Submitted to the Annals of Applied Statistics

NETWORK INDUCED LARGE COVARIANCE MATRIX ESTIMATION

By Shuo Chen* , Yishi Xing , and Donald Milton

University of Maryland, College Park

In this paper, we consider to estimate network/community induced large covariance matrices. The massive biomedical data (e.g. gene expression or neuroimaging data) often include networks where features are correlated with each other, and thus the covariance matrix in a complex yet organized topology. Although, current large covariance matrix and precision matrix estimation methods using thresholding or shrinkage strategies could provide satisfactory overall covariance/precision matrix estimation, they are limited for automatic network/topology detection and network induced covariance matrix estimation. To fill the gap, we propose a new network induced covariance estimation method (NICE) to simultaneously detect highly correlated networks and estimate the covariance matrix by leveraging an adaptive and graph topology oriented thresholding strategy. The novel thresholding strategy can reduce both false positive and false negative discovery rates by using the whole graph topological information which allows edges to borrow with each other. Moreover, we propose a novel 'quality and quantity' objective function to shrink the covariance matrix towards a parsimonious model while retaining most of the information. Simulation study results show that our method outperform the competing thresholding and shrinkage methods. We further illustrate the application of our new method by analysis of a serum mass spectrometry proteomics data set.

1. Introduction. We consider a large data set $X_{n\times p}$ with the sample size n and the feature dimensionality of p. The estimation of the covariance matrix (Σ) is fundamental to understand the inter-relationship between variables of the large data set $X_{n\times p}$ (Fan *et al*, 2015).

When the dimensionality is high (p is large), the estimation by using sample covariance is known to have poor performance and regularization is needed. Therefore, ℓ_1 regularized sparse inverse covariance selection methods (e.g. graphical lasso) have been developed to estimate the precision matrix Σ^{-1} assuming that $X_{n\times p}$ follows a multivariate normal distribution and the inverse covariance matrix $\Omega = \Sigma^{-1}$ is sparse (Friedman *et al*, 2008; Banerjee

*Corresponding author: shuochen@umd.edu

Keywords and phrases: graph, large covariance matrix, network, srhinkage, thresholding, topology

et al, 2008; Yuan and Lin, 2007; Lam and Fan, 2009; Yuan, 2010; Cai *et al*, 2011). In addition, the covariance matrix thresholding methods have been developed to directly regularize the sample covariance matrix (Bickel and Levina, 08; Rothman *et al*, 2009; Cai and Liu, 2011; Fan *et al*, 2013; Liu *et al*, 2014). Mazumder and Hastie (2012) and Witten *et al* (2011) find the two sets of methods are naturally linked regarding vertex-partition of the whole graph.

The regularization strategies including both covariance matrix thresholding and inverse covariance matrix shrinkage methods are often implemented on individual edges rather than considering the interaction between edges and the whole graph topology. Hence, a universal thresholding value or regularization standard is applied to to all edges without accounting for the dependency between edges. Cai and Liu (2011) propose adaptive thresholding methods by considering the column-wise marginal standard deviations, yet such thresholding strategy still applies the same decision rule to each column independently without accounting for the whole graph topology.

Notations: Graph theory notations and definitions are often used to delineate the relationship between the p variables of $X_{n\times p}$ (Yuan and Lin, 2007; Mazumder and Hastie, 2012). A finite undirected graph $G = \{V, E\}$ consists two sets, where the vertex set V represents the variables $\mathbf{X} = (X_1, \cdots, X_p)$ with $|V| = p$ and the edge set E denote relationships between the vertices. In practice, the edge set E (our main interest) is represented by a symmetric 0 and 1 adjacency matrix. Two nodes i and j are directly connected if $E_{ij} = 1$, otherwise unconnected. Under the sparsity assumption, the regularization algorithms assign most edges as 0s, and G is decomposed to a set of maximal connected subgraphs $G = \cup_{k=1}^{K} G_k$ and $\cap_{k=1}^{K} G_k = \emptyset$ with $G_k = \{V_k, E_k\}$ (Witten *et al*, 2011; Mazumder and Hastie, 2012).

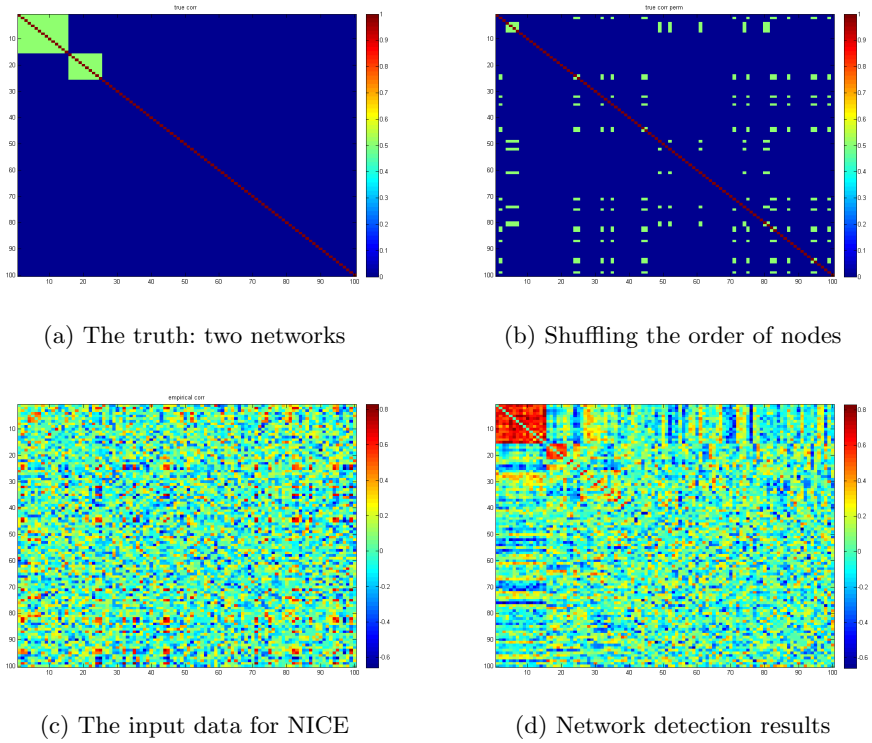
It has been widely recognized that the interactions between genes and neural processing units exhibit organized network graph topological properties (i.e. non Erdös-Rényi random graph), and it is crucial to recognize the topological patterns in a data-driven fashion. In this paper, we assume that subsets of variables of $X_{n\times p}$ display community/network structures, and variables are more correlated with variables within the same network than the others. Therefore, the truly correlated edges (i.e. $E_{ij} = 1$ for thoses edges should not be regularized) in the covariance matrix are G not randomly distributed rather in an (latent) organized fashion, for instance in several latent subgraphs and in each subgraph $\{\sum_{e_{ij}^k \in E_k} I(e_{ij}^k = 1)\}/|E_k|$ is high. Such model structure is naturally linked to the stochastic block model (SBM, Bickel and Chen, 2009; Karrer and Newman, 2011; Zhao *et al*, 2011; Choi *et al*, 2012; Nadakuditi and Newman, 2012; Lei and Rinaldo, 2014).

Conventionally, we apply SBM to binary graph data sets (e.g. social network data) to detect communities by using algorithms based on profile likelihood, spectral graph, and modularity maximization (Karrer and Newman, 2011; Zhao *et al*, 2011; Nadakuditi and Newman, 2012). However, the high-dimensional biomedical data tend to be different from the SBM model because it is likely that not all features belong to communities. Therefore, we consider a model structure that $G = G_1 \cup G_0$ where subgraph G_1 follows a SBM and G_0 is a Erdös-Rényi random graph. Hence, we are not only interested in community detection but also learning the whole graph topological structure which we rely on for the following adaptive regularization towards the network induced covariance matrix. We propose a new **N**etwork **I**nduced **C**ovariance **E**stimation (NICE) method to detect the graph topology (including latent community detection) and estimate for network induced covariance via graph topology oriented regularization.

We implement the NICE model in three steps by jointly using several advanced statistical techniques. In step 2 (see section 2.2 for details), we propose a novel statistical strategy/concept of graph topology based regularization, which is distinct from the current popular methods (e.g. ℓ_1 or ℓ_2 penalty terms). The new strategy aims to capture most informative edges in minimal communities, and naturally incorporates automatic selection of the optimal number of communities with regularization level selection in a data driven fashion. In addition, we introduce permutation test based inference framework to identify the true communities. All these new techniques in turn could be good additions to the SBM (and community detection) methods and theories.

The paper is organized as follows. Section 2 describes the NICE algorithm, followed by the simulation studies and model evaluation/comparison in Section 3. In Section 4, we apply our method to mass spectrometry proteomics data data, concluding remarks are summarized in Section 5.

2. Methods. We propose a new network induced covariance matrix estimation method to 1) detect the maximal connected subgraphs $G = \cup_{k=1}^{K} G_k$ as networks of interest, and 2) to estimate Σ. We consider the sample covariance $\mathbf{S}$ and the correlation matrix $\mathbf{R}$=diag$(\mathbf{S})^{-1/2}$ $\mathbf{S}$ diag$(\mathbf{S})^{-1/2}$ as our input data (Qi and Sun, 2006; Liu *et al*, 2014; Fan *et al*, 2015). Directly regularizing the correlation matrix $\mathbf{R}$ could provide estimation of the binary edge E matrix by using $\hat{E}_{ij} = I(R_{ij}^{\mathcal{T}} > 0)$, and $\hat{E}$ determines the set of maximal connected components $G = \cup_{k=1}^{K} \hat{G}_k$ ($\hat{G}_k$ could be a singleton). However, applying a universal regularization/thresholding rule to each element (or column) would introduce false positives and false negatives for



(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 covariance matrix: $|V|$=100 nodes and $|E|$=4950 edges, there are two networks (a) and in practice they are implicit (b) especially hard to recognize when looking at the sample covariance matrix (c); however, with the knowledge/estimation of topological network structures detected by NICE (d) the regularization strategy should take them into account.

$\hat{E}$ and lead to failure to detect networks of interest. Therefore, we propose the NICE method with a heuristic to taking the topological structures into account regarding regularization. Figure 1 illustrates an example synthetic data and some results of our new method.

The NICE method consists three steps: i) calculate the posterior probability weight matrix $\mathbf{W} = g(\mathbf{R})$ with $\mathbf{W}_{ij} = \text{Prob}(E_{ij} = 1|\mathbf{R})$ as a fuzzy logic metric; ii) detect the maximal connected components $G = \cup_{k=1}^{K} \hat{G}_k$ by using a 'quality and quantity rule'; iii) applying the adaptive threholding rules to the edges within outside networks. To illustrate our goal, we use a synthetic example: suppose there are two networks of interest where all

nodes are highly correlated, yet the networks are implicit : the nodes in the networks are not originally adjacent but randomly distributed.

2.1. *Calculate posterior probability based fuzzy logic weight matrix* $\boldsymbol{W}$. Rather than directly thresholding/binarizing the sample correlation matrix $\mathbf{R}$, we calculate a fuzzy logic metric $W_{ij} = \text{Prob}(E_{ij} = 1|\mathbf{R})$ (Chen *et al*, 2015a). The fuzzy logic metric avoids arbitrary choice of the cut-off threshold value and provides a more appropriate scale for maximal connected component detection. Let z_{ij} be equal to the Fisher's Z transformed correlation coefficient R_{ij} for example (Kendall's Tau or other pairwise relationship metrics could also be applied), and z_{ij} follows two distinct distributions dependent on E_{ij} that $z_{ij}|(E_{ij} = 0) \sim f_0(z_{ij})$ and $z_{ij}|(E_{ij} = 1) \sim f_1(z_{ij})$. The f_1 represents the distribution of correlations corresponding to connected edges, and f_0 for the unconnected edges. In practice, the label E_{ij} is not available, and thus the sample correlations follow a mixture distribution $z_{ij} \sim p_0 f_0(z_{ij}) + p_1 f_0(z_{ij})$ with $p_0 + p_1 = 1$. We adopt the empirical Bayes method to obtain $\hat{p}_0, \hat{p}_1, \hat{f}_0, \hat{f}_1$, and we would refer the readers for the details to the original works (Efron, 2004; Wu *et al*, 2006; Efron, 2007). The Bayes posterior probability that a case belongs to the connected edge set given z_{ij}, by definition, is:

$$\begin{aligned} E(W_{ij}) &\equiv \text{Prob}\{E_{ij} = 1\} = p_1 f_1(z_{ij})/f(z) \\ &= (f(z) - p_1 f_1(z_{ij}))/f(z) \\ &= 1 - \text{fdr}(z_{ij}) \end{aligned} \tag{2.1}$$

In practice, we calculate the fuzzy logic metric $W_{ij} = 1 - \widehat{\text{fdr}}(z_{ij})$ based on the estimated $\widehat{\text{fdr}}$ by using the existing statistical package e.g. '*lcfdr*' in the R software.

Therefore, the Bayes posterior probability based fuzzy logic metric provides a non-parametric transformation to map the original correlation metric value of z_{ij} to a probability based 0 to 1 scale (the probability of $E_{ij} = 1$ and should not be thresholded). The distribution of W_{ij} often shows a large proportion of edges with the fuzzy logic values close to zero, and a small proportion greater than zero and close to one, which clearly improves the signal to noise ratio and greatly enhances the following graph topology structure detection. Furthermore, the fuzzy logic metric avoids to binarize the edges and instead passes the posterior belief of each edge being connected to the following analysis without information loss.

2.2. *Graph topological structure detection.* In step 2, we seek to estimate the maximal connected components $\{\hat{G}_k\}$ that $\cup_{k=1}^{K}\hat{G}_k = G$ based on the input fuzzy logic metric weight matrix $\mathbf{W}$, which is equivalent to a generalized clustering problem (SBM G_1 and random graph G_0 separation and then community detection). Under the assumption that the G include induced complete subgraphs (networks) as shown in 1a, the edges within the networks are 1 and outside the network are 0 and thus the fuzzy logic metrics $W_{ij}|(E_{ij} = 1)$ are distributed in the same pattern. However, the networks are often implicitly embedded in G and the sample correlation or covariance matrix has no explicit network structure (1c). Thus, step 2 aims to automatically recognize such graph topological structures (from 1c to 1d). One straightforward pathway to achieve this goal is to 1) first shuffle the nodes by allocating the connected ones to be adjacent to each other in order, and 2) identify each network by cutting the edges connecting the network with the rest of the graph. In fact, to perform 2) would automatically implement 1) above because only connected nodes could be adjacent in order. Then, our goal turns out to be a edge cutting problem, likely in glasso and adaptive thresholding we also need to determine which edge to threshold or set to zero. The ideal edge cutting results will provide both community detection results and G_1 and G_0 separation because nodes in G_0 could be cut out as singletons and edges connecting nodes in G_1 and elsewhere will be cut as well.

One naive strategy is to cut edge by applying a threshold for the continuous fuzzy logic metric and wish that the topological structure will show itself. However, it is well known that the sample correlation/covariance matrix could include a large proportion of false positive and negative noises even after the transformation by step 1. In practice, the conditional fuzzy logic metric $W_{ij}|(E_{ij} = 0)$ values could be false positively high and $W_{ij}|(E_{ij} = 1)$ could be false positively low because of the noises. Such noises will make the graph topology detection more complex and difficult. When applying covariance matrix thresholding rules (e.g. Bickel and Levina, 08; Friedman *et al*, 2008; and Cai and Liu, 2011), and the thresholding results could mislead the detection of the true topological structures because such noises. For a toy example, suppose G_{k1} and G_{k2} are the only two separate networks in G, and $E_{ij} = 1|(E_{ij} \in G_{k1}$ or $E_{ij} \in G_{k2})$ and $E_{ij} = 0$ otherwise. Due to the noises, some values of W_{ij} ($i \in V_{K1}$ and $j \in V_{K2}$) could be false positively high, and thus it is possible the individual edge/column based decision rules would produce a few false positive result of $\hat{E}_{ij} = 1|W_{ij}$. Let $|V_1| = |V_2| = 20$ and there are 400 edges between the two networks, and we assume the false positive discovery rate is 1% (very low), then the probability of failure to

separate the two networks is equal to $1-(1-0.99)^{400}=0.982$ (assuming that the edges are independent). In addition, the other nodes (not in G_1 and G_2) could be falsely connected to the networks G_{k1} and G_{k2}, and thus the accurate network detection could become a formidable task. Therefore it is almost impossible for the graph topological structures to show themselves by direct thresholding even the universal thresholding tuning parameter (λ) is optimal. Therefore, we need to take the whole graph topological structure into consideration for edge-cutting and develop an algorithm robust to the substantial amount of false positive and false negative noises.

2.2.1. *Spectral clustering for edge cutting.* Spectral clustering algorithms have been applied to optimize the edge cutting problem for example using Ratio-Cut and Normalized Cut algorithms (von Luxburg, 2007; Nadakuditi and Newman, 2012). Given an appropriate number of clusters K, the spectral clustering methods provide a promising solution to cut edges and detect networks. However, different from the traditional goal of spectral clustering that allocates all nodes into K classes, we try to detect a graph G that could only include a handful of networks (G_1) and the rest of G can be considered as a Erdös-Rényi random graph (G_0). To detect such graph topological structure, we not only need to cut the edges between communities but also most edges in G_0 and between G_0 and G_1 because highly correlated edges in the random graph are distributed randomly (with no community structure) and thus G can be considered as a set of singletons. To cut edges of G with the above topological structure, the NormalizedCut algorithm seems not to work because it may fail to detect many singletons of the the Erdös-Rényi random graph (von Luxburg, 2007). We construct the objective function based on the RatioCut algorithm. The edge cutting objective function is to minimize:

$$\text{argmin}_{\{G_k\}_{k=1}^K} \sum_{k=1}^{K} \frac{\sum_{i\in G_k, j\notin G_k} W_{ij}}{|V_k|}, \tag{2.2}$$

It minimizes the fuzzy logic metric values of the edges (W_{ij}) between networks (including network of size one). The denominator $|V_k|$ prevents to generate with one major community and $K-1$ singletons. Thus, the spectral clustering based decision rule is to cut those between network edges whose sum of fuzzy logic metric values is minimum. The network detection and edge cutting are obtained simultaneously when implementing minimization of 2.2. It is NP complex to implement the optimization of 2.2, and fortunately computational algorithms have been successfully developed for solutions (Shi and Malik, 2000; Chen *et al*, 2015b). It has been established

(Chung, 1997) that the 2.1 is equivalent to

$$\operatorname{argmin} \sum_{k=1}^{K} h_i L h_i = \operatorname{Tr}(H' L H), \tag{2.3}$$

where $H_{p\times K}$ is indicator matrix and a column h_k of H is a binary $p \times 1$ vector (elements with entry value 1 indicate that they belong to the kth network). Estimating $H_{p\times K}$ provides the network detection results. L is the Laplacian matrix, which is defined by:

$$L = D - \mathbf{W}, \tag{2.4}$$

where D=diag$(\sum_j^n W_{ij})$ and $i = 1, \cdots, p$. We implement the optimization of 2.4 to estimate $H_{p\times K}$ by using unnormalized spectral clustering (von Luxburg, 2007) and details are provided in the algorithm table. Furthermore, Lei and Rinaldo (2014) provides proof of consistency with regard to spectral clustering for SBM. Thus, the estimated $\hat{G}$ is consistent to G, and when K is appropriately selected both communities and singletons can be detected.

2.2.2. *K selection and regularization: a quantity and quality criterion.* It is crucial to select the number of K, because it not only influences the allocation of nodes but also the number of edges to cut. Cutting an edge highly increases the chance to perform hard thrsholding on that edge, and thus K determines the regularization stringent level naturally (with similar function to the parameter λ of glasso). For example, if K = $|V|$ then all edge will be cut in G (then G is an Erdös-Rényi random graph) while if K=1 then no edge will be cut. Thus, it is important to select K wisely because it is the key to reveal the true graph topological structure and perform graph topology oriented regularization. Our heuristic to choose K is to maximally include informative edges (W_{ij} with larger values) into community networks (the quantity criterion) when ensuring that the detected networks has a high proportion of informative edges (the quality criterion). We consider W_{ij} with larger values as informative edges because the edges with higher W_{ij} values are more rare and more likely for E_{ij} to be one (generally $Prob(E_{ij} = 1) > Prob(E_{i'j'} = 1)$ if $W_{ij} > W_{i'j'}$). Therefore, the quantity and quality criterion could help to select the K that is able to capture the graph topological structure if informative edges are non-randomly distributed. To implement the quantity and quality criterion, we define an objective function for K selection:

$$\frac{\sum_{k=1}^{K}\sum_{i\in G_k, j\in G_k} W_{ij}}{\sum_{i<j} W_{ij}} \cdot \frac{\sum_{k=1}^{K}\sum_{i\in G_k, j\in G_k} W_{ij}}{\sum_{k=1}^{K}\sum_{i\in G_k, j\in G_k} 1}. \tag{2.5}$$

The first term (the proportion of informative edges included in the network with contrast to the informative edges in the whole graph) reflects the quantity criterion and the second term for the quality criterion (the proportion of informative edges in the networks with contrast to the total number of edges in those networks). The quantity criterion ensure to detect the networks when informative edges are in organized structures (power), and the quality criterion ensure the selected networks with all nodes well connected (to reduce false positive network detection). In practice, if the dimension p is massive W_{ij} could be replaced by $I(W_{ij} > p_0)$ (e.g. $p_0 = 0.05$) to avoid the accumulation of noises (e.g. false positive errors). In addition, since we apply different regularization rules for correlation metrics inside and outside networks (details in the following subsection), thus the regularization is closely linked to the selection of K. We implement K optimization by grid searching because K can only be integers from two to n. Therefore, the NICE algorithm selects K objectively, and the regularization stringent level is based on the optimized K because our following regularization rule is guided by the detected topological structures.

After performing the edge cutting step using spectral clustering algorithm with optimized K, we obtain the estimated clusters as $\cup_{k=1}^{K}\hat{G}_k = G$ (where many $\hat{G}_k$ could be singletons). We could further examine the detected community networks with size greater than two are true networks by using permutation tests (see details in the algorithm table). The underlying heuristic is that if the detected network is a true community, then there are higher proportion of edges with $E_{ij} = 1$ within the detected network and $\sum_{i,j\in G_k} W_{ij}$ is larger than the whole graph average level. We consider a detected network is a true community structure when the permutation p value is less than α/K_s where α is significance level (e.g. 0.05) and K_s is number of non-singleton clusters. Since the number of edges is at the power order of number of nodes, the true communities tend to have very small p-values and even the stringent Bonferroni correction (for multiple network testing adjustment) has very slight impact on the results. When p is large, there could be many natural large blocks in G (Witten *et al*, 2011), and we could apply our graph topological structure detection method to each large block.

2.3. *Graph topology oriented threshohlding: a empirical Bayesian model with informative priors.* The graph topological structure detection results

could provide us prior knowledge which edges in G are connected. The prior probability follow:

$$\text{Prob}(E_{ij=1}|(i \in G_k, j \in G_k) = 1,$$

where G_k is the detected community network with significant permutation test p value. The prior probability for the rest of edges $\text{Prob}(E_{ij=1}) = 0$. Based on the prior knowledge from the graph topology structure detection, we could estimate the probability distributions of Fisher's Z transformed correlations inside and outside the networks as $\hat{f}_{in}(z_{ij})$ and $\hat{f}_{out}(z_{ij})$, and the proportions of edges inside and outside of the communities $\hat{p}_{in}$ and $\hat{p}_{out}$. This procedure is more straightforward than the mixture model estimation in $lcfdr$ because the labels of edges are given based on the graph topological structure. Thus, our thresholding rule with accounting for the induced network topological structure is:

If $i \in G_k$ and $j \in G_k$,

$$\hat{\rho}_{ij} = \begin{cases} R_{ij} & \text{if } \hat{f}_{in}(z_{ij}) \geq \hat{f}_{out}(z_{ij}); \\ 0 & \text{otherwise.} \end{cases}$$

else if $i \notin G_k$ or $j \notin G_k$,

$$\hat{\rho}_{ij} = \begin{cases} R_{ij} & \text{if } \hat{p}_{in}\hat{f}_{in}(z_{ij}) \geq q\hat{p}_{out}\hat{f}_{out}(z_{ij}); \\ 0 & \text{otherwise.} \end{cases}$$

In general, the peaks of $\hat{f}_{in}(z_{ij})$ and $\hat{f}_{out}(z_{ij})$ are apart and the first is much larger than the latter. The inside-network edges with smaller z_{ij} (at the left tail of $\hat{f}_{in}(z_{ij})$) are thresholded as 0s, and the outside-network edges with larger z_{ij} (at the right tail of $\hat{f}_{out}(z_{ij})$) are unchanged. Therefore, we apply a graph topology oriented adaptive hard-thresholding strategy (Donoho *et al*, 1995; Bickel and Levina, 08; Cai and Liu, 2011). For the outside network edges, the decision rule is similar to the hard-thresholding, but with a different thresholding value is used. The thresholding value is closely linked to the local fdr (Efron, 2007), yet fortunately our network detection step provides explicit labels for all edges and thus the null component and the non-null component are pre-determined. The constant q is related to the local fdr threshold value, and if the local fdr threshold is 0.2 (suggested by Efron, 2007), then q=4. For the inside network edges, we use a different thresholding rule which is less stringent (than outside network), because we have a stronger prior belief that for inside network edges $E_{ij} = 1$. The prior belief is based on the previous network detection step, and it has verified by statistical tests that the graph topology of these networks are non-random

but organized. Therefore, the graph topology information is used for thresholding and both false positive and false negative discovery rates, as the thresholding is not performed on a individual edge but based on the whole graph topology structure.

We determine both inside and outside network edge threshold values by using Bayes factor (BF, Kass and Raftery, 1995). The posterior probability distribution of all edges are $\hat{p}_{in}\hat{f}_{in}(z_{ij}) + \hat{p}_{out}\hat{f}_{out}(z_{ij})$. For outside edges, the BF threshold is $\hat{f}_{in}(z_{ij})/\hat{f}_{out}(z_{ij}) \geq \hat{p}_{out}/\hat{p}_{in} \times q$; but for the side edges, the BF threshold is $\hat{f}_{in}(z_{ij})/\hat{f}_{out}(z_{ij}) \geq 1$. The difference reflects the prior belief odds ratio:

$$\frac{Prob(E_{inside} = 1)/Prob(E_{inside} = 0)}{Prob(E_{outside} = 1)/Prob(E_{outside} = 0)} = \hat{p}_{out}/\hat{p}_{in} \times q,$$

where the only tuning parameter is q (4 is recommended by Efron, 2007). The prior information incorporates the network allocation (i.e. topological structure information) and influences the final thresholding and regularization via a Bayesian framework.

Remarks: the statistical inferences on large covariance matrices involve multiple aspects such as topological structures and covariance matrix estimation. It is limited to simply apply a one (or two) step universal decision rule for knowledge/data mining. Recognizing the networks provides a fundamental understanding of the interactive relationships between multivariate variables. In return the topological structure could become prior knowledge to lead topological structure assisted large covariance matrix regularization and estimation.

3. Simulation Studies. We conduct numerical studies to evaluate the performance of NICE algorithm for detecting the correlated networks and estimating covariance matrix, and compare it with the other popular large covariance matrix shrinkage and thresholding method.

3.1. *Simuluation datasets.* We simulate a data set with $p = 100$ variables, and thus $|V| = 100$ and $|E| = \binom{100}{2} = 4950$. We assume that the covriance matrix includes two induced correlated networks, and the first include 15 nodes and the second 10 nodes. The induced networks are complete subgraphs 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 sets that the network structures are implicit (Fig 1b), which represents the true edge set E. Let X_p follow a multivariate normal distribution, with zero mean and covariance

Algorithm 1 NICE algorithm

procedure NICE–ALGORITHM
  Obtain the empirical Bayes fuzzy logic matrix **W**=g(**R**)
  Calculate the Laplacian matrix L=D-**W**
  for cluster number $K = 2 : |V| - 1$ **do**
    Compute the first K eigenvectors $[u_1, \cdots, u_K]$ of L, with eigenvalues ranked from the smallest.
    Let $U = [u_1^T, \cdots, u_K^T]$ be a $|V| \times K$ matrix containing all K eigenvectors.
    Perform K-means clustering algorithm on U with K to cluster $|V|$ nodes into K networks
    Calculate the quality and quantity criterion for each K.
  end for
  Adopt the clustering results using the K of the maximum score of the quality and quantity criterion.
  Identify the networks with significantly high proportion of larger W values by permutation test: for each network
i) calculate the $T^0 = \sum_{i,j \in G_k} W_{ij}$;
ii) permute the labels of the nodes in the detected network for M (e.g. 10,000) times and calculate T^m for each iteration;
iii) calculate the percentile of T^0 in $\{T^m\}$, if it is less than the α level the network is considered as true community network
  Implement the topological structure oriented thresholding strategies for covariance entries inside and outside networks (see details in 2.3)
end procedure

matrix $\Sigma_{p\times p}$, and the sample size is N. σ_{ij} is an entry at the ith row and jth column of Σ, $\sigma_{ij} = 1$ if $i = j$, and $\sigma_{ij} = \rho | E_{ij} = 1$ for edges within the networks and $\sigma_{ij} = 0 | E_{ij} = 0$ for edges outside the networks. We simulate 100 data sets at different signal to noise (SNR) levels by using sample sizes N and different values of ρ (Fig 1c) as a) a larger sample size reduces the asymptotic variance of $\hat{\sigma}_{ij}$ and thus the noise level is lower (for larger N); b) a higher absolute value of ρ represents higher signal level. We compare our method with glasso, CLIME, and adaptive thresholding by comparing the false positive and negative rates of $\hat{E}_{ij}$ with contrast to the ground truth regarding the network detection and edge set E estimation. We assess the performance of each method by estimating the number of false positive (FP) edges $\hat{E}_{ij} = 1 | \hat{E}_{ij} = 0$ and false negative (FN) edges $\hat{E}_{ij} = 0 | \hat{E}_{ij} = 1$. In our simulated data sets, 150 edges are connected $E_{ij} = 1$ and 4800 edges are unconnected $E_{ij} = 0$. For inverse covariance matrix shrinkage and covariance matrix thresholding methods we treat the the non-zero entry as $\hat{E}_{ij} = 1$ and then summarize the FP and FN edges because they provide the same estimate of E (Mazumder and Hastie, 2012).

TABLE 1
Median and quantiles of FP and FN

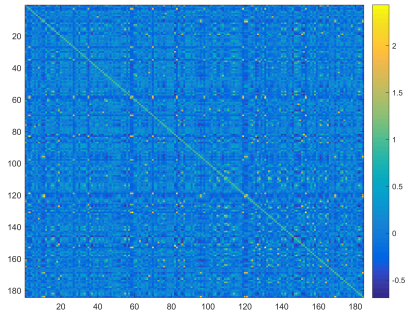
Parameters	$\sigma = 0.5$	*sample size* $= 25$	$\sigma = 0.5$	*sample size* $= 50$	$\sigma = 0.7$	*sample size* $= 25$
	FP	FN	FP	FN	FP	FN
$glasso_{0.1}$	1673(1648,1702)	59(55,64)	1621(1591.5,1640)	44(40,46)	1581.5(1557,1606)	45.5(42,48)
$glasso_{0.2}$	1008.5(989,1025)	59(53.5,64.5)	630(610,644)	38(33.5,43)	932.5(920,955.5)	36(32,40)
$glasso_{0.3}$	546(529.5,560)	56(48,63.5)	151(141,162.5)	38(30.5,43)	500.5(490,516)	28(23.5,33)
$glasso_{0.4}$	211.5(200.5,222.5)	60(50.5,72)	19(16,21)	48.5(38,58)	194(186,204.5)	24.5(20,29)
$glasso_{0.5}$	51(46,59)	80.5(66,96)	1(0,2)	82.5(67,96.5)	47(41.5,54)	28(22.5,35)
$glasso_{0.6}$	7(5,10)	112.5(97,125.5)	0(0,0)	130(118.5,137)	6(5,8.5)	41(31,51)
$glasso_{0.7}$	0(0,1)	140(131.5,146)	0(0,0)	149(147,150)	0(0,1)	75(61.5,89)
$glasso_{0.8}$	0(0,0)	149(148,150)	0(0,0)	150(150,150)	0(0,0)	127(119.5,135)
$glasso_{0.9}$	0(0,0)	150(150,150)	0(0,0)	150(150,150)	0(0,0)	149(149,150)
$glasso_{1}$	0(0,0)	150(150,150)	0(0,0)	150(150,150)	0(0,0)	150(150,150)
$CLIME_{0.1}$	1082.5(1047.5,1108)	56(48,64.5)	993.5(981,1024)	39(32,45.5)	1054(1021,1079)	48.5(40,56)
$CLIME_{0.2}$	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)
$CLIME_{0.3}$	63(57,69)	110(98.5,115)	25(22,29)	92(84.5,100)	64(59,68)	98(87,103)
$CLIME_{0.4}$	0(0,1)	140(135,144)	0(0,0)	130(124,135)	0(0,1)	134(129,139)
$CLIME_{0.5}$	0(0,0)	150(150,150)	0(0,0)	150(150,150)	0(0,0)	150(150,150)
$Thres_{0.1}$	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)
$Thres_{0.3}$	1292.50(1252,1331)	2(0,5)	1249.5(1220.5,1288.5)	0(0,0)	1293.5(1251,1341.5)	1(0,3)
$Thres_{0.5}$	721.5(699,752)	5(1,12)	689(673.5,721)	0(0,1)	722(693,756)	3(1,10.5)
$Thres_{0.7}$	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)
$Thres_{0.9}$	132(121,143.5)	30(18,45)	129.5(121,142)	3(1,7)	133(123.5,146)	24(12,39.5)
$Thres_{1.1}$	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)
$Thres_{1.3}$	9(6,10)	92(74,112)	10(8,12)	25(13,37)	9(6,11)	78(54.5,89)
$Thres_{1.5}$	1(0,2)	126(112.5,137)	2(1,3)	50.5(32.5,68)	1(0,2)	106(92.5,114)
$Thres_{1.7}$	0(0,0)	145(138.5,148)	0(0,0)	85.5(67,102.5)	0(0,0)	132.5(120.5,138.5)
$Thres_{1.9}$	0(0,0)	150(149,150)	0(0,0)	120.5(105,130)	0(0,0)	147(144,149)
$AThres_{0.3}$	2593(2566.5,2627.5)	2(0,5)	2538.5(2509.5,2571)	0(0,0)	2594(2563,2619)	1(0,3)
$AThres_{0.5}$	1460(1421.5,1486)	5(1,12)	1412.5(1379.5,1440)	0(0,1)	1453(1419.5,1491)	3(1,10.5)
$AThres_{0.7}$	691.5(667,717)	14(7,26.5)	668.5(646,697)	1(0,2)	695.5(665.5,720)	10(3,21.5)
$AThres_{0.9}$	271.5(258,291.5)	30(18,45)	265(252,283.5)	3(1,7)	270.5(255.5,288)	24(12,39.5)
$AThres_{1.1}$	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)
$AThres_{1.3}$	18(15,21)	92(74,112)	22(18.5,25.5)	25(13,37)	18(14.5,22)	78(54.5,89)
$AThres_{1.5}$	2(1,4)	126(112.5,137)	4(3,6)	50.5(32.5,68)	3(1,3)	106(92.5,114)
$AThres_{1.7}$	0(0,0)	145(138.5,148)	0(0,1)	85.5(67,102.5)	0(0,0)	132.5(120.5,138.5)
$AThres_{1.9}$	0(0,1)	150(149,150)	0(0,0)	120.5(105,130)	0(0,0)	147(144,149)
$NICE$	44(15,98)	3(0,27)	11(1,30)	0(0,4)	32.5(13.5,71)	14(4,38.5)

3.2. *Numerical Results.* The simulation results are summarized in Table 1. Rather than selecting a single tuning parameter λ for glasso by cross-validation, we explore all possible choices in the reasonable range. We utilize 25%, 50%, and 75% of the FP and FN edges of the 100 simulation data sets to evaluate the performance of the methods. The results show the NICE algorithm outperform the other methods even when comparing with the optimal tuning parameters. One possible reason could be the NICE algorithm thresholds the covariance matrix based on the topological structure rather than the a universal shrinkage or thresholding strategy. More importantly, our NICE method is the only method can automatically detect the underlying network structures. The NICE method also outperform the other models. Thus, the numerical results demonstrate that our new method not only provides more accurate estimation of the covariance matrix and the edge set E than the competing method, but also automatically detects the networks where high connectivity edges distribute in an organized fashion.

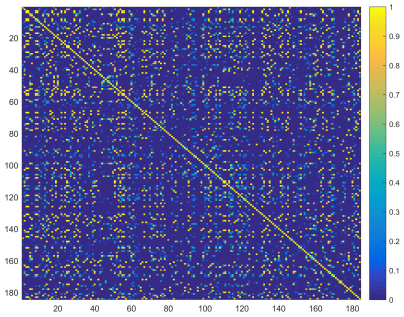
4. Data example. We further apply our method to a publicly available mass spectrometry proteomics data set (Yildiz *et al*, 2007). The study collected matrix-assisted laser desorption ionization mass spectrometry (MALDI

MS) data sets to obtain the most abundant peptides in the serum that may distinguish lung cancer cases from matched controls. The study included 182 subjects in the training data set and 106 in the testing data set. The raw data were 288 mass spectra for all subjects, and each raw spectrum consists roughly 70,000 data points. After preprocessing steps including MS registration, wavelets denoising, alignment, peak detection, quantification, and normalization (Chen *et al*, 2009), 184 features are considered 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 intended to utilize the proteomics data to enhance disease diagnosis and prediction. In this paper, we estimate the covariance/correlation matrix to investigate the relationship between these features. We use the training data set to calculate the sample correlation matrix as our input data $\mathbf{X}_{n\times p}$.

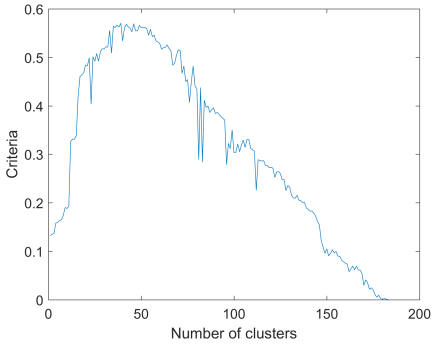
Next, we apply our NICE method to estimate the covariance matrix and to detect the correlated peptide/protein networks based on the (Fisher's Z transformed) sample correlation matrix (2a). Next, we calculate the fuzzy logic weight matrix as shown in 2b. The quality and quantity criterion is implemented and K=39 is selected, and the procedure is demonstrated in 2c. The network detection is performed to cut edges, allocate correlated features to each other 2d. Based on the permutation test, 8 networks are detected. Furthermore, inside and outside network edges show distinct distributions of Fisher's Z transformed correlations 2e. We apply the topology/network oriented adaptive decision/thresholding rules to estimate $\hat{E}$ and the covariance/correlation matrix 2f. Note that 2e only identify two components of positively correlated edges and the null, without the negatively correlated edge component. As a result, the negative edges are thresholded. Since the null distribution in 2e seems to be symmetric, based on the Bayes factor decision rule we are confident that edges with high negative correlations are false positive. The network detection results could provide informative inferences about the between feature relationship. In this example data set, each network represents a group of related protein and peptides that could be confirmed by proteomics mass spectrometry literature. For example, the most correlated network 4 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 15127, 15868, 7564, and 7934). The highly correlated networks of biomedical features could potentially provide a set of biomarkers for future research that allow to borrow power between each other.



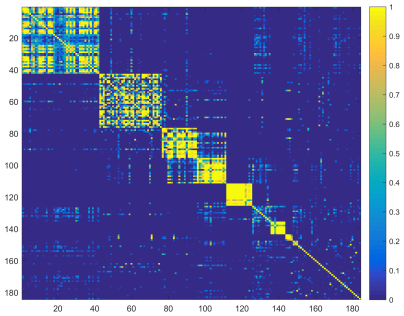
(a) Sample correlation



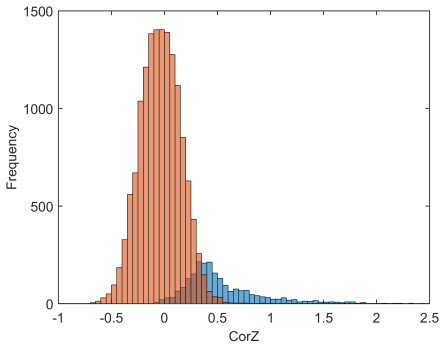
(b) Fuzzy logic weight matrix



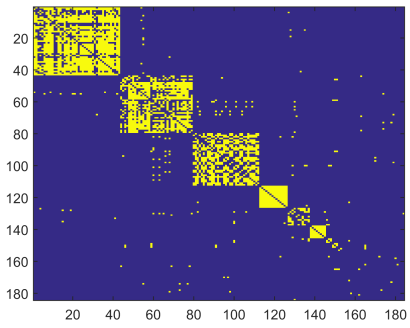
(c) K selection



(d) Network detection: redordered W.



(e) Edges inside and outside networks



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

Fig 2: Application of the NICE to the example data set.

5. Discussion and Conclusion. The large covariance/correlation matrix of data sets from high-throughput biomedical assays often demonstrate complex, yet highly organized, topology of the underlying physiological and biological machinery. There have been unmet needs of statistical methodologies to simultaneously estimate the covariance matrix and reveal the underlying topological structure. The regularization methods (e.g. shrinkage or thresholding) are exploited to estimate the covariance matrix. On the hand, the community detection algorithms (e.g. profile likelihood and modularity maximization based on SBM) are employed to explore the graph topological structures, and the results heavily depend on two factors: the similarity metrics and number of clusters. We develop a novel strategy to bridge the covariance estimation and graph topological structure for network induced covariance matrix estimation.

The NICE framework at least makes three novel contributions. First, the fuzzy logic metric provides a sensitive and robust scale to detect the networks because the accuracy of clustering findings primarily depends the similarity metrics. Second, the quality and quantity criterion is a efficient and data-driven heuristic not only to objectively select optimal tuning parameter (i.e. K), but also to guide regularization with graph topology oriented adaptive thresholding stringent levels. A larger K will cut more edges and keep less edges within the networks, and the edges outside and inside of the networks are subject to more stringent thresholding levels because 1) the number of inside of edges are less; and 2) the mean of two distributions are more separate (see 2e). Note that different from all community detection algorithms, our selected tuning parameter K (by the quantity and quality criterion) does not determine the final number of detected networks and a network must pass the permutation test significance level. For example, our example set has K=39 but only 8 networks are detected. Therefore, the quality and quantity criterion is a new regularization computational strategy which is distinct from the lasso type ℓ_1 and ℓ_2 shrinkage (e.g. glasso) and adaptive thresholding algorithms because it implement the regularization based on the topology structure constraint and select optimal tuning parameter objectively (less ad-hoc). Last, our covariance matrix estimation (thresholding strategy) could reduce both false positive and false negative discovery rate by leveraging the graph topological structures. The adaptive thresholding rules depends on the inherent topology of G, for example, if the graph is totally random our strategy would be similar to hard thresholding (Bickel and Levina, 08). In this case, the univariate correlation thresholding method seems to provide satisfactory performance (Friedman *et al*, 2010). In our applications, only positive (correlation) edges are organized in graph topol-

ogy and the negative (correlation) edges are randomly distributed. Based on the Bayes factor decision rule, none of negative (correlation) edges are suprathrhold. In the future, we will examine whether the graph topology structures could include organized negative (correlation) edges, and negative (correlation) edges and positive (correlation) edges could jointly comprise organized topology structures.

The simulation studies and example data set application have demonstrated excellent performance of the NICE algorithm. The computational cost of NICE algorithm is low (for our simulation example the algorithm only takes 40 seconds using i7 CPU and 24G memory), and thus it is ready to scale up for larger data sets. In addition, the NICE algorithm is not restricted for multivariate Gaussian distributed data and it is straightforward to extend the sample correlation matrix to other sample metrics, for example maximal information coefficients (**?**) for continuous data and polychoric correlation coefficient for categorical data (**?**) because both fuzzy logic metric and graph topology oriented thresholding are based on the empirical distribution of the coefficients.

Acknowledgements. 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.

References.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research* **9**, 485-516.

Bickel, P.J., Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, no. 6, 2577–2604.

Bickel, P. J., Chen, A. (2009). A nonparametric view of network models and NewmanGirvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**(50), 21068-21073.

Bonett, D. G., Price R. M. (2005). Inferential Methods for the Tetrachoric Correlation Coefficient. *Journal of Educational and Behavioral Statistics*, **30**, 213.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594–607.

Cai, T., Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** (494), 672–684.

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.

Chen, S., Kang, J., Wang, G. (2015). An empirical Bayes normalization method for connectivity metrics in resting state fMRI. *Frontiers in neuroscience*, **9**, 316-323.

Chen, S., Kang, J., Xing, Y., Wang, G. (2015). A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Human brain mapping*, **36**(12), 5196-5206.

Choi, D. S., Wolfe, P. J., Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, **99**, 273-284.

Chung, F. R. (1997). Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92), American Mathematical Society.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 301-369.

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, **99**, 96-104.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, **35**(4), 1351-1377.

Efron, B., Turnbull, B., Narasimhan, B. (2008). locfdr: Computes local false discovery rates. *R package*, 195.

El Karoui, N. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: risk underestimation. *The Annals of Statistics*, **38**, 3487–3566.

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**, no. 4, 603–680.

Fan, J., Liao, Y., Liu, H. (2015). Estimating Large Covariance and Precision Matrices. arXiv preprint arXiv:1504.02995.

Friedman, J., Hastie, T., Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. (2008).*Biostat.* **9**(3), 432–441.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models (pp. 1-22). Technical report, Stanford University.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* **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. *Journal of biomolecular techniques: JBT*, **22**(3), 90.

Lei, J., Rinaldo, A. (2014). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, **43**(1), 215-237.

Liu, H., Wang, L. and Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, **23**, 439-459.

Kass, R. E., Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, **90**(430), 773-795.

Karrer, B., Newman, M. E. (2011). Stochastic block models and community structure in networks. *Physical Review E*, **83**(1), 016107.

Kinney, J. B., Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, **111**(9), 3354-3359.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, **13**(1), 781-794.

Nadakuditi, R. R., Newman, M. E. (2012). Graph spectra and the detectability of community structure in networks. *Physical review letters*, **108**(18), 188701.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177-186.

Qi, H. and Sun, D. (2006). A quadratically convergent newton method for computing the nearest correlation matrix. S*IAM journal on matrix analysis and applications* **28** 360-385.

Shi, J., Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(8), 888-905.

von Luxburg, U. A tutorial on spectral clustering. (2007), *Stat. Comput.* **17** (4), 395–416.

Witten, D. M., Friedman, J. H., Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, **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. *Journal of thoracic oncology*, **2**(10), 893-915.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, **94**, 19–35.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, **11**, 2261-2286.

Zhao, Y., Levina, E., Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, **108**(18), 7321-7326.

Shuo Chen and Yishi Xing
Department of Epidemiology and Biostatistics
University of Maryland, College Park
E-mail: shuochen@umd.edu
E-mail: ustc.xingys@gmail.com

Donald Milton
Maryland Institute for Applied Environmental Health
University of Maryland, College Park
E-mail: dmilton@umd.edu