent, network topological structures. Estimating the non-sparse large covariance matrix of these high-dimensional biomedical data while preserving and recognizing the latent network topology are challenging. A two step procedure is proposed that first detects latent network topological structures from the sample correlation matrix by implementing new penalized optimization and then regularizes the covariance matrix by leveraging the detected network topological information. The network topology guided regularization can reduce false positive and false negative rates simultaneously because it allows edges to borrow strengths from each other precisely. Empirical data examples demonstrate that organized latent network topological structures widely exist in high-dimensional biomedical data across platforms and identifying these network structures can effectively improve estimating covariance matrix and understanding interactive relationships between biomedical features.



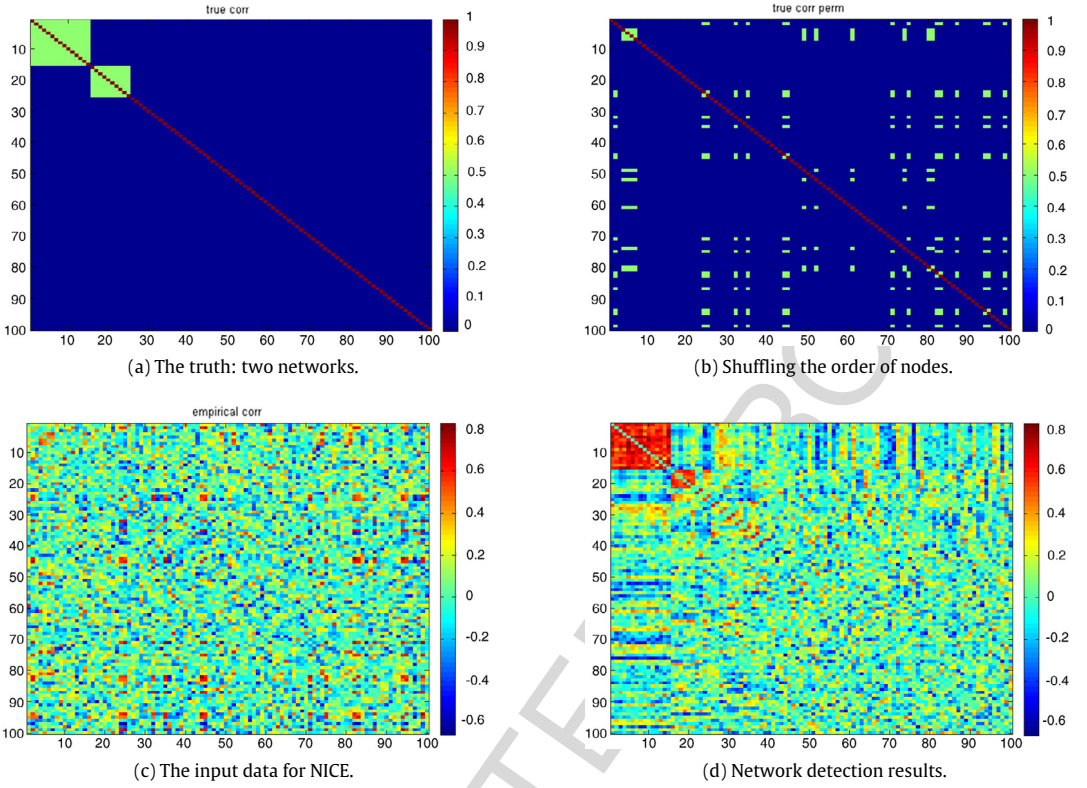
## 1. Introduction

Recent advances in bio-technologies allow measuring multi-dimensional biological features simultaneously in genomics, proteomics, and neuroimaging research. The underlying biological machinery is often associated with coordination between high-throughput features (Emilsson et al., 2008). For a large biomedical data set $\mathbf{X}_{n\times p}$ with the sample size $n$ and $p$ variables, estimating large covariance matrix $\boldsymbol{\Sigma}$ or correlation matrix $\mathbf{R}$ is fundamental to understand the interactive relationships between the biomedical features (Fan et al., 2015).

Regularization methods have been developed to estimate the high-dimensional covariance/correlation and precision matrix. For instance, the $\ell_1$ norm penalized maximum likelihood has been utilized to estimate the sparse precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ (Friedman et al., 2008; Banerjee et al., 2008; Yuan and Lin, 2007; Lam and Fan, 2009; Yuan, 2010; Cai and Liu, 2011; Shen et al., 2012) and the covariance matrix thresholding methods to directly regularize the sample covariance matrix (Bickel and Levina, 2008; Rothman et al., 2009; Cai et al., 2011; Zhang, 2010; Fan et al., 2013; Liu et al., 2014). The thresholding regularization techniques have also been applied to correlation matrix $\mathbf{R}$ estimation (Qi and Sun, 2006; Liu et al., 2014; Cui et al., 2016). We consider the estimation of standard deviations and correlation matrix are independent, and

* Corresponding author at: Division of Biostatistics and Bioinformatics, School of Medicine, University of Maryland, Baltimore, MD, USA.
*E-mail address:* shuochen@som.umaryland.edu (S. Chen).

https://doi.org/10.1016/j.csda.2018.05.008
0167-9473/




(a) The truth: two networks. (b) Shuffling the order of nodes.

(c) The input data for NICE. (d) Network detection results.

**Fig. 1.** An example of a network induced correlation matrix: $|V| = 100$ nodes and $|E| = 4950$ edges, there are two networks (a) and in practice they are implicit (b); it may be difficult to recognize the latent $G^1 \cup G^0$ mixture structure when looking at the sample correlation matrix (c); the proposed objective function is robust to false positive noise and identify the latent $G^1 \cup G^0$ mixture structure from the sample correlation matrix.

thus regularizing the large correlation and covariance matrix are exchangeable by $\hat{\boldsymbol{\Sigma}} = \text{diag}(\mathbf{S})^{-1/2}\hat{\mathbf{R}}\text{diag}(\mathbf{S})^{-1/2}$, where $\mathbf{S}$ is the sample covariance matrix (Barnard et al., 2000; Khondker et al., 2013; Fan et al., 2015). Graph notations and definitions are used to describe the relationship between the $p$ variables of $\mathbf{X}_{n\times p}$ (Yuan and Lin, 2007; Mazumder and Hastie, 2012). A finite undirected graph $G = \{V, E\}$ consists two sets, where the node set $V$ represents variables $\mathbf{X} = (X_1, \ldots, X_p)$ with $|V| = p$ and the edge set $E$ denotes relationships between the nodes. Let $e_{i,j}$ be the edge between nodes $i$ and $j$. Then $e_{i,j}$ is an connected edge if nodes $i$ and $j$ are dependent with each other in $G$. Under the sparsity assumption, the regularization algorithms assign most edges as unconnected, and $G$ can be decomposed to a set of maximal connected subgraphs (Witten et al., 2011; Mazumder and Hastie, 2012).

*Motivation: estimating large no-sparse covariance matrix with latent network topology.*

A key assumption for most aforementioned large covariance/precision estimation methods is the *sparsity* property that only a small proportion of edges are connected (variables are dependent), yet this assumption is not directly applicable in many biomedical applications (Fan et al., 2015). When analyzing high-dimensional omics data sets, we note that the interactions between biological features often interestingly exhibit *non-sparse* and organized network/graph topological patterns. The direct application of the regularization methods for large covariance/precision matrix estimation (with sparsity assumption) may miss interactions between features with network topology. Recently, the factor based large covariance matrix estimation methods have been developed to account for the common factors of the dependence structure between features (Fan et al., 2013, 2015, 2016). However, these methods may not explicitly provide inferences on the interactive relationships that reveal and reflect underlying network topological structures. Therefore, we propose a new statistical procedure that discovers the latent network topological structures and regularizes the covariance/correlation matrix with the guidance of the detected networks.

*Network topological structure and detection:* we frequently observe a specific $G^1 \cup G^0$ mixture structure (though it is latent) in omics and imaging data sets (see Fig. 1 and the two examples in Section 3). This topological structure denotes $G$ as a mixture of two components $G = G^1 \cup G^0$ where the first component $G^1 = \cup_{c=1}^{C_1} G_c^1$ is a stochastic block model structure and the second component $G^0 = \cup_{c=1}^{C_0} G_c^0$ ($G_c^0$ is a singleton) can be considered as an Erdös–Rényi random graph. We refer it as the $G^1 \cup G^0$ mixture structure. The $G^1 \cup G^0$ mixture structure is a special case of the stochastic block model, which contains

Please cite this article in press as: Chen S., et al., Estimating large covariance matrix with network topology for high-dimensional biomedical data. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.05.008.