Bioinformatics, 2016, 1–10
doi: 10.1093/bioinformatics/btw410
Advance Access Publication Date: 29 June 2016
Original Paper

OXFORD

Systems biology

Network-based pathway enrichment analysis with incomplete network information

Jing Ma,[1,*] Ali Shojaie[2] and George Michailidis[3]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, PA 19104, USA, [2]Department of Biostatistics, University of Washington, Seattle, WA 98915, USA and [3]Department of Statistics, University of Florida, Gainesville, FL 32611, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 22, 2016; revised on June 21, 2016; accepted on June 22, 2016

Abstract

Motivation: Pathway enrichment analysis has become a key tool for biomedical researchers to gain insight into the underlying biology of differentially expressed genes, proteins and metabolites. It reduces complexity and provides a system-level view of changes in cellular activity in response to treatments and/or in disease states. Methods that use existing pathway network information have been shown to outperform simpler methods that only take into account pathway membership. However, despite significant progress in understanding the association amongst members of biological pathways, and expansion of data bases containing information about interactions of biomolecules, the existing network information may be incomplete or inaccurate and is not cell-type or disease condition-specific.

Results: We propose a constrained network estimation framework that combines network estimation based on cell- and condition-specific high-dimensional Omics data with interaction information from existing data bases. The resulting pathway topology information is subsequently used to provide a framework for simultaneous testing of differences in expression levels of pathway members, as well as their interactions. We study the asymptotic properties of the proposed network estimator and the test for pathway enrichment, and investigate its small sample performance in simulated and real data settings.

Availability and Implementation: The proposed method has been implemented in the R-package netgsa available on CRAN.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: jinma@upenn.edu

1 Introduction

Recent advances in high-throughput technologies have transformed biomedical research by enabling comprehensive monitoring of complex biological systems. By profiling the activity of different molecular compartments (genomic, proteomic, metabolomic), one can delineate complex mechanisms that play key roles in biological processes or the development of distinct phenotypes. These technological advances have thus motivated new methodological developments, most notably the adaptation of systems perspectives to analyze biological systems. Pathway analysis represents a key component in the analysis process and has been used successfully in generating new biological hypotheses, as well as in determining whether specific pathways are associated with particular phenotypes. Examples include analysis of pathways involved in initiation and progression of cancer and other complex diseases (Wilson *et al.*, 2010), discovering novel transcriptional effects and co-regulated genes (Green *et al.*, 2011), and understanding the basic biological processes in model organisms (Gottwein *et al.*, 2007; Houstis *et al.*, 2006). See Huang *et al.* (2008) for additional examples of applications.

Pathway analysis methods have evolved since the seminal work by Subramanian *et al.* (2005). As pointed out in the review article by Khatri *et al.* (2012), earlier techniques such as over-representation analysis (Al-Shahrour *et al.*, 2005), and gene set analysis (GSA) (Efron and Tibshirani, 2007; Subramanian *et al.*, 2005) treat each pathway as a set of biomolecules. These methods assess whether members of a given pathway have higher than expected levels of activity, either by counting the number of differentially active members, or by also accounting for the relative rankings of pathway members and/or the magnitude of their associations with the phenotype. On the other hand, more recent and statistically powerful methods also account for interactions between biomolecules. These interactions are increasingly available from carefully curated biological databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa and Goto, 2000), Reactome (Joshi-Tope *et al.*, 2003), RegulonDB (Huerta *et al.*, 1998) and BioCarta (Nishimura, 2001).

A network topology-based method that exhibits superior statistical power in identifying differential activity of pathways was proposed in Shojaie and Michailidis (2009, 2010). The Network-based Gene Set Analysis (NetGSA) method also allows testing for potential changes in the network structure under different experimental or disease conditions. However, it requires *a priori* knowledge of interactions among pathway members, which, despite rapid progress, remains highly incomplete and occasionally unreliable (see e.g. Zaki *et al.*, 2013 and references therein). Moreover, existing network information often determine molecular interactions in the normal state of the cell, and do not provide any insight into condition/disease-specific alterations in interactions amongst components of biological systems.

The increased availability of large sets of high-dimensional Omics data [e.g. from The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/], coupled with the development of network estimation techniques based on graphical models (Lauritzen, 1996) offers the possibility to validate and complement existing network information, and to obtain condition-specific estimates of molecular interactions. Such an approach for leveraging existing knowledge to enhance the analysis of low signal-to-noise biological datasets was advocated in Ideker *et al.* (2011).

The first contribution of this article is the development of a method for constrained network estimation from high-dimensional data, together with establishing the consistency of the resulting estimate. Estimation of high-dimensional networks subject to hard (or soft) constraints on conditional dependence relationships amongst random variables represents a canonical problem in the context of graphical models, and the proposed method for addressing this problem is of independent interest. By incorporating the condition-specific network estimates from the proposed method into the NetGSA framework, we also provide a rigorous statistical framework for assessing alterations in biological pathways, referred to as *differential network biology* (Ideker and Krogan, 2012).

The proposed framework accounts for two sources of uncertainty: the first concerns the reliability of the external information used for constructing the network estimate from data. The second is the variability of the network estimate, which can impact the pathway enrichment testing procedure. We establish that, under certain regularity conditions, consistent estimates of the network can be obtained, leading, in turn, to an asymptotically most power unbiased test for pathway enrichment analysis. Our theoretical analysis also sheds light into the potential improvements in accuracy and power by directly accounting for the amount of reliable external network information.

A second objective of this study is to scale up the NetGSA estimation algorithm to very large size networks. The main bottleneck in applying the NetGSA methodology arises from the estimation of mixed effects linear parameters—specifically the variance components—for thousands of variables. We develop efficient and stable computational methods for estimation of these parameters based on a profile likelihood approach. In particular, we employ a Cholesky factorization of the covariance matrices to speed up matrix inversions, and use it to develop a stable algorithm based on Newton's method with backtracking line search (Boyd and Vandenberghe, 2004: 487) for step size selection. To supply reliable starting points for this algorithm, we further develop an approximate method-of-moment-type estimator.

The proposed methods are illustrated on both metabolomics and gene expression data. For mass spectrometry metabolomics profiling one can obtain good quality measurements for a few hundred metabolites that do not provide complete coverage of the underlying biochemical pathways. The small number of metabolites in each pathway and the incomplete coverage of the metabolites particularly hinder the application of over-representation and GSA methods in this setting. In our experience, only topology-based pathway enrichment analysis methods, such as NetGSA, are capable of reliably delineating pathway activity, as illustrated in Section 4. Further, our investigation of previously analyzed gene expression data set on lung and breast cancer provides new useful insights.

The remainder of the article is organized as follows. Section 2.1 presents the new method for network estimation under external information constraints and establishes its consistency. Section 2.2 outlines the new computational algorithm for scaling up NetGSA, as well as the inference procedure for both pathway enrichment and differential network analysis. The performance of the developed methodology is evaluated in Section 3 and is examined on real data sets in Section 4.

2 Methods

Gaussian graphical models (Lauritzen, 1996, Chapter 5) are widely used in biological applications to model the interactions among components of biological systems (Dehmer and Emmert-Streib, 2008, Chapter 6). Specifically, partial correlation networks are commonly used to model interactions in molecular networks; these networks are represented by an undirected graph $G=(V,E)$ with node set V and edge set E corresponding to biomolecules interactions among them, respectively. The edge set E corresponds to the $p\times p$ precision, or inverse covariance, matrix $\mathbf{\Omega}$, whose nonzero elements $\omega_{ii'}$ refer to edges between nodes i and i', and indicate that i and i' are conditionally dependent given all other nodes in the network. The magnitude of the partial correlation $\mathbf{A}_{ii'}=-\omega_{ii'}/\sqrt{\omega_{ii}\omega_{i'i'}}$ determines the strength (positive or negative) of the conditional association between the respective nodes. In the sequel, the matrix $\mathbf{A}$ will also be called the weighted adjacency matrix, with $\mathbf{A}_{ii'}$ being the association weight between i and i'.

2.1 Network estimation under external information constraints

As discussed in Section 1, the availability of large collections of samples for different disease states and biological processes together with carefully curated information of biomolecular interactions enables the estimation of network structures within the setting of Gaussian graphical models. However, the availability of external network information provides a novel and unexplored modification

of the corresponding network estimation problem. Denote by E^c the set of node pairs not connected in the network, i.e. $\omega_{ii'} = 0$. Then, the external information can be represented by the following two subsets

$$E_1 = \{(i,i') \in E : i \neq i', \omega_{ii'} \neq 0\},$$
$$E_0 = \{(i,i') \in E^c : i \neq i', \omega_{ii'} = 0\}.$$

In words, E_1 contains known edges, while E_0 contains node pairs where it is known that no interaction exists between them. Note that $E_1 \subseteq E$ and $E_0 \subseteq E^c$. The external information available in E_1 does not imply exact knowledge of the magnitude of $\omega_{ii'}$ nor $\mathbf{A}_{ii'}$.

Suppose we observe an $m \times p$ data matrix $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$, where each row represents one sample from a p-variate Gaussian distribution $\mathcal{N}(0, \Omega^{-1})$ for a given biological condition (e.g. cancer or normal). Our goal is then to estimate the network structure, or equivalently the precision matrix Ω, subject to external information encoded in E_1 and E_0. Let $\mathbf{D} = \mathrm{diag}(\Omega)$ represent the diagonal matrix whose diagonal entries are the same as Ω and $\mathbf{I}_p$ be the p-identity matrix. Then, $\mathbf{A} = \mathbf{I}_p - \mathbf{D}^{-1/2}\Omega\mathbf{D}^{-1/2}$ is the partial correlation matrix. When $E_1 = E$ and $E_0 = E^c$, the problem becomes that of covariance selection (Dempster, 1972), which has been studied extensively in the literature. However, to the best of our knowledge, the problem of estimating Ω (and the partial correlation matrix $\mathbf{A}$) when E_1 and E_0 only contain partial information ($E_1 \subsetneq E$ and $E_0 \subsetneq E^c$) has not been investigated before.

In this section, we assume that the m observations used for estimating condition-specific networks are separate from those used for pathway enrichment analysis (highlighted by the use of $\mathbf{Z}_i$'s and m to denote the random variables and sample size, respectively). The framework introduced in this section reduces the potential bias in small sample settings, and takes advantage of the additional publicly available samples, in lieu of reliable network information. With large enough samples, network estimation and pathway enrichment can be performed using the same set of samples by incorporating sample splitting strategies. Although the problem considered in this section is seemingly similar to matrix completion (Candes and Recht, 2009), the two problems are fundamentally different in nature. In particular, in this setting, matrix completion corresponds to completing the remaining entries from the partially observed $p \times p$ matrix $\mathbf{A}$, under some structural assumptions on $\mathbf{A}$, such as low-rankness. On the other hand, in the setting of graphical models, the entries of the weighted adjacency matrix are estimated based on data on the nodes of the graph.

In biological settings, both the structure of the network, as well as strengths of associations may be condition-specific. Therefore, we need to accurately estimate the nonzero entries in Ω to recover both the structure of the network and the strength of associations between nodes. In the absence of any external information, the ℓ_1-penalized negative log-likelihood estimate of Ω is obtained by solving

$$\underset{\Omega \succ 0}{\mathrm{argmin}} \left\{ \mathrm{trace}(\Omega\widehat{\Sigma}) - \mathrm{logdet}\Omega + \lambda||\Omega||_1 \right\}, \tag{1}$$

wherein $\widehat{\Sigma} = \mathbf{Z}^T\mathbf{Z}/m$ is the empirical covariance matrix of the data, $||\Omega||_1 = \sum_{i \neq i'} |\omega_{ii'}|$ denotes the ℓ_1 norm of the parameters, and λ is the regularization parameter. In the presence of external information, the problem can be cast as the following constrained optimization one

$$\min_{\Omega \succ 0} \left\{ \mathrm{trace}(\Omega\widehat{\Sigma}) - \mathrm{logdet}\Omega \right\}, \tag{2}$$

subject to $\omega_{ii'} = 0$ for $(i,i') \in E_0$, $\omega_{ii'} \neq 0$ for $(i,i') \in E_1$, and $\sum_{i \neq i',\ (i,i') \notin E_0 \cup E_1} \omega_{ii'} | \leq t$.

In the following, we present a two-step procedure to solve the constrained optimization problem (2). The proposed approach combines the neighborhood selection technique (Meinshausen and Bühlmann, 2006) with constrained maximum likelihood estimation. It exploits the fact that the estimated neighbors of each node using neighborhood selection coincide with the nonzero entries of the inverse covariance matrix (Friedman *et al.*, 2008). Specifically, in neighborhood selection the network structure is estimated by finding the optimal set of predictors when regressing the random variable $\mathbf{Z}_i$ corresponding to node $i \in V$ on all other variables, using an l_1-penalized linear regression. The coefficients for this optimal prediction $\boldsymbol{\theta}^i$ are closely related to the entries of the inverse covariance matrix: for all $i' \neq i$, $\theta^i_{i'} = -\omega_{ii'}/\omega_{ii}$. The set of nonzero coefficients of $\boldsymbol{\theta}^i$ is thus the same as the set of nonzero entries in the row vector of $\omega_{ii'}$ ($i' \neq i$), which defines the set of neighbors of node i.

Let J^i_1 and J^i_0 denote the subsets $V \backslash i$ for which external information is available: J^i_1 is the set of nodes which are known to be in the neighborhood of i, and J^i_0 is the set of nodes which are known to be not connected to i. Let $\mathbf{Z}_{-i}$ denote the submatrix obtained by removing the ith column of $\mathbf{Z}$. Assume all columns of $\mathbf{Z}$ are centered and scaled to have norm 1. Denote by $\mathcal{S}^p_+$ the set of all $p \times p$ positive definite matrices and $\mathcal{S}^p_E = \{\Omega \in \mathbb{R}^{p \times p} : \omega_{ii'} = 0, \text{ for all } (i,i') \notin E \text{ where } i \neq i'\}$. The proposed algorithm proceeds in two steps.

i. Estimate the network structure $\widehat{E}$. For every node i, find $\widehat{\boldsymbol{\theta}}^i$ via the following steps.
   a. For $i' \in J^i_0$, set $\widehat{\theta}^i_{i'} = 0$.
   b. For $i' \in J^i_1$, find $\widehat{\theta}^i_{i'}$ using linear regression

$$\widehat{\boldsymbol{\theta}}^i_{J^i_1} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{|J^i_1|}}{\mathrm{argmin}} \frac{1}{m} ||\mathbf{Z}_i - \mathbf{Z}_{J^i_1}\boldsymbol{\theta}||^2_2. \tag{3}$$

   c. For $i' \in \tilde{J} \equiv V \backslash \{J^i_1 \cup J^i_0 \cup \{i\}\}$, find $\widehat{\theta}^i_{i'}$ using lasso

$$\widehat{\boldsymbol{\theta}}^i_{\tilde{J}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{|\tilde{J}|}}{\mathrm{argmin}} \frac{1}{m} ||\mathbf{W}_i - \mathbf{Z}_{\tilde{J}}\boldsymbol{\theta}||^2_2 + 2\lambda \sum_{i' \in \tilde{J}} |\theta_{i'}|, \tag{4}$$

   where $\mathbf{W}_i = \mathbf{Z}_i - \mathbf{Z}_{J^i_1}\widehat{\boldsymbol{\theta}}^i_{J^i_1}$ is the residual vector after regressing $\mathbf{Z}_i$ on the known connections.

   The edge set $\widehat{E}$ is estimated to be $\{(i,i') : \widehat{\theta}^i_{i'} \neq 0 \text{ OR } \widehat{\theta}^{i'}_i \neq 0\}$.

ii. Given the structure $\widehat{E}$, estimate the inverse covariance matrix $\widehat{\Omega}$ by

$$\widehat{\Omega} = \underset{\Omega \in \mathcal{S}^p_+ \cap \mathcal{S}^p_{\widehat{E}}}{\mathrm{argmin}} \left\{ \mathrm{trace}(\widehat{\Sigma}\Omega) - \mathrm{logdet}\Omega \right\}. \tag{5}$$

Remark 1. In step (i-b) of the algorithm, the coefficients $\boldsymbol{\theta}^i$ for known edges have not been penalized in (3). In settings where the external information may be unreliable, we can augment (3) with a lasso penalty $\lambda \sum_{i' \in J^i_1} t_{i'} |\theta_{i'}|$, where the penalty weights $t_{i'}$ ($i' \in J^i_1$) allow for different penalization depending on the reliability of existing information.

The second step focuses on estimation of the magnitude of nonzero entries in the precision matrix Ω, given the estimated network topology $\widehat{E}$. The optimization problems in both steps are convex and can be solved efficiently using existing software (e.g. glmnet and glasso in R).

The proposed estimator enjoys nice theoretical properties under certain regularity conditions. Before presenting the main result, we introduce some additional notations. Let Σ_0 be the true covariance matrix and $\Omega_0 = \Sigma_0^{-1}$. For $i = 1, \ldots, p$, denote by $||\boldsymbol{\theta}^i||_0 = \#\{i' : \theta^i_{i'} \neq 0\}$ the l_0 norm of $\boldsymbol{\theta}^i$. Write $s = \max_{i=1,\ldots,p} ||\boldsymbol{\theta}^i||_0$ and

$S_0 = \sum_{i=1}^{p} ||\boldsymbol{\theta}^i||_0$. For a subset $J \subset \{1, \ldots, p\}$, let $\mathbf{Z}_J$ be the submatrix obtained by removing the columns whose indices are not in J. We make the following assumptions.

Assumption 1. *There exist $\phi_1, \phi_2 > 0$ such that the eigenvalues of Σ_0 are bounded, i.e. $0 < \phi_2 \leq \phi_{\min}(\Sigma_0) \leq \phi_{\max}(\Sigma_0) \leq 1/\phi_1 < \infty$.*

Assumption 2. *There exists $\kappa(s) > 0$ such that*

$$\min_{|J| \leq s} \min_{\substack{\boldsymbol{\delta} \in \mathbb{R}^p \\ ||\boldsymbol{\delta}_{J^c}||_1 \leq 3||\boldsymbol{\delta}_J||_1}} \frac{1}{\sqrt{m}} \frac{||\mathbf{Z}\boldsymbol{\delta}||_2}{||\boldsymbol{\delta}_J||_2} \geq \kappa(s). \tag{6}$$

Assumption 1 is standard in high-dimensional settings. Assumption 2 corresponds to the restricted eigenvalue assumption introduced in Bickel *et al.* (2009), which is presented here for completeness.

Denote by $|E|$ the cardinality of the edge set E. Let $r \equiv (|E_0| + |E_1|)/\{p(p-1)/2\}$ represent the percentage of external network information available. Clearly, $0 \leq r < 1$. We are now ready to state our first result.

Theorem 1. Suppose Assumption 1 holds and Assumption 2 is satisfied with $\kappa(2s)$. For constants $c_1 > 4$ and $0 < k_1 < 1$, assume also that the sample size satisfies

$$m \geq \left\{ \frac{16c_1}{k_1 \phi_1 \kappa^2(2s)} \right\}^2 (1-r) S_0 \log(p - rp), \tag{7}$$

where S_0 is the total number of nonzero parameters excluding the diagonal. Consider $\widehat{\Omega}$ defined in (5). Then, with probability at least $1 - 2p^{2-c_1^2/8}$, under appropriately chosen λ, we have

$$||\widehat{\Omega} - \Omega_0||_2 \leq ||\widehat{\Omega} - \Omega_0||_F = O\left(\sqrt{\frac{S_0 \log(p - rp)}{m}} \right). \tag{8}$$

Remark 2. In addition to the improved sample complexity (7), the convergence rate in (8) indicates an improvement of the order of $\sqrt{S_0 \log(1-r)^{-1}/m}$ in the presence of external information. This improvement is particularly important for our analysis of power properties of NetGSA in Section 2.2.2, which requires norm consistency of adjacency matrix estimation. Although consistency can be established using a theoretical analysis similar to graphical lasso (Rothman *et al.*, 2008), our proofs in the Supplementary Materials Section A utilize the techniques from Bickel *et al.* (2009) and Zhou *et al.* (2011) to characterize the improvement in rates resulting from the external information.

Let $\mathbf{A}_0$ be the true partial correlation matrix, i.e. $\mathbf{A}_0 = \mathbf{I}_p - \mathbf{D}_0^{-1/2} \Omega_0 \mathbf{D}_0^{-1/2}$, where $\mathbf{D}_0 = \text{diag}(\Omega_0)$. The following corollary is an immediate result of Theorem 1.

Corollary 1. *Let assumptions in Theorem 1 be satisfied. Assume further that $S_0 = o(m/\log(p - rp))$. For $\widehat{\Omega}$ defined in (5), let $\widehat{\mathbf{A}}$ be the corresponding partial correlation matrix. Then, with probability at least $1 - 2p^{2-c_1^2/8}$, under appropriately chosen λ, we have*

$$||\widehat{\mathbf{A}} - \mathbf{A}_0||_2 = o(1).$$

Remark 3. Corollary 1 implies that, under certain regularity conditions, the error in the condition-specific network estimate $\widehat{\mathbf{A}}$ is negligible. This proves essential for establishing power properties of NetGSA with estimated network information, as shown in the next section. The proof of Corollary 1 is available in the Supplementary Materials Section A.

The tuning parameter λ in the first step of the proposed algorithm is important for selecting the correct structure of the network, which further affects the magnitude of the network interactions in the second step. Accurate estimation of these magnitudes is crucial for topology-based pathway enrichment methods. We propose to select λ using the Bayesian Information Criterion (BIC). Specifically, for a given λ, we define

$$\text{BIC}(\lambda) = \text{trace}(\widehat{\Sigma}\widehat{\Omega}_\lambda) - \text{logdet}(\widehat{\Omega}_\lambda) + \frac{\log(m)}{m} |\widehat{E}_\lambda|, \tag{9}$$

where $\widehat{\Omega}_\lambda$ is the estimated precision matrix from the data and $\widehat{E}_\lambda$ is the estimated edge set. The optimal tuning parameter is thus $\lambda^* = \text{argmin}_\lambda \text{BIC}(\lambda)$.

2.2 NetGSA with estimated network information

Next, we discuss how (condition-specific) estimates of bimolecular interactions from Section 2.1 can be incorporated into the NetGSA framework to obtain a rigorous inference procedure for both pathway enrichment and differential network analysis. To this end, we formally define the NetGSA methodology based on undirected Gaussian graphical models in Section 2.2.1. In Section 2.2.2, we discuss how the constrained-network estimation procedure of Section 2.1 can be combined with NetGSA to rigorously infer differential activities of biological pathways, as well as changes in their network structures.

2.2.1 The latent variable model

Consider p genes (proteins/metabolites) whose activity levels across n samples are organized in a $p \times n$ matrix $\mathcal{D}$. In the framework of NetGSA, the effect of genes (proteins/metabolites) in the network is captured using a latent variable model (Shojaie and Michailidis, 2009, 2010). Denote by $\mathbf{Y}$ an arbitrary column of the data matrix $\mathcal{D}$. Suppose the observed data can be decomposed into signal, $\mathbf{X}$, plus noise $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_p)$, i.e. $\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$. The latent variable model assumes that the signal $\mathbf{X}$ follows a multivariate normal distribution with partial correlation matrix $\mathbf{A}$. Based on the connection between linear recursive equations and covariance selection proposed in Wermuth (1980), there exists a lower triangular matrix Λ such that $\Lambda^{-1}\mathbf{X} = \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma_\gamma^2 \mathbf{I}_p)$ and $\Lambda\Lambda^T = (\mathbf{I}_p - \mathbf{A})^{-1}$. Note that the current version of the NetGSA model differs from the original model in Shojaie and Michailidis (2009, 2010). This difference is primarily manifested through the definition of Λ in the two models: Λ is defined here based on the *undirected partial correlation network* $\mathbf{A}$, whereas it was previously defined based on directed (physical) interactions among genes (proteins/metabolites) in Shojaie and Michailidis (2009, 2010).

Assuming that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are independent, the NetGSA model can then be summarized as

$$\mathbf{Y} = \Lambda\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{10}$$

The NetGSA methodology allows for more complex models, including time course observations. For expositional clarity, we present the methodology in the setting of two experimental conditions and consider the general case where $\mathbf{A}^{(1)} \neq \mathbf{A}^{(2)}$. Details of NetGSA under multiple conditions can be found in Shojaie and Michailidis (2010) and are applicable for the undirected networks presented in this work. Let $\mathbf{Y}_j^{(k)}$ $(j = 1, \ldots, n; k = 1, 2)$ be the j-th sample in the expression data under condition k (jth column of data matrix $\mathcal{D}$), with the first n_1 columns of $\mathcal{D}$ corresponding to condition 1 (control) and the remaining $n_2 = n - n_1$ columns to condition 2 (treatment).

Denote by $\Lambda^{(k)}$ the *influence matrix* and $\boldsymbol{\mu}^{(k)}$ the mean vector under condition k. The NetGSA framework considers a latent variable model of the form

$$\mathbf{Y}_j^{(1)} = \Lambda^{(1)}\boldsymbol{\mu}^{(1)} + \Lambda^{(1)}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, \quad (j = 1, \ldots, n_1),$$
$$\mathbf{Y}_j^{(2)} = \Lambda^{(2)}\boldsymbol{\mu}^{(2)} + \Lambda^{(2)}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, \quad (j = n_1 + 1, \ldots, n).$$

Here, $\boldsymbol{\gamma}_j$ is the vector of (unknown) random effects, and $\boldsymbol{\varepsilon}_j$ is the vector of random errors. They are independent and normally distributed with mean $\mathbf{0}$ and variances $\sigma_\gamma^2\mathbf{I}_p$ and $\sigma_\varepsilon^2\mathbf{I}_p$, respectively.

Inference in NetGSA requires estimation of the mean parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and variance components σ_γ^2 and σ_ε^2. The variance components can be estimated via maximum likelihood or restricted maximum likelihood, which can be computationally demanding for large networks. To extend the applicability of the NetGSA, we consider using Newton's method for estimating the variance parameters based on the profile log-likelihood to improve the computational stability. See Supplementary Materials Section B for more details.

2.2.2 Joint pathway enrichment and differential network analysis using NetGSA

To test for enrichment of a pre-specified pathway P, Shojaie and Michailidis (2009) propose the contrast vector (Searle, 1971) $\ell = (-\mathbf{b}\Lambda^{(1)} \odot \mathbf{b}, \mathbf{b}\Lambda^{(2)} \odot \mathbf{b})$, where $\mathbf{b}$ is a row binary vector determining the membership of genes in a pre-specified pathway P and $\odot$ denotes the Hadamard product. The advantage of this contrast vector is that it isolates influences from nodes outside the pathways of interest. Let $\boldsymbol{\beta} = (\boldsymbol{\mu}^{(1)T}, \boldsymbol{\mu}^{(2)\mathrm{T}})^{\mathrm{T}}$ be the concatenated vector of means. The null hypothesis of no pathway activity vs the alternative of pathway activation then becomes

$$H_0 : \ell\boldsymbol{\beta} = 0, \quad H_1 : \ell\boldsymbol{\beta} \neq 0. \tag{11}$$

The significance of individual contrast vectors in (11) can be tested using the following Wald test statistic

$$TS = \frac{\ell\widehat{\boldsymbol{\beta}}}{\mathrm{SE}(\ell\widehat{\boldsymbol{\beta}})}, \tag{12}$$

where $\mathrm{SE}(\ell\widehat{\boldsymbol{\beta}})$ represents the standard error of $\ell\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$. Both ℓ and $\mathrm{SE}(\ell\widehat{\boldsymbol{\beta}})$ depend on the underlying networks, which are estimated using data from the two experimental conditions. Under the null hypothesis, TS follows approximately a t-distribution whose degrees of freedom can be estimated using the Satterthwaite approximation method (Shojaie and Michailidis, 2010).

The above general framework allows for test of pathway enrichment in arbitrary subnetworks, while automatically adjusting for overlap among pathways. In addition, the above choice of contrast vector ℓ accommodates changes in the network structure. Such changes have been found to play a significant role in development and initiation of complex diseases (Chuang *et al.*, 2012), and NetGSA is currently the only method that systematically combines the changes in expression levels and network structures, when testing for pathway enrichment. However, the applicability of the existing NetGSA framework (Shojaie and Michailidis, 2009, 2010) is limited by the assumption of known network structure (namely $\Lambda^{(k)}, k = 1, 2$). In the current framework, we estimate $\Lambda^{(k)}$ $(k = 1, 2)$ from data as discussed in Section 2.1. We next show that NetGSA with estimated network information provides valid inference for pathway enrichment and differential network analysis.

For $k = 1$, 2, let $\mathbf{Z}^{(k)}$ of dimension $m_k \times p$ be the data matrix used to separately estimate the partial correlation matrix under condition k. Denote by S_k the number of nonzero off-diagonal entries in the true partial correlation matrix $\mathbf{A}_0^{(k)}$ and by r_k the percentage of available external information. We obtain the following result.

Theorem 2. Let assumptions in Theorem 1 be satisfied and $S_k = o(m_k/\log(p - r_k p))$ under each condition k $(k = 1, 2)$. Consider the inverse covariance matrices $\widehat{\boldsymbol{\Omega}}^{(k)}$ estimated from (5) of Section 2.1. Then the test statistic in (12) based on the corresponding networks $\widehat{\mathbf{A}}^{(k)}$ is an asymptotically most powerful unbiased test for (11).

Remark 4. Theorem 2.1 of Shojaie and Michailidis (2010) says that NetGSA is robust to uncertainty in network information. Specifically, Shojaie and Michailidis (2010) show that if the error in network information $\Delta_{\mathbf{A}_0^{(k)}} = \widehat{\mathbf{A}}^{(k)} - \mathbf{A}_0^{(k)}$ satisfies $||\Delta_{\mathbf{A}_0^{(k)}}||_2 = o_{\mathbb{P}}(1)$, then NetGSA is an asymptotically most powerful unbiased test for (11). The result in Theorem 2 establishes this property for (partially) estimated networks using the consistency of our proposed network estimation procedure in Theorem 1 and Corollary 1. A detailed proof can be found in the Supplementary Materials Section A.

3 Simulation results

We present two simulation studies to assess the performance of the proposed network estimation procedure, as well as its impact on NetGSA. We refer readers to the Supplementary Materials Section C for additional simulation scenarios—including validation of Type I errors and settings with a large number of variables p—and subsequent discussion.

Our first experiment is based on an undirected network of size $p = 100$. The network structure is extracted from the DREAM3 challenge (Prill *et al.*, 2010) corresponding to the Ecoli network (labeled Ecoli 1). The pathways of interest are determined through a community detection algorithm based on the leading nonnegative eigenvector of the modularity matrix of the network (Csardi and Nepusz, 2006). Under the null hypothesis, all nodes have the same mean expression values of 1. Under the alternative hypothesis, the mean expression levels of 0%, 30%, 40% and 60% of nodes in subnetworks 1, 3, 5 and 7 are increased by 0.5, respectively.

Our second experiment considers a network of size $p = 160$ with 8 subnetworks of equal sizes, all of which are generated from the same scale-free graph of size 20. To allow for interactions across subnetworks, there is 20% chance for the hub node in each subnetwork to connect to the hub node in another subnetwork. Mean expression values for all nodes are the same under the null hypothesis. Under the alternative hypothesis, we allow, respectively, 0, 40, 60 and 80% of the nodes to have positive mean changes of magnitude 0.5 for subnetworks 1–4. Subnetworks 5–8 follow the same pattern.

In both experiments, we also allow the structures in four subnetworks under the alternative hypothesis to differ from their null equivalent by a small amount, in order to simultaneously test pathway enrichment and differential networks. Figure 1 shows the network topologies as well as the structural changes for the chosen subnetworks from the null to the alternative hypothesis in the two experiments. Further, we study the robustness of NetGSA to model misspecification by including scenarios where a proportion (50% for $r = 0.2$ and 20% for $r = 0.8$ in experiment 1, and 60% in experiment 2) of the supplied structural information is incorrectly specified, i.e. they are not present in the true model.

To illustrate how external network information affects the estimation accuracy, we vary the percentage of information r from 0 to 1. When r is less than 1, we estimate the adjacency matrices using

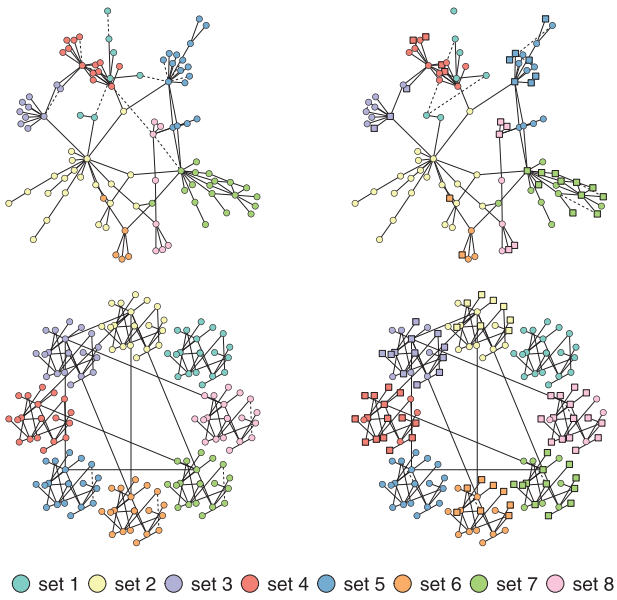

Fig. 1. The network and subnetwork topology in experiment 1 under the null (top left) and alternative (top right), and experiment 2 under the null (bottom left) and alternative (bottom right). Dashed lines represent edges that are present in only one condition. Nodes in square are associated with mean changes

the proposed two-step procedure and fill in the non-zero edges with the estimated weights. When full knowledge of the network topology is given ($r = 1$), we only apply the second step to estimate the edge weights. When there exist misspecified edges in the external information, we use two tuning parameters for network estimation, one for controlling the overall sparsity of the network and the other for correcting the misspecified edges. The optimal tuning parameters are selected over a grid of values using BIC defined in (9).

Table 1 compares the estimated networks with the true model under several deviance measures based on 100 simulation replications; in both experiments, the sample size for both null and alternative hypotheses is $m = 100$. The Matthews correlation coefficients improve significantly as the percentage of external information r increases from 20 to 80%, and the Frobenius norm loss shows a clear decreasing trend, both indicating the improvement in estimation accuracy when more external information is available. In cases where the information is misspecified [denoted by 0.2(m) and 0.8(m)], one can see that the performance of network estimation is not compromised by much after properly selecting the tuning parameters.

Next, we examine the performance of NetGSA in detecting pathway enrichment by comparing it with GSA (Efron and Tibshirani, 2007). GSA tests a competitive null hypothesis and compares the set of genes in the pathway with its complement in terms of association with the phenotype. The underlying model consists of both randomization of the genes and permutation of the samples, which are combined into the idea of 'restandardization'. This method is later denoted by GSA-c. In addition, we consider GSA with permutation of the samples only, later denoted by GSA-s, since this version of GSA compares the set of genes in the pathway with itself.

Tables 2 and 3 present the estimated powers for each subnetwork in the two experiments from 100 simulation replicates, respectively. Here we use $n_1 = n_2 = 25$ samples for each condition in experiment 1 and $n_1 = n_2 = 40$ in experiment 2, which are different from the datasets used for network estimation. The powers are calculated as the proportion of replicates that show differential

Table 1. False positive rate (FPR in percentage), false negative rate (FNR in percentage), Matthews correlation coefficient (MCC) and Frobenius norm loss (Fnorm) for network estimation in experiments 1 and 2

		$P = 100$				$P = 160$			
	r	FPR (%)	FNR (%)	MCC	Fnorm	FPR (%)	FNR (%)	MCC	Fnorm
Null	0.0	9.46	2.78	0.43	0.48	2.94	0.84	0.54	0.36
	0.2	7.64	5.83	0.45	0.46	2.77	1.03	0.55	0.34
	0.8	1.81	1.22	0.75	0.28	1.18	0.02	0.72	0.24
	0.2 (m)	7.91	4.85	0.45	0.46	2.76	0.95	0.55	0.34
	0.8 (m)	2.29	3.82	0.70	0.31	1.22	0.02	0.71	0.25
Alt	0.0	8.71	1.52	0.44	0.45	2.90	0.88	0.54	0.36
	0.2	7.09	3.82	0.47	0.42	2.73	0.88	0.55	0.35
	0.8	1.80	1.19	0.75	0.25	1.19	1.89	0.71	0.26
	0.2 (m)	7.29	2.62	0.47	0.42	2.72	0.78	0.55	0.34
	0.8 (m)	2.17	5.50	0.69	0.29	1.22	1.93	0.70	0.27

Table 2. Powers in experiment 1

	$P = 100$							
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2 (m)	0.8 (m)
1	0.03	0.03	0.08	0.06	0.15	0.04	0.03	0.02
2	0.08	0.08	0.08	0.06	0.09	0.00	0.09	0.09
3	0.36	0.33	0.43	0.46	0.24	0.00	0.40	0.38
4	0.38	0.26	0.09	0.07	0.26	0.05	0.37	0.24
5	0.91	0.91	0.95	0.97	0.95	0.00	0.92	0.89
6	0.27	0.24	0.24	0.26	0.37	0.00	0.26	0.25
7	0.72	0.80	0.99	0.99	0.98	0.14	0.69	0.86
8	0.45	0.61	0.63	0.57	0.87	0.00	0.51	0.58

FDR cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c. 0.2/0.8 refer to NetGSA with 20/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes based on 1000 permutations; 0.2/0.8(m) refer to NetGSA with 20/80% misspecified external information.

changes, based on the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg (1995). To facilitate comparison, different FDR cutoffs are used for GSA and NetGSA to ensure consistent type I error for the first pathway in both experiments. For NetGSA, we look at scenarios when there is 20 and 80% external structural information (with and without misspecification) and use the estimated networks to test enrichment for each subnetwork. We also include the scenario where the exact networks with correct edge weights are provided, in which case only the variance components and mean expression values are estimated from the mixed linear model. True powers for each subnetwork are calculated by replacing all unknown parameters with their corresponding known values.

For $p = 100$, the results from NetGSA with the exact networks agree with the true powers, indicating low powers for subnetworks 1, 2 and 4, slightly higher powers for 3, 6 and 8, high powers for 5 and 7 due to significant changes in mean expression levels and structures. When the exact networks are unknown, we see clear improvement in the estimated powers for subnetworks 4, 7 and 8 as the percentage of external information increases from 20 to 80%. GSA-s does reasonably well with overestimated powers for subnetwork 8. The last two columns in Table 2 show the estimated powers from NetGSA when the external information is misspecified. For both

Table 3. Powers in experiment 2

	$P = 160$							
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.04	0.06	0.02	0.05	0.06	0.02	0.04	0.06
2	0.37	0.36	0.30	0.36	0.41	0.00	0.36	0.36
3	0.88	0.94	0.96	0.99	0.99	0.00	0.89	0.94
4	0.97	0.99	1.00	1.00	1.00	0.23	0.97	0.99
5	0.36	0.25	0.11	0.11	0.13	0.15	0.35	0.25
6	0.38	0.27	0.03	0.07	0.26	0.01	0.35	0.27
7	0.66	0.72	0.92	0.92	1.00	0.00	0.67	0.72
8	0.90	0.95	1.00	1.00	1.00	0.13	0.91	0.94

FDR cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes based on 1000 permutations; 0.2/0.8(m) refer to NetGSA with 20/80% misspecified external information.

cases ($r = 0.2$ and $r = 0.8$), the results bear high similarity to those in the first two columns, which suggests that the proposed framework is robust to inaccuracy in network information.

For $p = 160$, the NetGSA estimated powers when 20% external information is available match the true powers reasonably well, with a small underestimation of powers for subnetworks 3, 7 and 8. We note marked improvement in the three corresponding values when the external information increases to 80%. Moreover, NetGSA is able to distinguish subnetworks 5–8 that have both changes in mean values and subnetwork topology from their corresponding counterparts 1–4. When the external information is misspecified, the last two columns indicate that NetGSA still returns valid powers that are comparable to those obtained with correctly specified structural information. GSA-s yields a small overestimation of powers for subnetwork 7.

In both experiments, GSA with randomization of the genes (GSA-c) fails to identify any of the differential subnetworks.

4 Applications to metabolomics and genomics data

We apply NetGSA to three Omics data sets to demonstrate its potential in revealing biological insights. In all three studies, the *P*-values were corrected for multiple comparisons using the FDR control procedure proposed in Benjamini and Yekutieli (2001) to account for the dependency among KEGG pathways.

Our first application is based on the metabolomics data set from (Putluri *et al.*, 2011) to examine changes in metabolic profiles associated with bladder cancer using untargeted mass spectrometry data acquisition strategy. The data consists of 31 cancer and 27 benign tissue samples and 63 detected metabolites. Here we focused on estimating the network of metabolic interactions, enhanced by information gleaned from KEGG (Kanehisa and Goto, 2000). For each condition, we used the BIC criterion to select the tuning parameter λ. At the optimal λ, we applied the proposed network estimation procedure to identify the metabolic network; see Figure 2 for an illustration of the estimated networks for the cancer and benign classes, respectively. It can be seen that there are numerous interactions between pathways that describe energy metabolism in the cancer state, due to the greater need of cancer cells for energy.

We tested for differential activity of biochemical pathways extracted from KEGG using the same set of data. Shown in Table 4

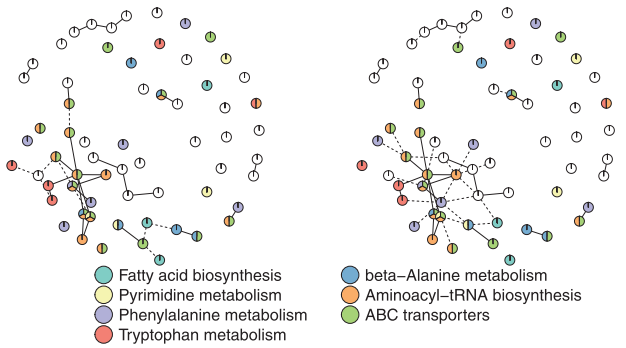

Fig. 2. The estimated network topology and enriched pathways in the metabolomics study for the benign class (left) and cancer class (right). Dashed lines represent edges that are present in only one class. Nodes in multiple colors are present in multiple pathways

Table 4. *P*-values for the pathways in the metabolomics study, with FDR correction at $q^* = 0.01$

Pathway	NetGSA	GSA-s	GSA-c
Tryptophan metabolism	$3e^{-5}$	0.00	1.00
beta-Alanine metabolism	$3e^{-5}$	0.00	1.00
Aminoacyl-tRNA biosynthesis	$2e^{-4}$	0.00	1.00
ABC transporters	$4e^{-4}$	0.00	1.00
Fatty acid biosynthesis	$2e^{-3}$	1.00	1.00
Pyrimidine metabolism	$2e^{-3}$	0.00	1.00
Phenylalanine metabolism	$4e^{-3}$	0.00	1.00

Here 0.00 represents a zero *P*-value produced out of finite permutations. GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes based on 3000 permutations.

are estimated *P*-values after FDR correction with a *q*-value of 0.01 for the significant pathways selected from NetGSA. These identified pathways include those that describe altered utilization of amino acids and their aromatic counterparts, as well as metabolism of fatty acids and intermediates of tricarboxylic acid cycle which were followed up for biological insights in the original study by Putluri *et al.* (2011). Among the selected pathways, fatty acid biosynthesis is not identified by GSA-s. Interestingly, GSA-c fails to report any pathway as being significantly enriched. This again confirms our hypothesis that incorporating pathway topology information allows sophisticated enrichment methods in detecting important regulatory pathways.

The second data set (Subramanian *et al.*, 2005) consists of gene expression profiles of 5217 genes for 62 normal and 24 lung cancer patients. We considered 47 KEGG pathways of size at least 5 that describe signaling and biochemical mechanisms and excluded genes that eitrher are not present in the 47 pathways, or without recorded network information. The number of genes that remain for pathway enrichment analysis is 303. Based on the external topology information from the BioGRID database, we applied the proposed network estimation procedure coupled with BIC to estimate the underlying interaction networks for both normal and cancer conditions. We then explored whether GSA and NetGSA with the estimated networks are able to detect enriched pathways using the same data set. After correcting for multiple comparisons, using a FDR of $q^* = 0.01$, none of the three methods identifies any pathway as being significantly differential enriched. The lack of statistical power in obtaining differential pathways was also noted in the original article of (Subramanian *et al.*, 2005); see Supplementary Table S9 for

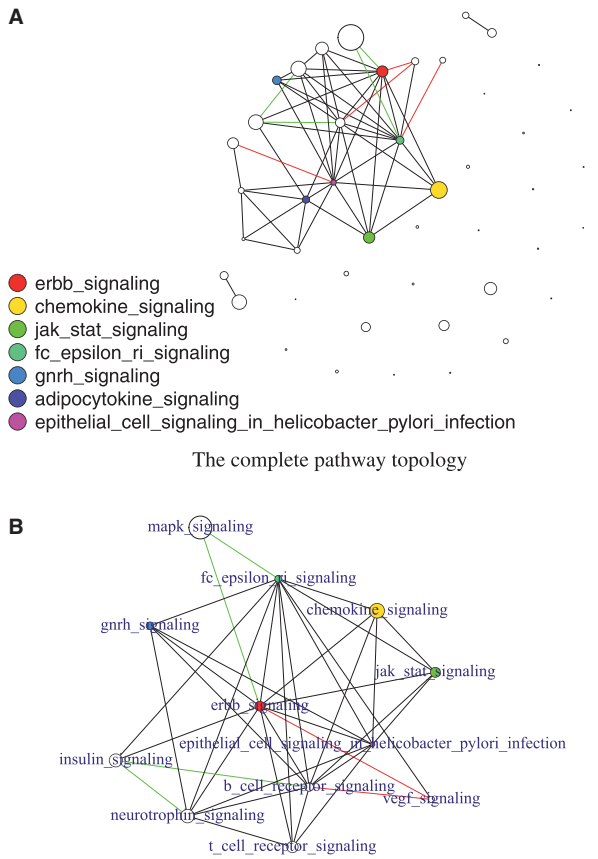

Fig. 3. The estimated pathway topology in the TCGA cancer study: (A) the complete pathway topology and (B) the subnetwork surrounding the ErbB pathway. Edges in black are present in both classes, whereas red and green edges are only present in ER positive and ER negative class, respectively. Node size is proportional to the size (number of genes) of the corresponding pathway.

Table 5. *P*-values for the differential pathways in the TCGA data, with FDR correction at $q^* = 0.01$

Pathway	NetGSA	GSA-s	GSA-c
Epithelial cell signaling in *Helicobacter pylori* infection	$5e^{-95}$	0.00	1.00
Cell cycle	$2e^{-47}$	0.00	1.00
Galactose metabolism	$3e^{-31}$	0.00	1.00
Glutathione metabolism	$1e^{-27}$	0.00	1.00
NOD-like receptor signaling pathway	$1e^{-24}$	0.00	1.00
Pyrimidine metabolism	$4e^{-23}$	0.00	1.00
Cysteine and methionine metabolism	$1e^{-22}$	0.00	1.00
Starch and sucrose metabolism	$1e^{-18}$	0.00	1.00
Toll-like receptor signaling pathway	$1e^{-18}$	0.00	1.00
Glycolysis/Gluconeogenesis	$3e^{-17}$	0.00	1.00
Jak-STAT signaling pathway	$9e^{-15}$	0.00	1.00
Chemokine signaling pathway	$3e^{-14}$	0.00	1.00
ErbB signaling pathway	$7e^{-13}$	0.00	1.00
p53 signaling pathway	$7e^{-12}$	0.00	1.00
Hedgehog signaling pathway	$5e^{-10}$	0.00	1.00
beta-Alanine metabolism	$1e^{-7}$	0.00	1.00
Fc epsilon RI signaling pathway	$5e^{-7}$	0.00	1.00
Fructose and mannose metabolism	$2e^{-6}$	0.00	1.00
Pentose phosphate pathway	$2e^{-6}$	0.00	1.00
PPAR signaling pathway	$5e^{-6}$	0.00	1.00
Adipocytokine signaling pathway	$4e^{-5}$	0.00	1.00
Purine metabolism	$6e^{-5}$	0.00	1.00
Valine, leucine and isoleucine degradation	$5e^{-4}$	$1e^{-3}$	1.00
GnRH signaling pathway	$2e^{-3}$	0.00	1.00
TGF-beta signaling pathway	$3e^{-3}$	0.00	1.00

Here 0.00 represents a zero *p*-value produced out of finite permutations. GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes based on 3000 permutations.

the complete list of FDR adjusted *P*-values. Nevertheless, using NetGSA and a relaxed FDR cutoff threshold of 0.30 (similar to the strategy adopted in Subramanian *et al.*, 2005), we obtain the following top three ranked signaling pathways: Jak-STAT, p53 and Wnt. All three are implicated in lung cancer, although the latter two are also implicated in multiple other types of human malignancies. However, the Jak/STAT pathway has been recently shown to play a key role in non small cell lung cancer cells (Song *et al.*, 2011).

Our third and final application is based on a data set from TCGA (2012). The data set contains RNA-seq measurements for 17 296 genes from 1033 breast cancer specimens, including ER positive, ER negative and other unevaluated cases. As in the previous gene microarray study, the external network information is extracted from the BioGRID database. We focused on a subset of the genes that have recorded network information and are present in KEGG pathways with at least 5 members. This leaves for further consideration 800 genes with 403 samples from the ER positive and 117 from the ER negative classes, spanning over 45 KEGG pathways. We then applied the constrained network estimation procedure with the tuning parameter selected via BIC in (9) to obtain the partial correlation networks for the ER positive and ER negative classes, respectively. Due to the large number of variables, visualization of the estimated networks at the individual gene level is challenging. Instead, we examine the interactions among pathways in Figure 3 to gain insight into their co-regulation behavior. The weighted pathway level network is defined as follows. Let each node in the network represent one pathway, with size proportional to the size of the corresponding pathway. A weighted edge between two pathways P_1 and P_2 is defined as the number of nonzero partial correlations between genes in P_1 and those in P_2 (normalized by the sizes of the two pathways). Links visualized in Figure 3 are the top 5% of the weighted edges, where ranking is based on edge weights.

Table 5 presents the FDR corrected *P*-values for the selected differential pathways using NetGSA based on the estimated partial correlation networks, as well as GSA-c and GSA-s. The complete table is presented in the Supplementary Materials Section D. At $q^* = 0.01$, NetGSA reports 25 out of the 45 KEGG pathways as significantly enriched, whereas GSA either rejects the null for all pathways (GSA-c) or fails to reject any pathway (GSA-s). Selected differential pathways identified by NetGSA are also highlighted in Figure 3. Of particular interest is the set of connected, enriched pathways centered around the ErbB pathway in Figure 3(b). This pathway contains receptors that signal through various pathways to regulate cell proliferation, migration, differentiation, apoptosis, and cell motility and play a key role in breast cancer (Howe and Brown, 2011), although its role in breast carcinogenesis not very well understood. Note that the Jak-STAT pathway is downstream of the ErbB one and can be activated by key epidermal growth factor receptors in the former to create signaling cascades (Henson and Gibson, 2006). Further, the GnRH signaling pathway has been reported to interact with the ErbB pathway receptors (Morgan *et al.*, 2011). All these

interconnected pathways are related to receptors that have been implicated in various studies with over-expression in the ER negative class and hence faster tumor growth and poorer clinical outcomes.

5 Discussion

This article introduces a constrained partial correlation network estimation method that seamlessly incorporates externally available interaction information for genes and other biomolecules. The end product is a reliable condition-specific estimate of the underlying networks. The resulting estimated network structures are then used for network-based pathway enrichment analysis. For the purpose of constrained network estimation, one might also try the one-step constrained maximum likelihood estimation (a functionality offered in the R-package glasso) to recover the underlying partial correlation network. However, this one-step approach requires sophisticated specification of the tuning parameters at positions for which structural information is available, and can be challenging to implement in practice.

Two sources of uncertainty can be identified in the proposed framework: one from the reliability of the external database information in the network estimation procedure and the other from the uncertainty regarding the estimated network itself, as well as how it propagates into the NetGSA testing procedure. As discussed in Remark 1, the proposed method can conveniently accommodate the first source of uncertainty by incorporating a non-zero penalty on parameters that are uncertain. Further, as shown in Theorem 2, the proposed test via the extended NetGSA framework is asymptotically unbiased and most powerful, given the consistency of the estimated network, and hence accounts for the second source of uncertainty. Nevertheless, in finite samples as the numerical work in the Supplementary Materials Section C illustrates, Type I errors may be slightly off in the presence of numerous errors in the estimated network (either due to misspecification of the external information or lack of samples for accurate estimation). The topic of dealing with network estimation errors and possible ways to address it is discussed in Narayan and Allen (2016).

Finally, the current framework of NetGSA uses the Cholesky decomposition of the covariance matrix of the underlying network. It is natural to ask whether the order of the variables affects the result of enrichment analysis. In simulations and the real data analyses, we find that the estimated powers/*P*-values from NetGSA are comparable after permutation of the variables.

Acknowledgements

The authors would like to thank the reviewers for many constructive comments and suggestions.

Funding

Ali Shojaie was supported by NSF award (DMS-1161565) and National Institutes of Health award (1K01HL124050-01A1). George Michailidis was supported by NSF awards (DMS-1228164 and DMS-1545277) and National Institutes of Health award (7R21GM10171903).

Conflict of Interest: none declared.

References

Al-Shahrour,F. *et al.* (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B*, **57**, 289–300.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Bickel,P.J. *et al.* (2009) Simultaneous analysis of lasso and dantzig selector. *Ann. Stat.*, **37**, 1705–1732.

Boyd,S. and Vandenberghe,L. (2004). *Convex Optimization*. Cambridge University Press, Cambrridge.

Candes,E.J. and Recht,B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

Chuang,H.Y. *et al.* (2012) Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood*, **120**, 2639–2649.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal. Compl. Syst*, 1695. http://igraph.org.

Dehmer,M. and Emmert-Streib,F. (2008). *Analysis of Microarray Data: A Network-Based Approach*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Dempster,A.P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Gottwein,E. *et al.* (2007) A viral microrna functions as an orthologue of cellular mir-155. *Nature*, **450**, 1096–1099.

Green,M.R. *et al.* (2011) Signatures of murine b-cell development implicate yy1 as a regulator of the germinal center-specific program. *Proc. Natl. Acad. Sci. USA*, **108**, 2873–2878.

Henson,E.S. and Gibson,S.B. (2006) Surviving cell death through epidermal growth factor (egf) signal transduction pathways: implications for cancer therapy. *Cell. Signal.*, **18**, 2089–2097.

Houstis,N. *et al.* (2006) Reactive oxygen species have a causal role in multiple forms of insulin resistance. *Nature*, **440**, 944–948.

Howe,L.R. and Brown,P.H. (2011) Targeting the her/egfr/erbb family to prevent breast cancer. *Cancer Prevent. Res.*, **4**, 1149–1157.

Huang,D.W. *et al.* (2008) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Huerta,A.M. *et al.* (1998) Regulondb: a database on transcriptional regulation in escherichia coli. *Nucleic Acids Res.*, **26**, 55–59.

Ideker,T. and Krogan,N.J. (2012) Differential network biology. *Mol. Syst. Biol.*, **8**, 565.

Ideker,T. *et al.* (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, **144**, 860–863.

Joshi-Tope,G. *et al.* (2003) The genome knowledgebase: A resource for biologists and bioinformaticists. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 237–244.

Kanehisa,M. and Goto,S. (2000) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Lauritzen,S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Meinshausen,N. and Bühlmann,P. (2006) High dimensional graphs and variable selection with the lasso. *Ann. Stat.*, **34**, 1436–1462.

Morgan,K. *et al.* (2011) Gnrh receptor activation competes at a low level with growth signaling in stably transfected human breast cell lines. *BMC Cancer*, **11**, 476.

Narayan,M. and Allen,G.I. (2016) Mixed effects models to find differences in multi-subject functional connectivity. *Front. Neurosci.*, **10**,

Nishimura,D. (2001) Biocarta. *Biotech Softw. Internet Rep.*, **2**, 117–120.

Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS One*, **5**, e9202.

Putluri,N. *et al*. (2011) Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression. *Cancer Res*., **71**, 7376–7386.

Rothman,A.J. *et al*. (2008) Sparse permutation invariant covariance estimation. *Electron. J. Stat*., **2**, 494–515.

Searle,S. (1971). *Linear Models*. John Wiley & Sons.

Shojaie,A. and Michailidis,G. (2009) Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol*., **16**, 407–426.

Shojaie,A. and Michailidis,G. (2010) Network enrichment analysis in complex experiments. *Stat. Appl. Genet. Mol. Biol*., **9**, 22.

Song,L. *et al*. (2011) Jak1 activates stat3 activity in non-small–cell lung cancer cells and il-6 neutralizing antibodies can suppress jak1-stat3 signaling. *Mol. Cancer Ther*., **10**, 481–494.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

TCGA (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Wermuth,N. (1980) Linear recursive equations, covariance selection, and path analysis. *J. Am. Stat. Assoc*., **75**, 963–972.

Wilson,B.G. *et al*. (2010) Epigenetic antagonism between polycomb and swi/snf complexes during oncogenic transformation. *Cancer Cell*, **18**, 316–328.

Zaki,N. *et al*. (2013) Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinformatics*, **14**, 163.

Zhou,S. *et al*. (2011) High-dimensional covariance estimation based on gaussian graphical models. *J. Mach. Learn. Res*., **12**, 2975–3026.