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.

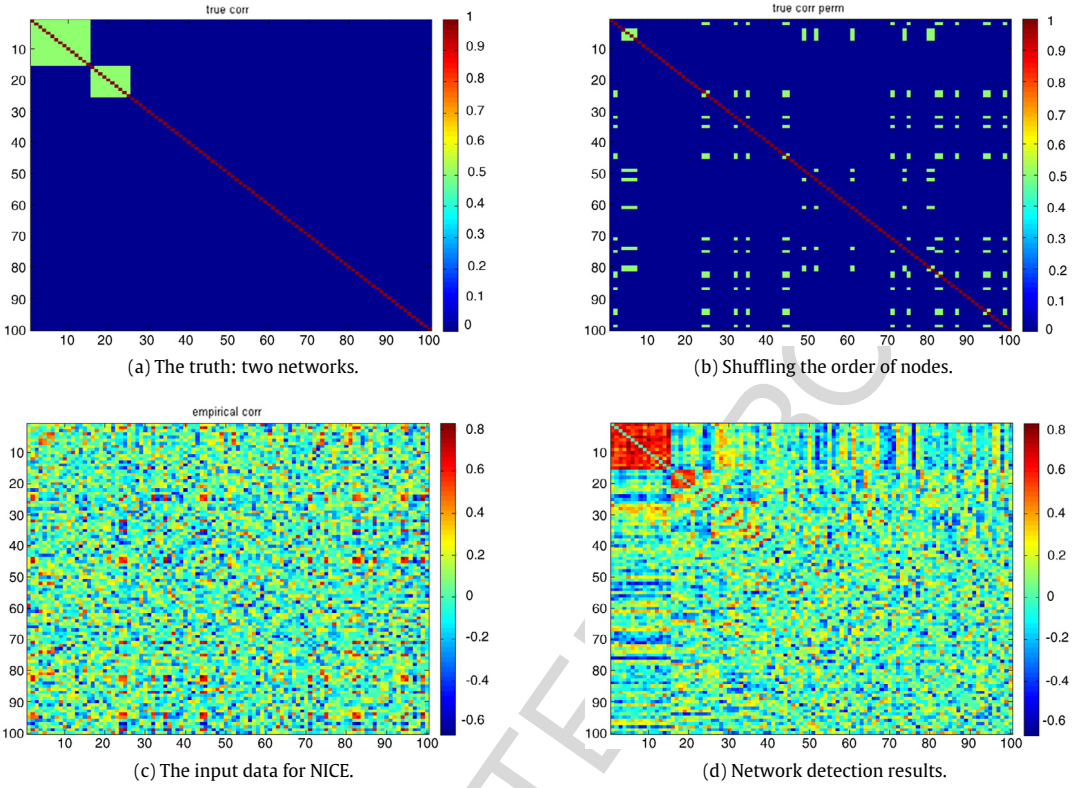
1. Introduction

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 ℓ_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.

many singletons and a number of communities (Bickel and Chen, 2009; Karrer and Newman, 2011; Nadakuditi and Newman, 2012). However, when exploring many commonly used community detection tools that can handle weighted/continuous similarity matrix (e.g. Newman, 2006; Pons and Latapy, 2006; Blondel et al., 2008; Fortunato, 2010 among others), we note that these methods cannot reveal the underlying $G^1 \cup G^0$ mixture structure due to the impact of noises (Tan et al., 2015; see detailed evaluation of noise effects in Section 3.2). To fill the gap, we propose a new parsimonious computational algorithm to effectively recognize the latent network topological structure from the sample correlation matrix, which is robust to false positive noises (edges). Our proposed method imposes a new penalty term on the sizes of networks so that the objective function maximizes the number of highly correlated edges in $\{G_c\}$ of G^1 while minimizing the sizes of edges in $\{G_c\}$. We consider the (sample) false positively correlated edges are often distributed in a random pattern rather in a community structure based on the fact that the graph combinatorics probability of false positively correlated edges forming an organized network is close to zero. Therefore, with the new penalty term these false positively correlated edges have little impact on organized network detection because including one false positive edge can greatly increase the sizes of detected networks. By implementing the penalized optimization, our proposed approach can effectively detect the latent $G^1 \cup G^0$ mixture structure and robust to noises. Therefore, the new latent network topology detection method alone would contribute as a new community detection algorithm that is robust to noises and well suited for detecting the $G^1 \cup G^0$ mixture structure. Since the detection of the latent $G^1 \cup G^0$ mixture structure is the foundation of the detection of many other topological structures and it widely exists in biomedical data sets across platforms, we focus on introducing the approach for $G^1 \cup G^0$ mixture structure detection in this article. Additionally, we include a more general and expandable algorithm to identify more sophisticated network topological structures in the Supplementary Materials.

Network topology guided large correlation matrix estimation: the detected network topology does not automatically yield large covariance matrix estimation, whereas it can provide useful guidance for thresholding decision. Estimating covariance/precision matrix involves regularizing edges that are entries of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ (Mazumder and Hastie, 2012). Edges in the large covariance matrix are constrained by nodes and naturally dependent with each other, and thus inferences/regularization decision on edges should incorporate the dependencies. However, the existing literature has not fully addressed this issue possibly because estimating the dependence between edges is computationally difficult, if feasible. We propose an new regularization strategy to threshold an edge based on both this edge's magnitude and the its 'spatial' *neighborhood* information via the detected graph topological structure: an edge is more likely to be connected when most of its neighbor edges in the same network are connected. Thus, the proposed network guided thresholding strategy allows edges to borrow strengths from each other (accounting for dependencies between edges) while avoiding the computationally difficult step to estimate *covariance between edges*. We determine the thresholds of edges within and edges outside detected networks in a data driven fashion by using an empirical Bayes approach that utilizes the detected graph topology as prior knowledge. Like many spatial statistics models, our proposed method can improve the efficiency of estimates without introducing false positive noises by taking network topology (spatial) dependence structure into account.

We name the two step network topology information guided regularization strategy as **N**etwork **I**nduced **C**orrelation matrix **E**stimation (NICE). The NICE method makes three contributions: (i) we propose a new penalized objective function that is well-suited to estimate latent network topological structures and robust to false positive noises; (ii) we fuse the network topological information and thresholding decision making procedure to simultaneously reduce false positive and false negative discovery rates; and (iii) we develop computationally efficient algorithms. In addition, the detected graph topological structures may also help to reveal underlying biological networks. The paper is organized as follows. Section 2 describes the NICE method. In Section 3, we apply our method to mass spectrometry proteomics data and gene expression data, and perform simulation studies for model evaluation/comparisons. Concluding remarks are summarized in Section 4.

2. Methods

We consider the sample covariance $\mathbf{S}$ and sample correlation matrix $\widehat{\mathbf{R}} = \text{diag}(\mathbf{S})^{-1/2}\,\mathbf{S}\,\text{diag}(\mathbf{S})^{-1/2}$ as our input data (Qi and Sun, 2006; Liu et al., 2014; Fan et al., 2015). We can directly perform hard thresholding on the sample correlation matrix to estimate $\mathbf{R}$ by using $R^{\mathcal{T}}_{i,j} = \{\widehat{R}_{i,j} I(|\widehat{R}_{i,j}| > T)\}$ without exploring the underlying network structure, where T is a pre-specified or calculated threshold (e.g. Bickel and Levina, 2008). However, applying the universal regularization rule (even when optimal T is provided) to each element (or column) may introduce numerous false positive and false negative findings when underlying topological structures exist and the large covariance/correlation matrix is not sparse. Therefore, we propose to leverage the information from the latent topological structure of the correlation matrix to assist the decision making process.

The NICE method consists two steps: (i) we first detect the latent topological structure of $G = G^1 \cup G^0$ mixture in G by applying the rule of parsimony; (ii) we then propose a new empirical Bayes based thresholding strategy to estimate the correlation matrix guided by the detected graph topology. We consider the sparsity assumption only holds for the outside network component, but not for the inside network component. In that, the convergence rates and false positive rates of thresholding for the outside network component are similar to the results of large sparse covariance matrix regularization (Rothman et al., 2009; Cai and Liu, 2011), and the interactions within network topological structures can be well captured and preserved.

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.

2.1. Estimating latent networks with parsimony from sample correlation matrix

We first define the weighted similarity matrix $\mathbf{W}$ based on the sample correlation matrix $\widehat{\mathbf{R}}$. An entry $w_{i,j}$ of $\mathbf{W}$ can be a transformed correlation coefficient between variables i and j that corresponds to the edge $e_{i,j}$ in G, for example, Fisher's Z transformation. $w_{i,j}$ is often a continuous metric. In the Supplementary Materials, we describe an empirical Bayes based procedure to calculate $w_{i,j}$ as a metric between 0 and 1. $\mathbf{W}$ is only used for the latent network detection rather than the regularization step.

We assume that G includes a set of community networks and many singletons as shown in Fig. 1a, and edges within the networks are more likely to be connected than edges outside networks. However, in practice this topological structure is latent and the sample correlation matrix does not automatically reveal graph topological pattern 1c. By implementing the objective function (1) we can recognize the latent graph topological structures 1d. We next perform permutation tests to evaluate the statistical significance for each G_c, and the statistically significant subgraphs $\{G_c\}$ are used to assist the estimation of the correlation matrix in the following step.

We propose to detect the latent $G^1 \cup G^0$ mixture structure from $\mathbf{W}$ by using penalized optimization. The heuristic is to identify a set of subgraphs $G^1 \cup G^0 = \cup_{c=1}^{C} G_c$ ($\{G_c\} = \{G_c^1\} \cup \{G_c^0\}$) that maximizes the sum of weights of edges in the community networks while minimizing the community network edge sizes. The penalty term is used to avoid the disruption of false positive noises (edges). A useful fact is that a singleton G_c contribute none to the edge size term, and when the number of subgraphs equal to the number of nodes $C = |V|$ the sum of community network edge sizes $\cup_{c=1}^{C}|E_c| = 0$. On the other extreme, $C = 1$ leads to $\cup_{c=1}^{C}|E_c| = |E|$. Therefore, C is related to the network size penalty term and a larger C can increase the parsimony level.

Formally, we propose the objective function:

$$\arg\max_{C,\{G_c\}} \sum_{c=1}^{C} \exp\left[\log\left\{\sum(w_{i,j}|e_{i,j} \in G_c)\right\} - \lambda_0 \log(|E_c|)\right], \tag{1}$$

with following definitions and conditions:

1. G_c ($c = 1, \ldots, C$) is a clique subgraph that $G_c = \{V_c, E_c\}$ and $|V_c| \geq 1$;
2. the size of a subgraph G_c is determined by the number of edges $|G_c| = |E_c|$;
3. $\cup_{c=1}^{C} V_c = V$, $\cap_{c=1}^{C} V_c = \emptyset$ and $\cup_{c=1}^{C} E_c \subseteq E$.

The objective function is non-convex and difficult to be directly solved. We develop iterative algorithm to optimize C and $\{G_c\}$ in the objective function, and we provide the detailed derivation and optimization algorithms for detecting the latent $G^1 \cup G^0$ mixture structure (and for other topological structures) in the Supplementary Materials. By applying the penalty term, the objective function often selects a relatively large $\hat{C}$ and include many G_c as singletons to shrinkage the subgraph sizes. This new objective function is well-suited to capture graph topological structures from the sample (noisy) correlation matrix because it is less affected by the false positive noises by implementing the new penalty term (see more details in Section 3.1). The objective function generally performs well when the correlated edges are distributed in organized latent topological structures which include most clustered patterns (e.g. communities, interconnected communities, $G^1 \cup G^0$ mixture, k-partite, and rich-club). However, (1) may fail to detect the non-clustered structures, for example, Bandable and Toeplitz structures (Cai et al., 2016b), which can be examined by exploratory data analysis. Given the fact that the clustered topological structure (e.g. $G^1 \cup G^0$ mixture) widely exists in high-dimensional biomedical data, the NICE network detection algorithm can be often used as the first step to examine whether the covariance matrix is sparse and possible latent network topological structure exists. If no clustered network structure is detected, we can directly apply the existing methods for large sparse covariance/precision matrix estimation and the Bandable and Toeplitz structured covariance matrix estimation (Cai et al., 2016b). Therefore, Optimizing (1) provides the estimates of underlying network topological structure within the large sample correlation matrix, which can be used to guide the large correlation matrix regularization as the prior knowledge.

2.2. Graph topology oriented correlation matrix thresholding

To estimate the correlation matrix $\mathbf{R}$, we perform graph topology guided thresholding on the sample correlation matrix $\widehat{\mathbf{R}}$ by using Bayes factors (BF). Let $z_{i,j}$ be the Fisher's Z transformed sample correlation coefficient of $\widehat{R}_{i,j}$ and it follows a mixture distribution that $z_{i,j} \sim \pi_0 f_0(z_{i,j}) + \pi_1 f_1(z_{i,j})$. The distribution assumption is well supported by the example data sets in Section 3.1.

Universal thresholding

Without considering prior information of the topology structure, the universal thresholding can be applied (Bickel and Levina, 2008). For instance, an empirical Bayes framework implements a Bayes factor based via the (Efron, 2004; Schäfer and Strimmer, 2005). The hard-thresholding rule (Cai et al., 2011; Fan et al., 2015) is often employed for this purpose, which sets an edge to zero unless

$$\frac{P(\delta_{i,j} = 1|z_{i,j})}{P(\delta_{i,j} = 0|z_{i,j})} = \frac{f_1(z_{i,j})\pi_1}{f_0(z_{i,j})\pi_0} \geqslant T,$$

T is a constant that is linked to local *fdr* cutoff, and π_0 and π_1 are the proportions of null and non-null distributions correspondingly. For example, $T = 4$ is equivalent to the cutoff of local *fdr* of 0.2 (Efron, 2007). For instance, given $\pi_0 = 0.9$ and $\pi_1 = 0.1$, the universal decision rule is that an edge is thresholded when Bayes factor is less than 36, i.e.,

$$BF_{i,j} = \frac{f_1(z_{i,j})}{f_0(z_{i,j})} \leq 36.$$

In practice, π_0 and π_1 are estimated based on the distribution of the statistics (e.g. $z_{i,j}$) and the Bayes factor cut-off is updated accordingly.

It has been well documented that the Bayes factor inferential models could adjust the multiplicity by adjusting the prior structure (Jeffreys, 1961; Kass and Raftery, 1995; Scott and Berger, 2006; Efron, 2007). The prior odds are tuned to control false positive rates, and a larger π_0 ($\pi_0 \to 1$) or a distribution of π_0 with larger mean leads to more stringent adjustment that may cause both low false positive discovery rates and high false negative discovery rates. Scott and Berger (2006) suggest a prior distribution with median value around 0.9 and numerical methods have been developed to estimate π_0 (Wu et al., 2006; Efron, 2007). However, the universal regularization methods (e.g. shrinkage or thresholding) face a trade-off between false positive and false negative findings and ignore the dependency structure between edges.

Network based thresholding

In a network induced correlation matrix, an edge with sample correlation value $z_{i,j}$ is more likely to be truly connected within than outside a network community because the within community 'neighbor' edges are more connected. Thus, we incorporate the topological location information of an edge into the regularization procedure. We first calculate the prior odds for edges (to be connected)within and outside community networks separately by:

$$\theta_{in} = \frac{\mathrm{P}(\delta_{i,j} = 0 | e_{i,j} \in G_c, \forall c)}{\mathrm{P}(\delta_{i,j} = 1 | e_{i,j} \in G_c, \forall c)} = \frac{\pi_0^{in}}{\pi_1^{in}},$$

$$\theta_{out} = \frac{\mathrm{P}(\delta_{i,j} = 0 | e_{i,j} \notin G_c, \forall c)}{\mathrm{P}(\delta_{i,j} = 1 | e_{i,j} \notin G_c, \forall c)} = \frac{\pi_0^{out}}{\pi_1^{out}}.$$

Clearly, the within community edges are more connected by and thus $\pi_1^{in} > \pi_1 > \pi_1^{out}$ and $\pi_0^{out} > \pi_0 > \pi_0^{in}$, and $\theta_{out} \geq \theta_{all} \geq \theta_{in}$.

Let edges inside and outside of the detected communities follow distributions:

$$z_{i,j} | e_{i,j} \in G_c \sim \pi_0^{in} f_0(z_{i,j}) + \pi_1^{in} f_1(z_{i,j}) \quad \text{and}$$

$$z_{i,j} | e_{i,j} \notin G_c \sim \pi_0^{out} f_0(z_{i,j}) + \pi_1^{out} f_1(z_{i,j}),$$

respectively.

The proportions are different for the inside and outside networks, as well as overall edges, yet we assume that the null $f_0(z_{i,j})$ and non-null $f_1(z_{i,j})$ distributions are identical (the distribution assumption is well supported by examples in Section 3.1). By using the identified the latent networks where edges in step one, we propose the network based thresholding rule. We denote $\widehat{R}_{i,j}^{T}$ the thresholded correlation estimates for nodes i and j. Let $\widehat{BF}_{i,j} = \widehat{f}_1(z_{i,j})/\widehat{f}_0(z_{i,j})$ be an estimate of $BF_{i,j}$, and $\widehat{\theta}_{in}$ and $\widehat{\theta}_{out}$ be the estimates of θ_{in} and θ_{out}, respectively. Then, the posterior odds update as $\frac{\mathrm{P}(\delta_{i,j}=1|z_{i,j}, e_{i,j} \in G_c, \forall c)}{\mathrm{P}(\delta_{i,j}=0|z_{i,j}, e_{i,j} \in G_c, \forall c)} = \frac{f_1(z_{i,j})\hat{\pi}_1^{in}}{f_0(z_{i,j})\hat{\pi}_0^{in}} = \widehat{BF}_{i,j}/\widehat{\theta}_{in}$, and the network topology guided decision rule is as follows.

If $e_{i,j} \in G_c$,

$$\widehat{R}_{i,j}^{T} = \begin{cases} \widehat{R}_{i,j} & \text{if } \widehat{BF}_{i,j} \geq T \cdot \widehat{\theta}_{in}; \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

else if $e_{i,j} \notin G_c$,

$$\widehat{R}_{i,j}^{T} = \begin{cases} \widehat{R}_{i,j} & \text{if } \widehat{BF}_{i,j} \geq T \cdot \widehat{\theta}_{out}; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Equivalently, the we provide estimate of the edge set $\widehat{E}$ by using:

$$\begin{aligned} \widehat{\delta}_{i,j}^{in} &= I\left(\widehat{BF}_{i,j} \geq T \cdot \widehat{\theta}_{in}\right) \\ \widehat{\delta}_{i,j}^{out} &= I\left(\widehat{BF}_{i,j} \geq T \cdot \widehat{\theta}_{out}\right), \end{aligned} \tag{4}$$

where $\delta_{i,j}$ is an indicator variable that $\delta_{i,j} = 1$ when variables i and j are correlated with each other, otherwise $\delta_{i,j} = 0$.

We obtain $\widehat{f}_1$, $\widehat{f}_0$, $\widehat{\theta}_{in}$ and $\widehat{\theta}_{out}$ by the following steps. First, assume the edge-specific Fisher's Z transformed sample correlation coefficients in $\widehat{\mathbf{R}}$ follow a mixture distribution: $f(z_{i,j}) = \pi_0^{all} f_0(z_{i,j}) + \pi_1^{all} f_1(z_{i,j})$. $\widehat{\pi}_0^{all}$, $\widehat{\pi}_1^{all}$, $\widehat{f}_0$ and $\widehat{f}_1$ can be estimated by using algorithms for local *fdr* (Efron, 2007). Next, using $\widehat{f}_0$ and $\widehat{f}_1$ in the previous step we estimate $\widehat{\pi}_0^{in}$ for in-network edges $e_{i,j} \in G_c$ as the only parameter in $f^{in}(z_{i,j}) = \pi_0^{in} f_0(z_{i,j}) + \pi_1^{in} f_1(z_{i,j})$ via maximum likelihood estimation. In results,

we calculate $\widehat{\theta}_{in} = \widehat{\pi}_0^{in}/\widehat{\pi}_1^{in}$ ($\widehat{\pi}_1^{in} = 1 - \widehat{\pi}_0^{in}$). For edges outside of networks ($z_{i,j}$ that $e_{i,j} \notin G_c$), we estimate $\widehat{\pi}_0^{out}$ in $f^{out}(z_{i,j}) = \pi_0^{out} f_0(z_{i,j}) + \pi_1^{out} f_1(z_{i,j})$ by following steps above, and calculate $\widehat{\theta}_{out} = \widehat{\pi}_0^{out}/\widehat{\pi}_1^{out}$ by following the same procedure.

In practice, our graph topological structure detection algorithm produces a very small odds ratio $\widehat{\theta}^{in}/\widehat{\theta}^{out}$ when the informative edges are distributed in an organized pattern. Thus, the choice of T has little impact on thresholding.

The detected graph topology provides the prior knowledge of the 'neighborhood' and 'location' for each edge. A network defines a neighborhood (spatial closeness) of *edges* with explicit boundaries and edges within the same neighborhood can borrow power from each other. Many statistical models are developed to account for dependency based on the neighborhood definition, for example, the Ising prior and conditional autoregressive (CAR) model (Besag and Kooperberg, 1995). Nevertheless, unlike data in spatial or imaging statistics the sample correlation/covariance matrix of large biomedical data often include no available information about the exact spatial location or closeness on edges. Alternatively, the detected graph topological structure provides a graph topological 'closeness' of edges and thus accounts for the dependency between edges. Last, when no network topological patterns are detected and the sparsity assumption is valid, the existing large covariance/precision matrix estimation methods can be applied.

2.3. Reduced false positive and negative discovery rates by using NICE thresholding

We show that under a mild regularity condition, the NICE method can simultaneously reduce false positive and false negative finding rates.

Condition 1. *Let* $\omega = (\sum_{c=1}^{C} |V_c| \times (|V_c| - 1)/2)/(|V| \times (|V| - 1)/2)$ *the proportion of edges inside community networks and* $\int_{z_0}^{\infty} f(z_{i,j}) = F(z_0)$. z_0 *is the universal threshold cut-off value,* $z_{0,in}$ *is the within networks threshold cut-off value,* $z_{0,out}$ *is the within networks threshold cut-off value.*

$$\frac{F_0(z_0) - F_0(z_{0,out})}{F_0(z_{0,in}) - F_0(z_0)} > \frac{\omega \pi_0^{in}}{(1-\omega)\pi_0^{out}}$$

$$\frac{F_1(z_0) - F_1(z_{0,out})}{F_1(z_{0,in}) - F_1(z_0)} < \frac{\omega \pi_1^{in}}{(1-\omega)\pi_1^{out}}.$$

This condition is generally valid for network induced correlation matrix because by implementing the parsimonious estimation of network topological structure f^{in} is distinct from f^{out} (see Figs. 2 and 3). Thus, we have $\pi_0^{in} \ll \pi_0^{out}$ and $\pi_1^{in} \gg \pi_1^{out}$, and condition holds.

Theorem 1. *Suppose Condition 1 holds, we have both (1)* $E(\sum_{i<j} I(\widehat{\delta}_{ij}^{NICE} = 1 | \delta_{ij} = 0)) \leq E(\sum_{i<j} I(\widehat{\delta}_{ij}^{Univ} = 1 | \delta_{ij} = 0))$ *the expected false positively thresholded edges by using the network topology oriented thresholding (NICE) method are less than the universal thresholding method; (2)* $E(\sum_{i<j} I(\widehat{\delta}_{ij}^{NICE} = 0 | \delta_{ij} = 1)) \leq E(\sum_{i<j} I(\widehat{\delta}_{ij}^{Univ} = 0 | \delta_{ij} = 1))$*, the expected false negatively thresholded edges by using the NICE method are less than the universal thresholding method.*

The proof is in the Supplementary Materials.

Uncovering the graph topological structure is important to understand the interactive relationships between multivariate variables (nodes) and the dependency between edges. We show that the detected topological structure can also provide prior knowledge to assist large covariance/correlation matrix regularization and estimation. The network based regularization approach utilizes the additional yet latent graph structure information and reduces false positive and negative discovery rates simultaneously. We summarize the overall NICE algorithm of both steps in details in the Supplementary Materials.

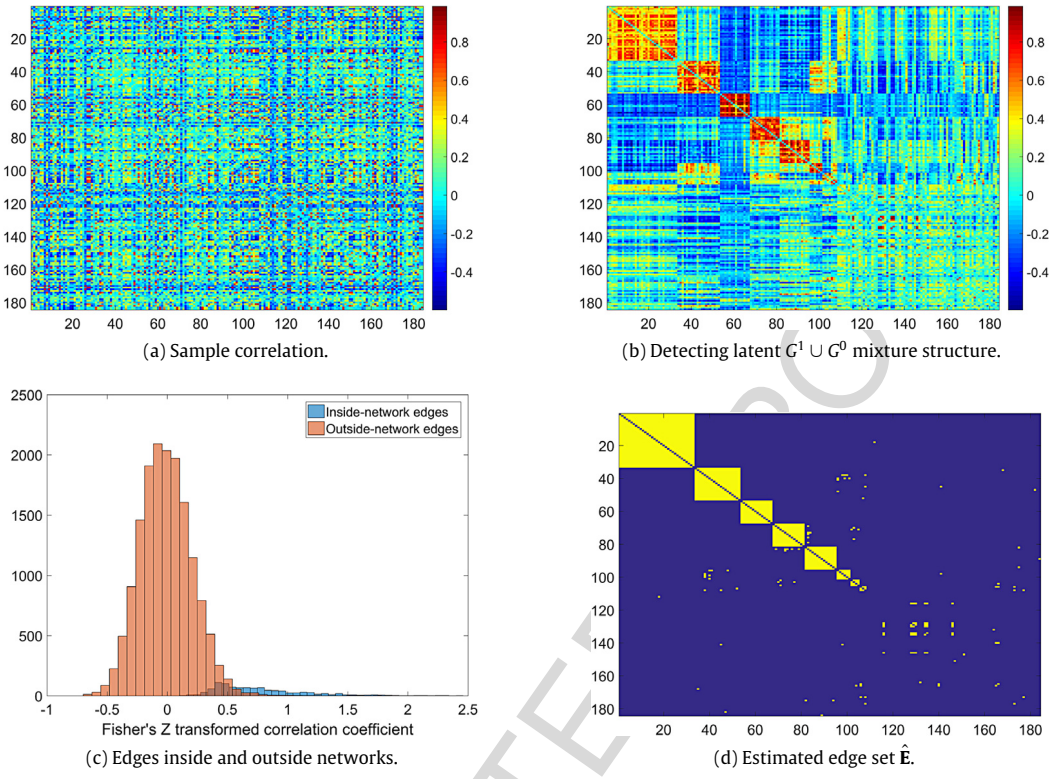
3. Data analysis

3.1. Data examples

We apply the NICE method to two publicly available high-dimensional biomedical data sets. Along with these two examples, we find that the latent $G^1 \cup G^0$ mixture structure widely exists in data across platforms (e.g. proteomics, genomics, and imaging data, yet due to space limitation we only demonstrate two data types).

3.1.1. Proteomics data

We first focus on a matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) proteomics data from human 288 subjects (Yildiz et al., 2007). The data assess the relative abundance of peptides/proteins in human serum. Each raw mass spectrum consists roughly 70,000 data points. After preprocessing steps including registration, wavelets denoising, alignment, peak detection, quantification, and normalization (Chen et al., 2009), 184 features are extracted to represent the most abundant protein and peptide features in the serum. Each feature is located at a distinct m/z value that could be linked to a specific peptide or protein with some ion charges (feature id label). The original paper



(a) Sample correlation.
(b) Detecting latent $G^1 \cup G^0$ mixture structure.
(c) Edges inside and outside networks.
(d) Estimated edge set $\hat{\mathbf{E}}$.

Fig. 2. Application of the NICE to the example data set one. (a) Is the heatmap of sample correlation matrix; (b) demonstrates the latent $G^1 \cup G^0$ mixture structure by reordering the variables in the heatmap; (c) shows the distributions of edges inside and outside the networks; (d) is the estimated $\hat{E}$ based on the NICE thresholding.

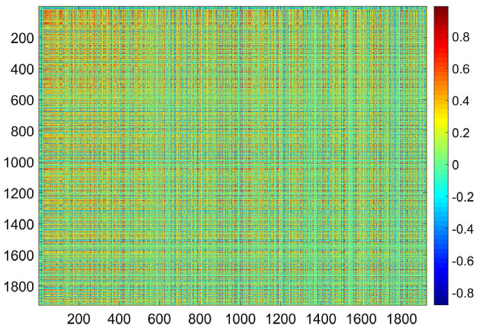
utilizes the proteomics data to enhance understanding of lung cancer pathology at the molecular level. In this paper, we focus on estimating the correlation matrix to investigate interactive relationships between these features.

We apply the NICE approach to detect correlated peptide/protein networks and estimate correlation matrix based on the Fisher's Z transformed sample correlation matrix (Fig. 2a). First, the penalized objective function (1) is implemented to capture the latent $G^1 \cup G^0$ mixture structure. The estimation results are $\hat{C} = 77$, and that seven significant community networks (G_1) are detected and the rest are singletons (G_0) (see Fig. 2b). Fig. 2b reorders features Fig. 2a by the detected topological structure. Generally, features within networks are more correlated than features outside networks

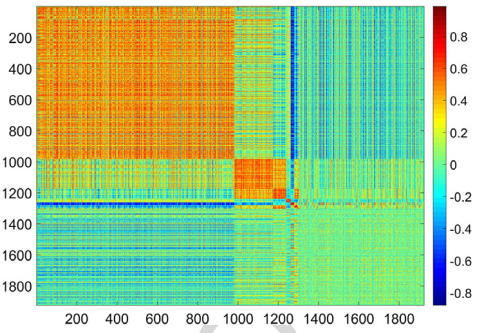
We show that the distributions of edges inside and outside networks in Fig. 2c. Clearly, f^{in} and f^{out} show distinct distributions, and f^{out} is close to the null distribution for which all edges are not connected. We estimate $\hat{\pi}_0^{all} = 0.78$, $\hat{\pi}_0^{out} = 0.83$, and $\hat{\pi}_0^{in} = 0.001$. Then, we apply the network based thresholding to estimate $\hat{E}$ and the correlation matrix. The estimate $\hat{E}$ and thresholding rule $\{\hat{\delta}_{i,j}\}$ are shown in Fig. 2d. The network detection results provide informative inferences on interactive relationships between these proteomic features. In this data example, each network represents a group of related protein and peptides that can be confirmed by proteomics mass spectrometry literature. For example, the most correlated network three consists a list of proteins of normal and variant hemoglobins with one and two charges (Lee et al., 2011) including normal hemoglobins α and β with one charge and two charges (at m/z 15 127, 15 868, 7564, and 7934). We also note that a few edges connecting nodes from different communities and the random graph component. The correlated biomedical features may provide guidance to identify a set of biomarkers for future research that allow to borrow power between each other.

3.1.2. Gene expression data

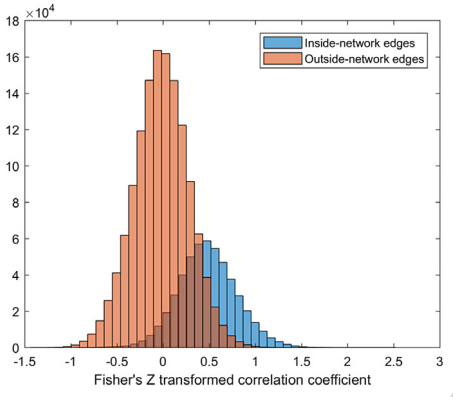
The second empirical data example is gene expression profiling data based on Affy Human Genome U133A 2.0 array. The data is publicly available at Gene Expression Omnibus (GEO) with accession code: GSE17156, GSE30550, GSE52428. Blood samples were collected for 110 healthy controls at baseline. We focus on 1924 gene expression features that are commonly observed in human blood, and normalized data is used for analysis. The input data for our model is a 1924×1924 sample correlation matrix (Fig. 3a). The sample correlation matrix shows no explicit organized topological structures. By applying the penalized objective function in (1), we identify the latent $G^1 \cup G^0$ mixture structure (Fig. 3b). Note that Fig. 3b is a



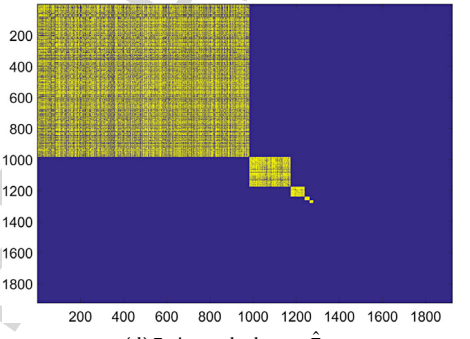
(a) Sample correlation.



(b) Weight matrix **W**.



(c) Edges inside and outside networks.



(d) Estimated edge set $\hat{\mathbf{E}}$.

Fig. 3. Application of the NICE to the example data set two. (a) Is the sample correlation matrix; (b) demonstrates the latent $G^1 \cup G^0$ mixture structure by reordering the variables in the heatmap; (c) shows the distributions of edges inside and outside the networks; (d) is the estimated $\hat{E}$ based on the NICE thresholding.

isomorphic graph to Fig. 3a with reordered nodes. With $\hat{C} = 613$, four large networks and a long list of singletons and small networks (with 2 or 3 nodes) are detected because of the penalty term.

Fig. 3c shows that edges inside and outside community networks follow distinct distributions. We estimate $\hat{\pi}_0^{all} = 0.84$, $\hat{\pi}_0^{out} = 0.99$, and $\hat{\pi}_0^{in} = 0.05$. The distribution of edges outside of community networks is also close to the null distribution of non-connected edges, whereas the distribution of edges inside networks again centers around 0.5. By applying the network guided thresholding, we obtain the estimated correlation matrix and $\hat{E}$ as shown in Fig. 3d. Due to the page limit, we do not provide the long list of inter-related gene expression features which involve a large number of features related to the proinflammatory chemokines and cytokines in the peripheral blood cell that are well-documented in precedent studies (e.g. Baechler et al., 2003). We note that the community network sizes in the gene expression data are larger and a larger proportion of inside edges are thresholded to zeros.

Interestingly, the $G^1 \cup G^0$ mixture structure both can be detected by NICE from both proteomics and genomics data sets, which is also discovered in large data from many other platforms including neuroimaging activation and connectivity data, DNA methylation data, and etc. Chen et al. (2016). Hence, these results further verify our belief that underlying organized network topological structures widely exist in biomedical large data. Network guided large covariance/correlation estimation can provide better inferences on interactive relationships between the massive biomedical features with network topological structures. In contrast, when we apply the methods for large *sparse* covariance/precision matrix estimation (e.g. *glasso*), many edges of the latent organized topological structures are (false negatively) regularized to zeros and the latent network topological structure can neither be correctly identified based on the estimated covariance/precision matrix (see details in the Supplementary Materials). We also note that the estimation results of NICE and these methods are similar for the outside network edges.

3.2. Simulation studies

We conduct numerical studies to evaluate the performance of our approach, and compare it with several existing methods.

Table 1
Comparisons of network topology detection results: medians along with 25% and 75% quantiles of FP and FN.

Method	$\rho = 0.5, n = 25$				$\rho = 0.5, n = 50$				$\rho = 0.7, n = 25$			
	FP		FN		FP		FN		FP		FN	
	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles
Newmann	1651.5	(1479, 1792.5)	15	(0, 35)	1595.5	(1440.5, 1739)	0	(0, 26)	1652.5	(1512, 1791,5)	9	(0, 24.5)
Random walk	714.5	(704.5, 719)	129.5	(119.5, 134)	714	(708, 721)	129	(123, 136)	713.5	(707, 721)	128.5	(122, 136)
Hierarchical	571.5	(402.5, 673.5)	16.5	(0, 45)	461.5	(342.5, 553.5)	0	(0, 0)	558.5	(412, 669)	0	(0, 45)
NICE	44	(15, 98)	3	(0, 27)	11	(1, 30)	0	(0, 4)	0	(0, 0)	0	(0, 14)

3.2.1. Synthetic data sets

We simulate each data set with $p = 100$ variables, and thus $|V| = 100$ and $|E| = \binom{100}{2} = 4950$. We assume that the correlation matrix includes two community networks, and the first include 15 nodes and the second 10 nodes. The induced networks are complete subgraphs (cliques) that all edges are connected within these two networks and no other edges are connected outside the two networks (Fig. 1a). Next, we permute the order of the nodes to mimic the practical data where the topological structure is latent. Fig. 1b represents the connected edges in the matrix. Let vector $\mathbf{x}^k_{p\times 1}$ follow a multivariate normal distribution, with zero mean and covariance matrix $\boldsymbol{\Sigma}_{p\times p}$, and the sample size is n. $\sigma_{i,j}$ is an entry at the ith row and jth column of $\boldsymbol{\Sigma}$, $\sigma_{i,j} = 1$ if $i = j$ (then $\boldsymbol{\Sigma} = \mathbf{R}$), and $\sigma_{i,j} = \rho$ if $e_{i,j} \in G_c$ (inside network edges) and $\sigma_{i,j} = 0$ when $e_{i,j} \notin G_c$ (outside network edges). We simulate 100 data sets at four different settings by using various sample sizes n and values of ρ. A larger n reduces the asymptotic variance of $\widehat{\sigma}_{i,j}$ and a higher absolute value of ρ represents higher signal level, which jointly lead to more distinct empirical distributions of $\widehat{\sigma}_{i,j}|e_{i,j} \in G_c$ and $\widehat{\sigma}_{i,j}|e_{i,j} \notin G_c$. All large covariance matrix thresholding methods and precision matrix shrinkage methods are expected to perform better with larger n and ρ. Fig. 1c demonstrates a calculated correlation matrix based on a simulated data set. The simulated data mimic the practical high-dimensional omics data well when comparing it to the aforementioned data examples.

In our simulated data sets, 150 edges are connected and 4800 edges are unconnected, which together represent the graph edge skeleton E. Since the NICE thresholding method relies on the network topology detection, we first compare the NICE network topology detection method to three of the popular community detection methods that can handle continuous similarity matrix. Next, we compare NICE with the existing large covariance matrix thresholding and precision matrix shrinkage methods. Note that we simulate the covariance matrix (instead of the precision matrix) based on known network topolical structure. Based on the results by Mazumder and Hastie, 2012 (Mazumder and Hastie, 2012), the estimated graph edge skeleton $\hat{E}$ by large covariance matrix thresholding and precision matrix shrinkage are comparable. Therefore, we treat a non-zero entry $\widehat{\delta}_{i,j} = 1$ as a connected edge if $\hat{\Sigma}_{i,j} > 0$ via large covariance matrix thresholding or $\hat{\Omega}_{i,j} > 0$ via precision matrix shrinkage, and compare NICE with both methods. We summarize the false positive (FP) edges $\widehat{\delta}_{i,j} = 1$ when the edge is not connected and $e_{i,j} \notin G_c$ and false negative (FN) edges $\widehat{\delta}_{i,j} = 0$ when the edge is connected and $e_{i,j} \in G_c$. We compare FN and FP counts of each method by contrasting the estimated $\widehat{E}$ with the truth E. We compare our method with *glasso*, universal thresholding (Thresh), adaptive thresholding (AThres), l_1 minimization for inverse matrix estimation (CLIME), and adaptive CLIME (ACLIME) by comparing the FP and FN edges of estimating the graph edge skeleton E (Bickel and Levina, 2008; Friedman et al., 2008; Cai et al., 2011; Cai and Liu, 2011; Cai et al., 2016a). Additionally, we compare the matrix loss $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{True}\|_F$ of NICE with the existing large covariance matrix thresholding methods by using the Frobenius matrix norm.

3.2.2. Simulation study results

Network topology detection: we compare the proposed method with three popular community detection algorithms by Newman (2006), Pons and Latapy (2006), Blondel et al. (2008) and we denote them as Newman, Hierarchical, and Random walk algorithms correspondingly in the tables and figures. The numbers of communities are determined by the embedded functions of these algorithms. We demonstrate the performance of the four methods in Fig. 4 using the setting $n = 50$, $\rho = 0.5$, and the proposed network detection method can better identify the latent network structure than the other methods. The objective functions of the competing methods can be affected by noises to include false positive edges and miss the underlying network topological patterns while the penalty term of the proposed method can avoid the impact of noises. For the network detection analysis, we set $\rho = 0.5, 0.7$ because most network detection methods handling continuous adjacency matrices assume ρ is sufficiently larger than 0 (Newman, 2006).

In addition, we summarize the 25%, 50%, and 75% quantiles of the number of FP and FN edges based on the detected network topological structure to assess the performance of each method cross the 100 simulation data sets in Table 1. We find that our proposed method is robust to false positive noise and can effectively detect the well organized network topological structure as a subgraph in G. The other three methods often include a large number of false positive edges and miss the latent structure across all settings. With the large number of false positive edges, the detected network topology cannot provide precise 'spatial' neighborhood information between edges and thus may not assist covariance matrix estimate. Therefore, the proposed network detection method provides a more robust pathway to detect informative latent network topological structure with higher accuracy to recover the latent network structure. The simulation results also indicate that the iterative process in optimizing the objective function (1) is fast because the maximum number of subgraphs is bounded by $|V|$. For each simulation data set, the computational cost is around 40 s by using an i7 3.4G HZ CPU and 24G memory desktop.

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.

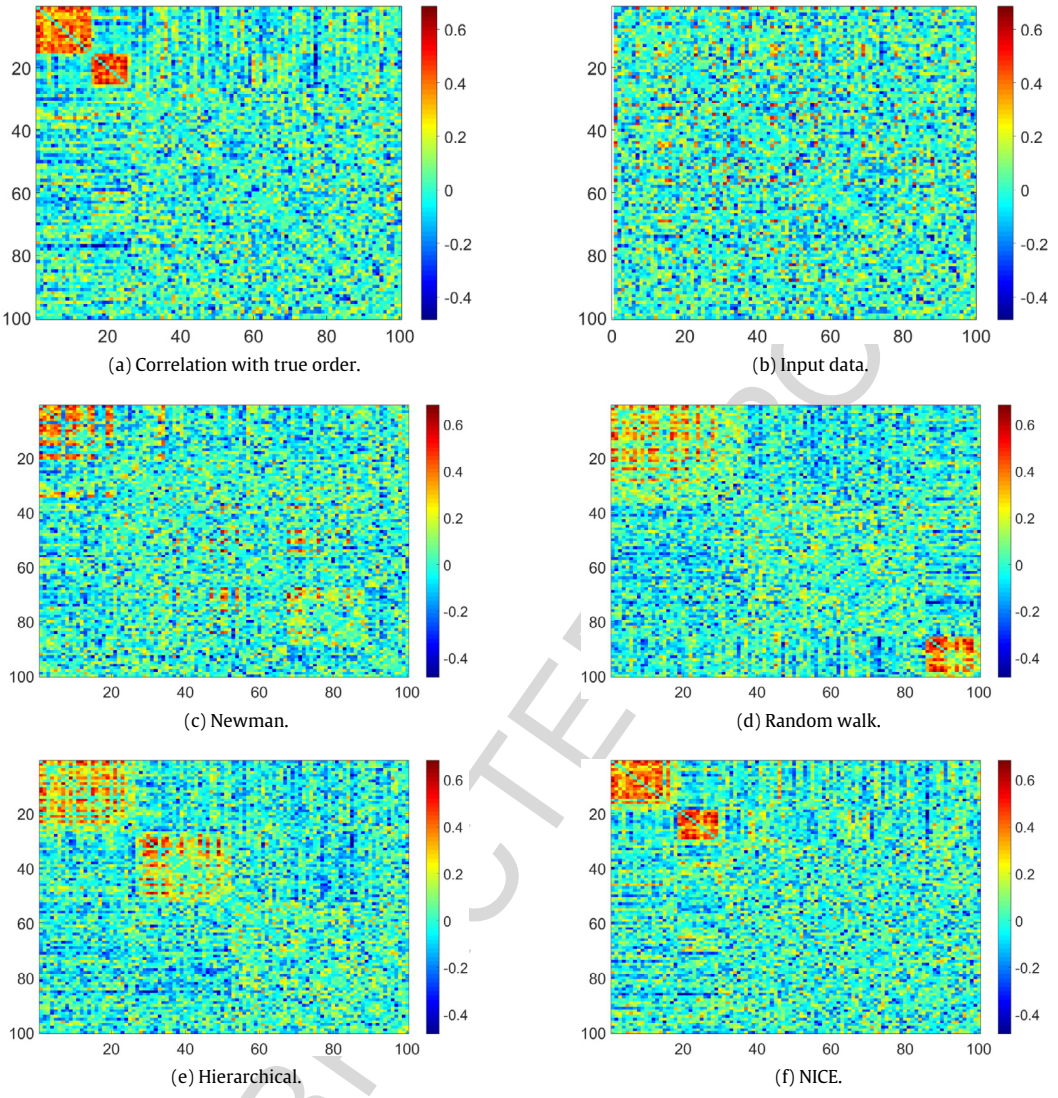


Fig. 4. Comparing the NICE network topology detection algorithm with the popular network detection methods. (a) Is the correlation matrix by listing nodes in communities first (with known network topological structure); (b) is the input data for all methods with unknown underlying network topological structure; (c) results of Newman algorithm; (d) results of Random walk algorithm; (e) results of Hierarchical algorithm; (f) results of NICE network detection algorithm. The results show that the NICE network detection algorithm is more robust to noises by applying the rule of parsimony and can better detect the underlying $G^1 \cup G^0$ mixture topological structure.

Comparisons of large covariance/precision matrix estimation results: we compare the graph edge skeleton $\hat{E}$ estimated by large covariance matrix thresholding and precision matrix shrinkage (see Table 2) across all simulation settings. Since in practice we rarely observe that both n and ρ are very large, we focus on the four settings in Table 2. Rather than selecting a single tuning parameter λ for *glasso* and other methods by cross-validation, we explore all possible choices within a reasonable range and use the one with best performance for comparison. For ACLIME, we let the tuning parameter equal to 2 as suggested (Cai et al., 2016a). Cross the 100 simulation data sets, we summarize the 25%, 50%, and 75% quantiles of the number of FP and FN edges to assess the performance of each method. The results show that the NICE algorithm outperforms the competing methods even when optimal tuning parameters are used (after comparing with the truth) for these methods. One possible reason could be that the NICE algorithm thresholds the correlation matrix based on the topological structure rather than a universal shrinkage or thresholding strategy. The difference of performance decreases when n and ρ are smaller. Moreover, we find that (i) adaptive thresholding performs better than the universal thresholding, and (ii) among precision matrix shrinkage methods ACLIME shows more robust and improved performance comparing to CLIME and *glasso* due to its minmax optimal convergence rate. Additionally, our approach is the only method can automatically detect the underlying $G^1 \cup G^0$ mixture topological structure. When the graph topological structure does not exist, the performance of all methods

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.

Table 2
Comparisons of large covariance/precision matrix estimation results: medians along with 25% and 75% quantiles of FP and FN.

Method		$\rho = 0.3, n = 25$				$\rho = 0.5, n = 25$				$\rho = 0.5, n = 50$				$\rho = 0.7, n = 25$			
	Tuning	FP		FN		FP		FN		FP		FN		FP		FN	
	Par.	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles
Glasso	0.1	4800	(4800, 4800)	0	(0, 0)	1673	(1648, 1702)	59	(55, 64)	1621	(1591.5, 1640)	44	(40, 46)	1581.5	(1557, 1606)	45.5	(42, 48)
	0.2	4800	(4800, 4800)	0	(0, 0)	1008.5	(989, 1025)	59	(53.5, 64.5)	630	(610, 644)	38	(33.5, 43)	932.5	(920, 955.5)	36	(32, 40)
	0.3	4128	(3959, 4322)	0	(0, 4.5)	546	(529.5, 560)	56	(48, 63.5)	151	(141, 162.5)	38	(30.5, 43)	500.5	(490, 516)	28	(23.5, 33)
	0.4	10	(2, 25)	120	(84, 143)	211.5	(200.5, 222.5)	60	(50.5, 72)	19	(16, 21)	48.5	(38, 58)	47	(41.5, 54)	28	(22.5, 35)
	0.5	(251, 727)	55.5	(25, 76.5)	51	(46, 59)	80.5	(66, 96)	1	(0, 2)	82.5	(67, 96.5)	525	194	(186, 204.5)	24.5	(20, 29)
	0.6	0	(0, 1)	146.5	(137, 149)	7	(5, 10)	112.5	(97, 125.5)	0	(0, 0)	130	(118.5, 137)	6	(5, 8.5)	41	(31, 51)
	0.7	0	(0, 0)	150	(149, 150)	0	(0, 1)	140	(131.5, 146)	0	(0, 0)	149	(147, 150)	0	(0, 1)	75	(61.5, 89)
	0.8	0	(0, 0)	150	(150, 150)	0	(0, 0)	149	(148, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	127	(119.5, 135)
	0.9	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	149	(149, 150)
	1.0	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)
Thres	0.1	2306	(2250.5, 2334)	11	(5, 16)	2017.5	(1963.5, 2067.5)	0	(0, 2)	1978.5	(1944.5, 2021.5)	0	(0, 0)	2021.5	(1968.5, 2061)	0	(0, 1)
	0.3	181	(161.5, 198)	79	(61, 94.5)	1292.50	(1252, 1331)	2	(0, 5)	1249.5	(1220.5, 1288.5)	0	(0, 0)	1293.5	(1251, 1341.5)	1	(0, 3)
	0.5	3	(2, 5)	138	(129.5, 145)	721.5	(699, 752)	5	(1, 12)	689	(673.5, 721)	0	(0, 1)	722	(693, 756)	3	(1, 10.5)
	0.7	0	(0, 0)	150	(149, 150)	344.5	(325, 360)	14	(7, 26.5)	328.5	(311.5, 349.5)	1	(0, 2)	342.5	(324, 363)	10	(3, 21.5)
	0.9	0	(0, 0)	150	(150, 150)	132	(121, 143.5)	30	(18, 45)	129.5	(121, 142)	3	(1, 7)	133	(123.5, 146)	24	(12, 39.5)
	1.1	0	(0, 0)	150	(150, 150)	41.5	(35, 46)	55.5	(40, 78.5)	40.5	(36.5, 47.5)	10	(4.5, 17)	40	(35.5, 46.5)	49.5	(28, 63)
	1.3	0	(0, 0)	150	(150, 150)	9	(6, 10)	92	(74, 112)	10	(8, 12)	25	(13, 37)	9	(6, 11)	78	(54.5, 89)
	1.5	0	(0, 0)	150	(150, 150)	1	(0, 2)	126	(112.5, 137)	2	(1, 3)	50.5	(32.5, 68)	1	(0, 2)	106	(92.5, 114)
	1.7	0	(0, 0)	150	(150, 150)	0	(0, 0)	145	(138.5, 148)	0	(0, 0)	85.5	(67, 102.5)	0	(0, 0)	132.5	(120.5, 138.5)
	1.9	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(149, 150)	0	(0, 0)	120.5	(105, 130)	0	(0, 0)	147	(144, 149)
AThres	0.3	3641	(3623, 3679)	3	(1, 6)	2593	(2566.5, 2627.5)	2	(0, 5)	2538.5	(2509.5, 2571)	0	(0, 0)	2594	(2563, 2619)	1	(0, 3)
	0.5	1749	(1693, 1788)	15.5	(9, 23.5)	1460	(1421.5, 1486)	5	(1, 12)	1412.5	(1379.5, 1440)	0	(0, 1)	1453	(1419.5, 1491)	3	(1, 10.5)
	0.7	639	(594, 665)	41.5	(27.5, 56)	691.5	(667, 717)	14	(7, 26.5)	668.5	(646, 697)	1	(0, 2)	695.5	(665.5, 720)	10	(3, 21.5)
	0.9	180	(161, 196)	79	(61, 94.5)	271.5	(258, 291.5)	30	(18, 45)	265	(252, 283.5)	3	(1, 7)	270.5	(255.5, 288)	24	(12, 39.5)
	1.1	37	(31.5, 45)	112.5	(96, 127)	83	(75, 95)	55.5	(40, 78.5)	85	(75.5, 95.5)	10	(4.5, 17)	82	(74, 89.5)	49.5	(28, 63)
	1.3	6.5	(4, 9)	133	(123, 142)	18	(15, 21)	92	(74, 112)	22	(18.5, 25.5)	25	(13, 37)	18	(14.5, 22)	78	(54.5, 89)
	1.5	1	(0, 1)	145	(140, 148)	2	(1, 4)	126	(112.5, 137)	4	(3, 6)	50.5	(32.5, 68)	18	(14.5, 22)	78	(54.5, 89)
	1.7	0	(0, 0)	149	(147, 150)	0	(0, 0)	145	(138.5, 148)	0	(0, 1)	85.5	(67, 102.5)	3	(1, 3)	106	(92.5, 114)
	1.9	0	(0, 0)	150	(149, 150)	0	(0, 1)	150	(149, 150)	0	(0, 0)	120.5	(105, 130)	0	(0, 0)	132.5	(120.5, 138.5)
CLIME	0.1	1073	(1069.5, 1075)	117	(113.4, 119)	1082.5	(1047.5, 1108)	56	(48, 64.5)	993.5	(981, 1024)	39	(32, 45.5)	1054	(1021, 1079)	48.5	(40, 56)
	0.2	430	(427, 431.5)	137	(134, 138.5)	353	(339.5, 367.5)	79.5	(69, 87.5)	241.5	(231.5, 251.5)	61	(54, 67.5)	345	(328, 359)	70	(59, 78)
	0.3	130	(128, 131)	146	(144, 147)	63	(57, 69)	110	(98.5, 115)	25	(22, 29)	92	(84.5, 100)	64	(59, 68)	98	(87, 103)
	0.4	29	(29, 30)	149	(149, 150)	0	(0, 1)	140	(135, 144)	0	(0, 0)	130	(124, 135)	0	(0, 1)	134	(129, 139)
	0.5	0	(0, 0)	150		0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)	0	(0, 0)	150	(150, 150)
ACLIME	2	1	(0, 2)	138	(124, 143.5)	34	(17, 30)	76	(58, 91)	2	(1, 4.5)	29.5	(16, 51.5)	47	(35, 60)	5	(2, 8.5)
NICE	None	0	(0, 0)	77.5	(35, 148)	42	(14, 92)	3	(0, 22)	9	(1, 32)	0	(0, 3)	0	(0, 0)	0	(0, 14)

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.

Table 3
Comparisons of large covariance matrix estimation methods using the Frobenius norm of matrix loss: medians along with 25% and 75% quantiles.

Method	$\sigma = 0.3, n = 50$		$\sigma = 0.5, n = 25$		$\sigma = 0.5, n = 50$		$\sigma = 0.7, n = 25$	
	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles	Med.	Quantiles
Thresh	11.27	(11.27, 11.27)	11.25	(11.24, 11.27)	11.27	(11.27, 11.27)	11.25	(11.24, 11.27)
AThresh	6.15	(6, 6.29)	7.77	(7.52, 8.12)	6.35	(6.22, 6.58)	9.49	(8.27, 10.85)
NICE	5.4	(4.67, 5.61)	6.95	(6.01, 8.2)	5.16	(4.7, 5.75)	8.98	(6.84, 10.17)

are similar across all settings. We note that methods with the sparsity assumption and conditional independence may miss many connected edges (false negative discovery rates are higher) even when small tuning parameter is used (false positive rates are high). Therefore, when a latent topological structure exists the sparsity assumption may not be valid because a cluster of features within a network are all correlated with each other and many of them can be conditionally independent by data sample calculation.

Last, we compare the performance of large covariance matrix estimation using the Frobenius norm $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{True}\|_F$. The large covariance matrix thresholding methods (Thresh and Athresh) are used for the comparison because we simulate data using the true covariance matrix (instead of the precision matrix) based on the known network topological structure. Table 3 shows the comparison results. The matrix loss of NICE is smaller than the Thresh and Athresh (based on their best performance across all tuning parameters) because NICE can fully leverage the additional information from the detected latent network topological structure.

In summary, the numerical results demonstrate that our method can not only provide more accurate estimation of the correlation matrix and the edge set E, but also automatically detect latent networks where highly correlated edges distribute in an organized topological structure.

4. Discussion and conclusion

We develop NICE approach to bridge large non-sparse covariance matrix estimation and underlying graph topological structure detection via a flexible empirical Bayesian framework. Recognizing the latent network topological structure can effectively reveal underlying biological pathway and further provide informative guidance to the decision making procedure of regularization.

Organized network topological structures exist widely in high-throughput biomedical data across platforms, however, the conventional network detection and clustering algorithms may not detect them due to the impact of false positive noises. For instance, a few false positive edges may lead to detecting a large networks with low proportion of highly connected edges as the conventional objective functions often do not penalize the sizes of networks. The proposed penalized network estimation objective function can identify the mixture structure by applying a new graph size penalty term. In general, the NICE latent network topology detection step can be applied before performing any large covariance/precision matrix estimation algorithms because it can assist to examine the model assumptions (e.g. sparsity) and guide the regularization decision by providing 'spatial' closeness information of edges. If no network topological structure is detected, the performance of the NICE algorithm would be similar to the existing large covariance/precision matrix estimation methods. In addition, efficient optimization algorithms of the penalized objective function is developed, and thus it is ready to scale up for larger data sets and regularly used.

The new Bayes factor based thresholding approach naturally incorporates detected network topological structure from step one as prior knowledge. The updated thresholding values are determined by each edge's 'location' on the detected graph topological 'map'. Therefore, edges can borrow strengths with each other with higher precision based on the detected topological structure, which also provides a flexible pathway to account for the dependencies between edges. With additional information from the detected topological structure and appropriate modeling strategy, our new thresholding approach reduces false positive and false negative rates simultaneously when topological structures exist. Clearly, the performance of graph topological structure detection influences the accuracy of correlation matrix thresholding because it determines the empirical distributions of $z_{i,j}^{in}$ and $z_{i,j}^{out}$ and thus $\widehat{\theta}_{in}$ and $\widehat{\theta}_{out}$. Therefore, the two steps of the NICE algorithm are seamlessly connected as the parsimonious property of the network detection ensures the efficiency and accuracy of the following regularization step.

Limitations: as NICE heavily depends on the topological structure detection, the proposed method may not perform well when the no latent network topological structure exists or the structure is not clustered (e.g. the Bandable and Toeplitz structures). In these settings, existing methods should be applied (Fan et al., 2015; Cai et al., 2016a). Therefore, we would recommend performing exploratory data analysis and permutation tests (see the Appendix Algorithm 1) to ensure that the network topological structure is correctly detected.

In summary, the NICE method provides a viable solution to estimate interactive relationships between massive biomedical features when underlying organized network structures exist. It can effectively capture the latent network structures and thus accurately estimate interactive relationships between features inside networks, while maintain outside network performance similar to those methods for large sparse covariance/precision matrix estimation. The numerical studies and

data example applications have demonstrated excellent performance of the NICE approach regarding false positive/negative findings and latent network detection. In our application, only positive (correlation) edges are distributed in an organized graph topology and the negative (correlation) edges are randomly distributed and thus negative (correlation) edges are thresholded. Nevertheless, our method is ready to be extended to the scenario that negatively correlated edges show a organized topological structure. The codes and example data are available at https://github.com/shuochenstats/Network_program/tree/master/NICE_folder.

Uncited references

Chen et al. (2015), von Luxburg (2007)

Acknowledgments

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via DJF-15-1200-K-0001725.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.05.008.

References

Baechler, E.C., Batliwalla, F.M., Karypis, G., …, Gregersen, P.K., 2003. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc. Natl. Acad. Sci. 100 (5), 2610–2615.
Banerjee, O., El Ghaoui, L., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. 9, 485–516.
Barnard, J., McCulloch, R., Meng, X., 2000. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statist. Sinica 10, 1281–1311.
Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82 (4), 733–746.
Bickel, P.J., Chen, A., 2009. A nonparametric view of network models and Newman–Girvan and other modularities. Proc. Natl. Acad. Sci. 106 (50), 21068–21073.
Bickel, P.J., Levina, E., 2008. Covariance regularization by thresholding. Ann. Statist. 36 (6), 2577–2604.
Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008 (10), P10008.
Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. J. Amer. Statist. Assoc. 106 (494), 672–684.
Cai, T., Liu, W., Luo, X., 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J. Amer. Statist. Assoc. 106, 594–607.
Cai, T.T., Liu, W., Zhou, H.H., 2016a. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. Ann. Statist. 44 (2), 455–488.
Cai, T.T., Ren, Z., Zhou, H.H., 2016b. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electron. J. Stat. 10 (1), 1–59.
Chen, S., Bowman, F.D., Mayberg, H.S., 2016. A Bayesian hierarchical framework for modeling brain connectivity for neuroimaging data. Biometrics 72 (2), 596–605.
Chen, S., Kang, J., Xing, Y., Wang, G., 2015. A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. Hum. Brain Mapp. 36 (12), 5196–5206.
Chen, S., Li, M., Hong, D., Billheimer, D., Li, H., Xu, B.J., Shyr, Y., 2009. A novel comprehensive wave-form MS data processing method. Bioinformatics 25 (6), 808–814.
Cui, Y., Leng, C., Sun, D., 2016. Sparse estimation of high-dimensional correlation matrices. Comput. Statist. Data Anal. 93, 390–403.
Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. J. Amer. Statist. Assoc. 99, 96–104.
Efron, B., 2007. Size, power and false discovery rates. Ann. Statist. 35 (4), 1351–1377.
Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., …., Mouy, M., 2008. Genetics of gene expression and its effect on disease. Nature 452 (7186), 423.
Fan, J., Feng, Y., Xia, L., 2016. A conditional dependence measure with applications to undirected graphical models. arXiv preprint arXiv:1501.01617.
Fan, J., Liao, Y., Liu, H., 2015 Estimating Large Covariance and Precision Matrices. arXiv preprint arXiv:1504.02995.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 603–680.
Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486 (3), 75–174.
Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostat. 9 (3), 432–441.
Jeffreys, H., 1961. Theory of Probability, third ed. Clarendon Press, Oxford.
Karrer, B., Newman, M.E., 2011. Stochastic block models and community structure in networks. Phys. Rev. E 83 (1), 016107.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90 (430), 773–795.
Khondker, Z.S., Zhu, H., Chu, H., Lin, W., Ibrahim, J.G., 2013. The Bayesian covariance lasso. Stat. Interface 6 (2), 243.
Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37, 42–54.
Lee, B.S., Jayathilaka, G.L.P., Huang, J.S., Vida, L.N., Honig, G.R., Gupta, S., 2011. Analyses of in vitro nonenzymatic glycation of normal and variant hemoglobins by MALDI-TOF mass spectrometry. J. Biomol. Tech.: JBT 22 (3), 90.
Liu, H., Wang, L., Zhao, T., 2014. Sparse covariance matrix estimation with eigenvalue constraints. J. Comput. Graph. Statist. 23, 439–459.
Mazumder, R., Hastie, T., 2012. Exact covariance thresholding into connected components for large-scale graphical lasso. J. Mach. Learn. Res. 13 (1), 781–794.
Nadakuditi, R.R., Newman, M.E., 2012. Graph spectra and the detectability of community structure in networks. Phys. Rev. Lett. 108 (18), 188701.
Newman, M.E., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. 103 (23), 8577–8582.
Pons, P., Latapy, M., 2006. Computing communities in large networks using random walks. J. Graph Algorithms Appl. 10 (2), 191–218.
Qi, H., Sun, D., 2006. A quadratically convergent newton method for computing the nearest correlation matrix. SIAM J. Matrix Anal. Appl. 28, 360–385.
Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Statist. Assoc. 104, 177–186.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4 (1).
Scott, J.G., Berger, J.O., 2006. An exploration of aspects of Bayesian multiple testing. J. Statist. Plann. Inference 136 (7), 2144–2162.
Shen, X., Pan, W., Zhu, Y., 2012. Likelihood-based selection and sharp parameter estimation. J. Amer. Statist. Assoc. 107, 223–232.
Tan, K.M., Witten, D., Shojaie, A., 2015. The cluster graphical lasso for improved estimation of Gaussian graphical models. Comput. Statist. Data Anal. 85, 23–36.
von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416.
Witten, D.M., Friedman, J.H., Simon, N., 2011. New insights and faster computations for the graphical lasso. J. Comput. Graph. Statist. 20 (4), 892–900.
Wu, B., Guan, Z., Zhao, H., 2006. Parametric and nonparametric FDR estimation revisited. Biometrics 62 (3), 735–744.
Yildiz, P.B., Shyr, Y., Rahman, J.S., Wardwell, N.R., Zimmerman, L.J., Shakhtour, B., …, Massion, P.P., 2007. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J. Thorac. Oncol. 2 (10), 893–915.
Yuan, M., 2010. High dimensional inverse covariance matrix estimation via linear programming. J. Mach. Learn. Res. 11, 2261–2286.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the gaussian graphical model. Biometrika 94, 19–35.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 89, 4–942.

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.