*Statistical Applications in Genetics and Molecular Biology*

*Volume* 9, *Issue* 1 2010 *Article* 22

# Network Enrichment Analysis in Complex Experiments

Ali Shojaie[*] George Michailidis[†]

[*]University of Michigan - Ann Arbor, shojaie@umich.edu

[†]University of Michigan - Ann Arbor, gmichail@umich.edu

# Network Enrichment Analysis in Complex Experiments*

Ali Shojaie and George Michailidis

## Abstract

Cellular functions of living organisms are carried out through complex systems of interacting components. Including such interactions in the analysis, and considering sub-systems defined by biological pathways instead of individual components (e.g. genes), can lead to new findings about complex biological mechanisms. Networks are often used to capture such interactions and can be incorporated in models to improve the efficiency in estimation and inference. In this paper, we propose a model for incorporating external information about interactions among genes (proteins/metabolites) in differential analysis of gene sets. We exploit the framework of mixed linear models and propose a flexible inference procedure for analysis of changes in biological pathways. The proposed method facilitates the analysis of complex experiments, including multiple experimental conditions and temporal correlations among observations. We propose an efficient iterative algorithm for estimation of the model parameters and show that the proposed framework is asymptotically robust to the presence of noise in the network information. The performance of the proposed model is illustrated through the analysis of gene expression data for environmental stress response (ESR) in yeast, as well as simulated data sets.

**KEYWORDS:** gene network, enrichment analysis, gene set analysis, complex experiments, spatio-temporal model, mixed linear model, systems biology

*The work of George Michailidis was partially supported by NIH grant 1RC1CA145444-0110 and MEDC grant GR-687 and Ali Shojaie was partially supported by NIH grant T32 CA083654. We would like to thank the editor and the two anonymous referees for helpful comments and suggestions.

# 1 Introduction

Recent advances in high throughput technologies have facilitated the simultaneous study of components of complex biological systems. Microarray technologies provide information about the expression levels of virtually all genes in the genome of a given specie; the patterns of changes in these expressions over large groups of genes can determine how living organisms respond to their environment. However, genes interact with each other in an orchestrated fashion and analysis of individual genes without taking into account their interactions (*single gene analysis*) may result in reduced efficiency and bias. We provide next an overview of two classes of methods that aim to overcome this shortcoming and discuss advantages and disadvantages of each of the methods.

## 1.1 Background

Two classes of models have been investigated by researchers in order to account for interactions among components of biological systems in the differential analysis of genes (proteins/metabolites). The first approach, known as *gene set analysis*, is to consider the joint effect of biologically related groups of genes. By performing gene set analysis, one hopes that the interactions among genes are preserved through considering the combined effect of genes in each set, and hence the resulting inference procedure implicitly includes such interactions. In addition, while individual genes may not show important changes of expression, the combined effect of changes in expressions of genes in a set (e.g. a genetic pathway) could unveil important changes in the state of the system. Hence, methods of gene set analysis offer improvements in both power, as well as interpretability of inference procedures. Examples of methods for gene set analysis include the Gene Set Enrichment Analysis (GSEA) of Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander et al. (2005) and its variants (Tian, Greenberg, Kong, Altschuler, Kohane, and Park, 2005, Efron and Tibshirani, 2007), which use a permutation-based procedure in order to evaluate the significance of gene sets.

The second class of methods aims to directly incorporate available information about interactions among genes and proteins into differential analysis. Gene networks are efficient tools to represent and model interactions among genes (Rahnenführer, Domingues, Maydt, and Lengauer, 2004) and have been used to improve the performance of differential analysis methods. Ideker, Thorsson, Ranish, Christmas, Buhler, Eng, Bumgarner, Goodlett, Aebersold, and Hood (2001) used integrated genomic and proteomic analysis of perturbed networks to discover interactions among genes. This was followed by proposing a method to test the significance of subnetworks through a permutation-based method (Ideker, Ozier, Schwikowski, and Siegel, 2002). Recently, Wei and Li (2007) and Wei and Pan (2008) have proposed Markov random field models to incorporate the network information in the differential analysis of genes. In these methods, connected genes in the networks are assumed to have "similar" expression levels and a Bayesian framework is developed using mixture models to evaluate whether each gene is differentially expressed.

A number of methods have recently been developed to combine the advantages of incorporating network information with strengths of enrichment analysis. Sanguinetti, Noirel, and Wright (2008) considered a mixture model on graphs (MMG) to account for network information in proteomic data and used a simple percolation algorithm to search for subnetworks of significant components. Shojaie and Michailidis (2009) discussed a method that incorporates network information through a latent variable model and used the framework of mixed linear models (MLM) to test whether *a priori* defined gene sets are differentially expressed. They considered two special classes of networks, namely directed acyclic graphs (DAGs), as well as sub-stochastic graphs and proposed a test statistic for the two-class inference problem (e.g. treatment and control).

The above models can all be viewed as attempts to incorporate the spatial correlation caused by the gene network into the analysis of differentially expressed genes. Another important aspect of gene expression is the dynamic behavior of genes in response to environmental conditions (Gasch, Spellman, Kao, Carmel-Harel, Eisen, Storz, Botstein, and Brown, 2000, Gasch and Werner-Washburne, 2002). The changes in gene expression levels over time may reveal unique features of biological systems that are not evident from studying gene expressions at a single time point. The temporal correlation among gene expressions can also be utilized to improve the efficiency of finding differentially expressed genes. Examples of models for time course gene expression data include Hong and Li (2006) and Yoneya and Mamitsuka (2007). *Spatio-temporal* models for gene expression analysis combine the advantages of both models. Wei and Li (2008) recently proposed a hidden spatio-temporal Markov random field model to account for both temporal correlation among expression levels, as well as spatial correlation among genes represented by the gene networks.

## 1.2 Outline

Currently available methods, reviewed above, focus either on incorporating network information for performing single gene analysis, or on gene set enrichment analysis for simple experimental conditions; e.g. treatment and control. Since methods of enrichment analysis are based on permutations tests (e.g. Ideker et al., 2002, Subramanian et al., 2005, Tian et al., 2005, Efron and Tibshirani, 2007), their extension to complex experimental settings, including the presence of temporal correlation among observations, is not straightforward.

In this paper, we generalize the framework of Shojaie and Michailidis (2009) in order to develop a flexible framework for analysis of gene sets in complex experimental conditions, while incorporating the known network information. In particular, we

(a) propose a generalization of the network influence to analyze arbitrary networks with both directed, and undirected edges,

(b) exploit the flexibility of mixed linear models to develop a general inference procedure that can be used to analyze changes in biological pathways in complex experiments, including experiments with multiple factors together with time course data, and

(c) describe an inference framework for simultaneous tests of multiple hypotheses for analysis of pathways in complex experiments.

In addition, in order to estimate the parameters of the resulting mixed linear model in real-world applications, we propose an iterative algorithm based on the block-relaxation technique (de Leeuw, 1994). Finally, we study the effect of noise in the underlying network information, e.g. when interactions among genes or the associated weights are estimated, and establish conditions under which the proposed inference procedure is asymptotically insensitive to such noise. Through analysis of simulated, as well as real, data examples, we illustrate the small sample properties of the proposed inference procedure and show that the model performs well in the presence of limited samples (the application discussed in Section 4 has a single sample per experimental condition and time point, and includes 3 time points) and also exhibits good performance in the analysis of small gene sets.

The remainder of the paper is organized as follows: in Section 2, the modeling framework is introduced and the mixed linear model representation is presented. The material in Sections 2.1 and 2.2 generalize the framework of Shojaie and Michailidis (2009) to analysis of general networks in complex experiments. Estimation and inference issues are discussed in Sections 2.3 and 2.4, respectively, and the asymptotic analysis of performance under noisy network information is presented in Section 2.5. The performance of the model is evaluated through simulation studies in Section 3. In particular, it is shown that while the performance of enrichment methods deteriorates in presence of temporal correlation, the proposed model can effectively handle the additional correlation. Finally, in Section 4, data from yeast environmental stress response (ESR) experiment of Gasch et al. (2000) are used to discover pathways that are differentially expressed in response to these stress factors. Section 5 summarizes the main findings and discusses some future research directions.

# 2 Model and Methods

Consider $p$ genes (proteins/metabolites) whose expression data $\mathscr{D}$ is organized in a $p \times n$ matrix, where each column of $\mathscr{D}$ represents a realization of the expression levels of genes in the study. In general, assume that there are $K$ different experimental conditions and each of conditions are studied for $J_k$ time points. Further, assume that for each combination of experimental condition and time, there exists $n_{jk}$ samples. Let $n = \sum_{k=1}^{K} \sum_{j=1}^{J_k} n_{jk}$ and denote by $Y$ an arbitrary column of the expression matrix $\mathscr{D}$. In other words, $Y$ consists of the expression levels of genes in the study for a given time point of a specific experimental condition.

## 2.1 The Latent Variable Model

In order to incorporate the network structure into the model, we represent the gene network by a directed graph $G = (V,E)$ with vertex set $V$, and edge set $E$. The edge set is captured in the $p \times p$ weighted adjacency matrix of the graph $A$, with positive and negative entries. Each nonzero element in the adjacency matrix, $A_{ij}$, represents a directed edge whose weight corresponds to the strength of association between the two vertices $i$ and $j$. Undirected graphs correspond to a special case, where $A_{ij} = A_{ji}$.

For any column of the gene expression matrix, suppose $Y = X + \varepsilon$, where $X$ represents the *signal* and $\varepsilon$ the measurement *noise*. It is assumed that the underlying expression level of each gene $X_i$ is a combination of its own individual effect and the influence of other genes. To that end, we define latent variables $\gamma_i$ that capture the individual gene contributions and assume that the signal for $i$ consists of $\gamma_i$ and the weighted sum of the expression levels of genes in the network that influence $i$. Finally, it is assumed that $\gamma$ and $\varepsilon$ are independent and normally distributed; specifically, $\gamma \sim N_p(\beta, \sigma_\gamma^2 I_p)$ and $\varepsilon \sim N_p(0,R)$, where $I_p$ denotes the $p$-identity matrix and $R$ is the covariance matrix of noise, which can account for other sources of dependence in the data, e.g. temporal correlation.

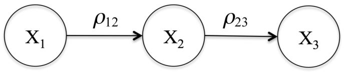

Figure 1: Illustration of the latent variable model for gene networks

To illustrate this model, consider the simple graph of Figure 1, for which we can write:

$$\begin{aligned} X_1 &= \gamma_1, \\ X_2 &= \rho_{12}X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2, \\ X_3 &= \rho_{23}X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3. \end{aligned}$$

More generally, these equations can be summarized in vector notation as:

$$Y = \Lambda\gamma + \varepsilon, \qquad \gamma \sim N_p(\beta, \sigma_\gamma^2 I_p), \qquad \varepsilon \sim N_p(0, R), \tag{2.1}$$

where $\Lambda$ is called the *Influence Matrix* of the graph and in the simple example of Figure 1 is given by:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}.$$

For the above model, we have $\mathbb{E}(Y) = \Lambda\beta$ and $\mathrm{Var}(Y) = \sigma_\gamma^2 \Lambda\Lambda' + R$, where $\Lambda'$ is the transpose of matrix $\Lambda$.

In most applications, the available network information comes in the form of the adjacency matrix, which determines the association between each gene and its immediate neighbors in the graph. On the other hand, the influence matrix represents the effect of each gene on all the other genes in the network and is given by $\Lambda = \sum_{r=0}^{\infty} A^r$, where $A^0 = I_p$. In the case of a directed acyclic graph (as in the toy example of Figure 1), Shojaie and Michailidis (2009) show that the relationship between $\Lambda$ and $A$ is given by $\Lambda = (I - A)^{-1}$. They also show that if the adjacency matrix of the network is sub-stochastic, i.e. its eigenvalues are smaller than 1 in magnitude, the above relationship between $A$ and $\Lambda$ still holds. This approach can also be adapted to define a latent variable model for chain graphs, where the network consists of undirected subgraphs that are connected by directed edges having no directed cycles (see e.g. Lauritzen, 1996). However, general gene networks, with both directed and undirected edges, may not satisfy the requirements of any of the above special classes of graphs. Therefore, an alternative approach is required to define the influence of the network for general graphs.

We start by normalizing the adjacency matrix $A$, by dividing its entries $A_{ij}$ by the corresponding row sum of the absolute values $\sum_{j=1}^{p} |A_{ij}|$. Formally, let

$$\mathscr{L}(\zeta)_{ij} = \frac{A_{ij}}{(\sum_{j=1}^{p} |A_{ij}|) + \zeta}, \qquad \text{for some } \zeta > 0. \tag{2.2}$$

Then by Gershgorin's Disk Theorem (see e.g. Friedberg, Insel, and Spence, 1996), the matrix $\mathscr{L}$ is sub-stochastic, and therefore, using the results of Shojaie and

Michailidis (2009), for each $\zeta > 0$ we get $\Lambda = (I - \mathscr{L}(\zeta))^{-1}$. Taking the limit, we get

$$\Lambda = \lim_{\zeta \to 0} (I - \mathscr{L}(\zeta))^{-1}.$$

This implies that, for general networks, the influence matrix of the graph can be defined as $\Lambda = (I - \mathscr{L})^{+}$, where $\mathscr{L}$ represents the normalized adjacency matrix with $\zeta = 0$ and $(I - \mathscr{L})^{+}$ denotes the Moore-Penrose pseudo-inverse of $I - \mathscr{L}$.

The normalization in (2.2) is motivated by the definition of the covariance matrix in Markov random fields (see e.g. Rue and Held, 2005). For undirected graphs with only positive weights in $A$, the matrix $I - \mathscr{L}$ also corresponds to a version of the Laplacian matrix of the graph (see e.g. Chung, 1997). Simulation studies show that small values of $\zeta$ (e.g. $\zeta \approx 0.01$) do not affect the outcome of the analysis, and $\Lambda = (I - \mathscr{L}(\zeta))^{-1}$ can be used to define the influence matrix.

## 2.2 Mixed Linear Model Representation

Consider the gene expression matrix of the previous section with $K$ experimental conditions, $J_k, k = 1 \cdots K$ time points and $n_{jk}$ observations at each combination of condition and time point. Let $\mathbf{Y}$, $\gamma$ and $\varepsilon$ represent the rearrangement of vectors $Y$, $\gamma$ and $\varepsilon$ into $np \times 1$ column vectors. Then, using the framework of mixed linear models, we can write

$$\begin{array}{rcll} \mathbf{Y} & = & \Psi\beta + \Pi\gamma + \varepsilon & \\ \varepsilon & \sim & N_{np}(\mathbf{0}, \mathbf{R}), & \mathbf{R} = \operatorname{diag}(R) \\ \gamma & \sim & N_{np}(\mathbf{0}, \mathbf{G}), & \mathbf{G} = \sigma_\gamma^2 \mathbf{I_{np}} \end{array} \tag{2.3}$$

Here $\beta$ and $\gamma$ are fixed and random effect parameters, and $\Psi$ and $\Pi$ are the corresponding design matrices of dimensions $np \times Kp$ and $np \times np$, respectively.

The precise form of these matrices depends on whether the influence matrix $\Lambda$ can change over time or over different experimental conditions (see Harbison, Gordon, Lee, Rinaldi, Macisaac, Danford, Hannett, Tagne, Reynolds, Yoo et al. (2004) for examples of changes in regulatory networks in different experimental conditions). To account for such changes in interactions among genes, let $\Lambda^{(jk)}$ denote the influence matrix of the network in condition $k$ at time point $j$. The changes in network influence over time or in different experimental conditions can then be directly incorporated in the model by replacing the matrix $\Lambda$ by the corresponding matrix $\Lambda^{(jk)}$ in design matrices $\Psi$ and $\Pi$, for fixed and random effect components in the mixed linear model of equation (2.3). Using this notation, $\Pi$ is a block diagonal matrix with $\Lambda^{(jk)}$ on the diagonal, while $\Psi$ is defined based on the setting of the experiment. More specifically, suppose $\chi$ is the design matrix of the linear regression model for a single gene, corresponding to $K$ experimental conditions and $J$ time points. The matrix $\Psi$ is then defined by replacing each $\chi_{jk}$ with $\chi_{jk}\Lambda^{(jk)}$. In the special case of $\Lambda^{(jk)} = \Lambda$, the matrices $\Psi$ and $\Pi$ are defined as

$$\begin{aligned} \Psi &= \chi \otimes \Lambda \\ \Pi &= I_n \otimes \Lambda \end{aligned} \tag{2.4}$$

where $\otimes$ denotes the Kronecker product of two matrices. Examples of the use of the mixed linear model for different experimental conditions are provided in Sections 3 and 4.

This model provides a general framework for evaluating changes in gene expressions in different experimental conditions over time. The structure of the experiment may be fairly complex, corresponding to a factorial design or a block design (Kerr and Churchill, 2001b,a, Yang and Speed, 2002). Examples of such designs arise in the experiments of Gasch et al. (2000), Causton, Ren, Koh, Harbison, Kanin, Jennings, Lee, True, Lander, and Young (2001) and Gasch and Werner-Washburne (2002). Further, the model facilitates the specification of correlations caused by both the gene network, as well as temporal dependence among gene expressions. In fact, using the covariance matrices **R** and **G**, a variety of correlation structures can be modeled (in Section 4, we provide detailed definition of design and covariance matrices for the analysis of yeast ESR data). In addition, the proposed model allows researchers to investigate the patterns of changes of expressions in different experimental conditions, and to study the expression profiles of gene sets over time, which could provide additional cues to the behavior of biological systems. Such experiments are hard to analyze using the permutation-based enrichment analysis methods.

## 2.3 Parameter Estimation

It is easy to see that for the mixed linear model of equation 2.3 $\mathbf{W} \equiv \mathrm{Var}(\mathbf{Y}) = \sigma_\gamma^2 \Pi\Pi' + \mathbf{R}$ and the maximum likelihood estimate of $\beta$ is given by (Searle, 1971):

$$\hat{\beta} = \left(\Psi'\hat{\mathbf{W}}^{-1}\Psi\right)^{-1}\Psi'\hat{\mathbf{W}}^{-1}\mathbf{Y} \tag{2.5}$$

These estimates depend on estimates of the variance components, $\sigma_\gamma^2$ and **R**, which are usually estimated via *Restricted Maximum Likelihood* (REML).

Lindstrom and Bates (1988) provide details of the Newton-Raphson and EM algorithms for estimation of parameters of MLM and presented evidence in favor of the former method. They also present a method of reducing the dimension of

**Algorithm 1** Block-Relaxation Algorithm for MLM Parameters

```
1. Find an initial estimate of $\hat{\beta}^{(0)}$ (e.g. using OLS)
2. Repeat until convergence $m = 1, 2, \cdots$
    2.1. $\mathbf{e} := \mathbf{e}^{(m+1)} = \mathbf{Y} - \Psi\hat{\beta}^{(m)}$
    2.2. $\hat{\theta}^{(m+1)} = \mathrm{argmax}_{\theta} -\frac{1}{2}\left\{\mathrm{logdet}\left(\mathbf{W}(\theta)\right) + \mathbf{e}'\mathbf{W}^{-1}(\theta)\mathbf{e}\right\}$
    2.3. $\hat{\beta}^{(m+1)} = \left(\Psi'\mathbf{W}^{-1}(\hat{\theta}^{(m+1)})\Psi\right)^{-1}\Psi'\mathbf{W}^{-1}(\hat{\theta}^{(m+1)})\mathbf{Y}$
```

the matrices involved in the calculation by breaking down the matrices $\Psi$ and $\Pi$ into smaller sub-matrices in case of repeated measures data. In dealing with specific problems, it may be possible to further reduce the dimension of these matrices by taking advantage of their structure and sparsity patterns. However, the size of the parameter vector and dimensions of the matrices involved in the calculations increase with the number of genes $p$. As a result, traditional methods available for estimation of parameters of MLM prove inefficient in mixed linear models for large gene networks. Therefore, estimation of MLM parameters in (2.3) requires efficient estimation procedures. Algorithm 1, which is a block-relaxation type algorithm (de Leeuw, 1994), makes the estimation of parameters tractable by partitioning the parameter space into smaller subspaces. To simplify the notation, we denote by $\theta = (\theta_{\varepsilon}, \theta_{\gamma})$ the vector of all variance parameters used do define $\mathbf{R}$ and $\mathbf{G}$ i.e. $\mathbf{R} = \mathbf{R}(\theta_{\varepsilon})$ and $\mathbf{G} = \mathbf{G}(\theta_{\gamma})$. Oberhofer and Kmenta (1974) proved the convergence of this algorithm under certain assumption on the estimates of the variance components. In fact, using the strict convexity of the negative log-likelihood function for mixed linear models, and the general theory of iterative algorithms (de Leeuw, 1994), it can be shown that this algorithm converges to the maximum likelihood estimates of the MLM parameters, provided the estimates of the covariance components result in a positive definite covariance matrix, and $\Psi$ has full column rank. For the model presented here, this is achieved if the variance components are estimated using the REML estimation criterion.

To further speed up the estimation process, one can also partition the estimation further over the subnetworks, which results in partitioning over both parameter, as well as observation spaces. It can be shown that, under specific conditions, estimates from such partitioning converge to the maximum likelihood estimates of the model parameters, however this is beyond the scope of the current paper.

## 2.4 Inference

A variety of hypotheses about fixed effect parameters of mixed linear models can be tested by considering tests of the form:

$$H_0 : l\beta = 0 \quad vs. \quad H_1 : l\beta \neq 0 \tag{2.6}$$

Here $l$ is in general any linear combination of $\beta$'s which meets the estimability requirement of Searle (1971). An example of such vectors is a *contrast vector*, which satisfies the constraint $\mathbf{1}'l = 0$. In the following discussion, any linear combination of $\beta$'s satisfying the estimability requirement is referred to as a *contrast vector*. In the setting of multiple experimental conditions, this inference framework allows tests of hypotheses of significance of parameters for each experimental condition, as well as tests of significant changes in responses of gene sets over time. More generally, different combinations of parameters can be tested using this framework, which allow researchers to fully investigate the behavior of gene sets of particular interest.

It was shown in Shojaie and Michailidis (2009), that for any given $1 \times p$ indicator vector $\mathbf{b}$ determining a specific subnetwork or gene set, the vector $(\mathbf{b}\Lambda \cdot \mathbf{b})\beta$ includes the effects of all the nodes in $\mathbf{b}$ on each other, but it is not affected by any node outside the set of nodes indexed by $\mathbf{b}$ (here $\cdot$ denotes the Hadamard or componentwise product of two vectors). In words, $\mathbf{b}\Lambda$ introduces the influence of genes indexed by $\mathbf{b}$ on each other, while the componentwise product with $\mathbf{b}$ excludes the effects of nodes not in $\mathbf{b}$. To illustrate this, consider again the simple network of Figure 2, where the subnetwork of interest consists of $X_2$ and $X_3$ i.e. $b = (0,1,1)$.

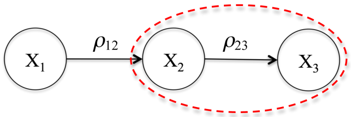

Figure 2: Illustration of the network contrast vector

It is then easy to see that

$$(\mathbf{b}\Lambda) = (\rho_{12} + \rho_{12}\rho_{23}, 1 + \rho_{23}, 1)$$

includes all the interactions among nodes connected to the subnetwork, while the proposed network contrast vector

$$(\mathbf{b}\Lambda \cdot \mathbf{b}) = (0, 1 + \rho_{23}, 1)$$

corresponds to the desired interactions. The change in $\Lambda$ in response to different experimental conditions or over time can be incorporated into this contrast vector by substituting $\Lambda$ by the influence matrix of the specific time and experimental conditions, $\Lambda^{(jk)}$. Hence, the contrast vector $l$ is formed by replacing in the general formula the influence matrix of the network under the specific conditions. As an example, suppose $\Lambda^{(j)}$ represents the influence matrix of the network at time $j$, $j = 1,\cdots,J$ and $\beta = (\beta^{(1)'},\cdots,\beta^{(J)'})'$. Then, the change in the expression levels of genes in the subnetwork indexed by $\mathbf{b}$ from time $j$ to $j+1$ can be tested using

$$l = (0,\cdots,0,-\mathbf{b}\Lambda^{(j)}\cdot\mathbf{b},\, \mathbf{b}\Lambda^{(j+1)}\cdot\mathbf{b},0,\cdots,0)$$

Letting $\mathbf{C} = (\Psi'\mathbf{W}^{-1}\Psi)^{-1}$, the significance of individual contrast vectors in (2.6) can be tested using the following Wald test statistic:

$$T = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l'}} \tag{2.7}$$

Under the null hypothesis, $T$ follows approximately a t-distribution whose degrees of freedom $\nu$ can be estimated using the Satterthwaite approximation method (McLean and Sanders, 1988)

$$\nu = \frac{2(l\hat{C}l')^2}{\tau' V \tau}$$

where $\tau = \frac{\partial}{\partial\theta} lCl'$, and $V$ is the empirical covariance matrix of $\theta$.

When analyzing complex experiments, often multiple contrast vectors of interest are considered for a specific subnetwork. In such situations, (2.7) can be used to test the significance of the contrast vector corresponding to each hypothesis of interest. The resulting p-values should then be adjusted for the total number of hypotheses tested amongst different subnetworks. Alternatively, one can combine these contrast vectors into a *contrast matrix* $L$, where each row of $L$ includes one of the contrast vectors. The significance of the subnetwork can then be tested using the following test statistic:

$$F = \frac{\hat{\beta}' L' (L\hat{C}L')^{-1} L\hat{\beta}}{q} \tag{2.8}$$

where $q$ is the rank of $L$. Under the null hypothesis of $L\beta = 0$, $F$ has an F-distribution with $q$ and $\eta$ degrees of freedom. To estimate $\eta$ using the Satterthwaite approximation method, one first needs to find matrices $P$ and $D$ such that $LCL' = PDP'$ (the eigen-decomposition of $LCL'$). Then, denoting the $m$th row of L by $l_m$, $\eta$ is calculated using:

$$\eta = \begin{cases} \frac{2E}{E-q} & E > q \\ 0 & o.w. \end{cases}$$

where

$$E = \sum_{m=1}^{q} \frac{\nu_m}{\nu_m - 2} I_{\{\nu_m > 2\}}, \qquad \nu_m = \frac{2D_m^2}{\tau_m' K \tau_m}.$$

The proposed F-test for the analysis of complex experiments reduces the number of hypotheses tested and offers a hierarchical testing approach. In particular, although some subnetworks may not show significant change with regard to individual hypotheses, the combined significance of the subnetwork due to multiple sources of differential expression may result in overall significance of the subnetwork. It is then possible to test the significance of individual hypotheses, in case the overall F-test for the subnetwork is significant. We illustrate this hierarchical testing procedure in Sections 3 and 4.

## 2.5 Uncertainty in Network Information

The method for network-based analysis of gene sets proposed here requires knowledge of interactions among genes (proteins/metabolites), as well as the corresponding association weights. In addition, to fully exploit the strength of the proposed methodology in testing the changes in the network structure, as well as the expression levels of genes, the adjacency matrix of the network should be available for different experimental conditions and time points. However, available network information may be noisy, and available resources often only determine the presence of interactions among genes, and do not provide information on the strength of associations. Therefore, it may be necessary to estimate the network information. Estimation of gene networks from high throughput observations is an important problem in systems biology and of independent interest (see e.g. Shojaie and Michailidis, 2010, for a review). It is important to note that since the network information is used in both estimation of parameters, as well as inference, to prevent unidentifiability and bias, the observations used for estimation of the underlying network should be independent from those used for analysis of differential expression.

In this section, we analyze the effect of uncertainty in the network information, by studying the asymptotic properties of the proposed test statistic. Our main result concerns the general case of error in network information in the case of a two-population test, described in Shojaie and Michailidis (2009). We also discuss the special case of estimating association weights, when the structure of the network is known.

In the following, we denote the available adjacency matrix of the network by $\tilde{A}$ and use the notation $||A||$ and $||A||_F$ to represent the matrix norm and Frobenius norm of $A$, respectively. Also, let $d_i^A$ denote the weighted *in-degree* of node $i$ based on the adjacency matrix $A$: $d_i^A = \sum_j |A_{ij}|$.

**Theorem 2.1.** *Suppose* $\tilde{A} = A + \Delta_A$*, where* $\|\Delta_A\| = o_P(1)$*, and assume that* $\min_i d_i^{\tilde{A}} \geq 1$.[1] *Then,* (2.7) *is an asymptotically most powerful unbiased test for* (2.6).

*Proof.* We consider here the special case where $\Lambda^C = \Lambda^T = \Lambda$ and only one gene set; i.e. the whole network is tested. This implies that $\mathbf{b} = \mathbf{1}'$ and the proposed network contrast vector $\mathbf{b}\Lambda \cdot \mathbf{b}$ reduces to $\mathbf{b}\Lambda$ (the general case of $\Lambda^C \neq \Lambda^T$ and $\mathbf{b} \neq \mathbf{1}'$ follows from a similar argument).

First, recall that for directed acyclic graphs (DAGs), $\Lambda = \sum_{r=0}^{\infty} A^r$, and for general graphs, $\Lambda = \sum_{r=0}^{\infty} \mathscr{L}^r$, where $\mathscr{L} = D_A^{-1}A$ and $D_A = \mathrm{diag}(d_i^A)$. Then $\tilde{A} = A + \Delta_A$ implies that for DAGs

$$\tilde{\Lambda} = \sum_{r=0}^{\infty} \tilde{A}^r = \sum_{r=0}^{\infty} A^r + \sum_{r=0}^{\infty}\sum_{s=1}^{\infty} A^r \Delta_A^s \equiv \Lambda + \Delta_\Lambda, \quad \|\Delta_\Lambda\| = o_P(1). \tag{2.9}$$

Similarly, for general graphs, we have

$$\tilde{\mathscr{L}} = D_{\tilde{A}}^{-1}\tilde{A} = D_{\tilde{A}}^{-1}(A + \Delta_A) \equiv \mathscr{L} + \Delta_{\mathscr{L}}$$

where

$$\|\Delta_{\mathscr{L}}\| \leq \|D_{\tilde{A}}^{-1}\|\|\Delta_A\| = 1/(\min_i d_i^{\tilde{A}})\|\Delta_A\| = o_P(1)$$

ex hypothesis. An argument similar to (2.9) implies that the following expression also holds for general graphs

$$\tilde{\Lambda} = \Lambda + \Delta_\Lambda, \quad \|\Delta_\Lambda\| = o_P(1) \tag{2.10}$$

Now, using the results in Shojaie and Michailidis (2009), the test statistic in (2.7) can be written as

$$T = \frac{\mathbf{b}(\bar{Y}^T - \bar{Y}^C)}{\sqrt{\mathbf{b}(n_1^{-1} + n_2^{-1})(\hat{\sigma}_\gamma^2 \tilde{\Lambda}\tilde{\Lambda}' + \hat{\sigma}_\varepsilon^2 I_p)\mathbf{b}'}} \tag{2.11}$$

where $\bar{Y}^T$ and $\bar{Y}^C$ represent the average expression of genes in the two experimental conditions and $n_1$ and $n_2$ represent the corresponding sample sizes. The test statistic in (2.11) represents the likelihood ratio test for testing the null hypothesis in (2.6), which is asymptotically most powerful unbiased, provided correct network information is given. Therefore, to establish the result, it suffices to show that the effect

[1] Note that $\min_i d_i^{\tilde{A}} \geq 1$ implies that the network is connected. However, the case of disconnected networks is a straightforward extension, as the networks can be analyzed separately.

of error in the network information is asymptotically negligible. However, since the numerator of the test in (2.11) does not depend on the network information, it suffices to show that the denominator is a consistent estimator.

To establish the consistency of estimates of the variance components, note that the negative log-likelihood function (up to a constant) for the two-population problem is given by

$$\ell(\theta) = n^{-1}\sum_{i=1}^{n} \operatorname{logdet}(W_i) + n^{-1}\sum_{i=1}^{n} r_i' W_i^{-1} r_i, \tag{2.12}$$

where $r_i = Y_i - \bar{Y}^C, i = 1, \dots, n_1$, $r_i = Y_i - \bar{Y}^T, i = n_1 + 1, \dots, n$ and $\theta$ is the vector of variance components. Then, using the fact that for the two-class problem with constant, but noisy network information $W_i = \mathrm{Var}(Y_i) = \sigma_\gamma^2 \tilde{\Lambda}\tilde{\Lambda}' + \sigma_\varepsilon^2 I_p$, we get

$$\ell(\theta;\tilde{\Lambda}) = \operatorname{logdet}\left(\sigma_\gamma^2 \tilde{\Lambda}\tilde{\Lambda}' + \sigma_\varepsilon^2 I_p\right) + n^{-1}\sum_{i=1}^{n} r_i' \left(\sigma_\gamma^2 \tilde{\Lambda}\tilde{\Lambda}' + \sigma_\varepsilon^2 I_p\right)^{-1} r_i. \tag{2.13}$$

Using (2.10) we can then approximate $\ell(\theta;\tilde{\Lambda})$ with its one-term Taylor expansion around $\Lambda$

$$\ell(\theta;\tilde{\Lambda}) = \ell(\theta;\Lambda) + \|\Delta_\Lambda\| \operatorname{trace}\left[(\nabla_\Lambda \ell(\theta;\Lambda))' \Delta_\Lambda / \|\Delta_\Lambda\|\right] + o(\|\Delta_\Lambda\|^2) \tag{2.14}$$

where $\nabla_\Lambda \ell(\theta;\Lambda)$ is the gradient of $\ell$ with respect to $\Lambda$ (see e.g. Dattorro (2005)-Appendix D for details on directional derivatives and approximations for functions of matrices).

But, for square positive definite matrices we have $\nabla \operatorname{logdet}(X) = X^{-1}\nabla X$ and $\nabla X^{-1} = -X^{-1}\nabla X X^{-1}$. Hence, noting that $\nabla_\Lambda \Lambda\Lambda' = (\Lambda' + \Lambda)$, by the chain rule and after some algebra, we can write

$$\begin{aligned}
\nabla_\Lambda \ell(\theta;\Lambda) \quad = \quad & \sigma_\gamma^2 (\sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p)^{-1} (\Lambda' + \Lambda) \\
& - n^{-1} \sigma_\gamma^2 \sum_{i=1}^{n} r_i' \left(\sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p\right)^{-1} (\Lambda' + \Lambda) \left(\sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p\right)^{-1} r_i.
\end{aligned}$$

Let $\tau^2 = \sigma_\varepsilon^2 / \sigma_\gamma^2$ and $\Gamma = \Delta_\Lambda / \|\Delta_\Lambda\|$, and denote

$$\begin{aligned}
g(\theta) \quad \equiv \quad & \operatorname{trace}\left[\Gamma' (\Lambda\Lambda' + \tau^2 I_p)^{-1} (\Lambda' + \Lambda)\right] \qquad (2.15) \\
& - n^{-1} \sigma_\gamma^{-2} \sum_{i=1}^{n} r_i' \left(\Lambda\Lambda' + \tau^2 I_p\right)^{-1} \Gamma' (\Lambda' + \Lambda) \left(\Lambda\Lambda' + \tau^2 I_p\right)^{-1} r_i.
\end{aligned}$$

Replacing (2.15) in (2.14) then gives

$$\ell(\theta;\tilde{\Lambda}) = \ell(\theta;\Lambda) + \|\Delta_\Lambda\| g(\theta) + o(\|\Delta_\Lambda\|^2).$$

However,

$$\begin{aligned}
|g(\theta)| &\leq \left|\mathrm{trace}\left(\Gamma'(\Lambda\Lambda'+\tau^2 I_p)^{-1}(\Lambda'+\Lambda)\right)\right| \\
&\quad + n^{-1}\sigma_\gamma^{-2}\sum_{i=1}^{n} r_i'(\Lambda\Lambda'+\tau^2 I_p)^{-1}\Gamma'(\Lambda'+\Lambda)(\Lambda\Lambda'+\tau^2 I_p)^{-1} r_i \\
&\equiv \mathrm{I}+\mathrm{II}.
\end{aligned}$$

Using von Neumann's inequality for the matrix trace (see e.g. Mirsky, 1975), and the relationship between singular values and matrix norms, we get

$$\begin{aligned}
\mathrm{I} &\leq \sum_{j=1}^{p} \kappa_{[j]}([\Lambda'+\Lambda]\Gamma')\kappa_{[j]}([\Lambda\Lambda'+\tau^2 I_p]^{-1}) \\
&\leq p\kappa_{[1]}([\Lambda'+\Lambda]\Gamma')\kappa_{[1]}([\Lambda\Lambda'+\tau^2 I_p]^{-1}) \\
&\leq p\kappa_{[1]}(\Lambda'+\Lambda)\kappa_{[1]}(\Gamma)\kappa_{[1]}([\Lambda\Lambda'+\tau^2 I_p]^{-1})
\end{aligned}$$

where $\kappa_{[j]}(A)$ denotes the $j$-th largest singular value of $A$. But, by definition, $\kappa_{[1]}(\Gamma)=1$. Moreover, by construction, $\kappa_{[1]}(\Lambda'+\Lambda)$ is bounded by say $M$, and $\kappa_{[1]}([\Lambda\Lambda'+\tau^2 I_p]^{-1}) = 1/(\lambda_{[p]}(\Lambda\Lambda')+\tau^2)$, where $\lambda_{[p]}(\Lambda\Lambda')$ is the smallest eigenvalue of $\Lambda\Lambda'$ and hence is positive (by definition of $\Lambda$). This implies that $\mathrm{I} < 2pM/\tau^2$.

On the other hand,

$$\begin{aligned}
\mathrm{II} &\leq \sigma_\gamma^{-2}\|(\Lambda\Lambda'+\tau^2 I_p)^{-1}\Gamma'(\Lambda'+\Lambda)(\Lambda\Lambda'+\tau^2 I_p)^{-1}\| n^{-1}\sum_{i=1}^{n} r_i' r_i \\
&\leq \sigma_\gamma^{-2}\|(\Lambda\Lambda'+\tau^2 I_p)^{-1}\|^2\|\Gamma'\|\|(\Lambda'+\Lambda)\|\|n^{-1}\sum_{i=1}^{n} r_i' r_i \\
&< \sigma_\gamma^{-2} 2\tau^{-4} M n^{-1}\sum_{i=1}^{n} r_i' r_i = 2\tau^{-2}\sigma_\varepsilon^{-2} M\mathbb{E}(\|r_i\|^2) \quad \text{w.p.1},
\end{aligned}$$

where the last step follows from the strong law of large numbers. This implies that provided the variance components are non-zero, with probability one $g(\theta)$ is bounded, and hence $\|\Delta_\Lambda\| g(\theta) = o_P(1)$. This in turn implies that $\ell(\theta;\tilde{\Lambda}) = \ell(\theta;\Lambda)+o_P(1)$.

Denote by $\mathscr{E}$ the event $[\ell(\theta;\tilde{\Lambda}) = \ell(\theta;\Lambda)]$. Then conditioning on $\mathscr{E}$, the estimates of the variance components are found by minimizing the negative log-likelihood function with true network information, which is a convex function of variance components. M-estimation results in Haberman (1989) imply that $\mathbb{P}(\hat{\theta}=\theta|\mathscr{E}) = 1$ and hence, $\hat{\theta}\rightarrow_P \theta$ as $\tilde{A}\rightarrow_P A$. However, this further implies that as $\tilde{A}\rightarrow_P A$, the denominator of the test statistic in (2.7) converges to the true value, and the result follows. □

*Remark* 2.2. In the general case of complex experiments, the estimates of the fixed effects are also dependent on the network information. A similar result will then follow upon deriving the asymptotic distribution of the numerator of the test statistic in (2.7). In Section 3.3, we provide empirical evidence in support of the insensitivity of the proposed inference framework to the presence of noisy network information.

The above theorem guarantees that as long as the error is small in magnitude, the network-based inference procedure correctly determines the significance of the gene sets. In other words, a necessary condition for the proposed method to work in presence of noise in the network information is that $\|\Delta_A\| = o_P(1)$. As mentioned earlier, the problem of estimation of network structure for directed, as well as undirected, networks is an important problem in multivariate statistics and researchers have studied asymptotic properties of network estimation for different classes of problems. Here, we consider a special case of the problem of estimating high dimensional networks, where the structure of the network is known, and the problem is reduced to estimating association weights among genes. The following corollary shows that the proposed network-based gene set analysis procedure is not sensitive to the estimation noise in this setting. It is important to note that the conditions of this result only limit the degree of nodes in the graph and no constraint is required on the total number of nodes in the graph. In the following, $d_i$ represents the unweighted in-degree of node $i$: the number of neighbors of $i$ in undirected graphs and the number of parents of $i$ in directed graphs.

**Corollary 2.3.** *Let $\mathscr{G}$ be a DAG or an undirected graph, with p nodes and adjacency matrix A. Assume that* $\max_i(d_i) = n^b$ *for some* $0 < b < 1$ *and* $\sum_{i\in\mathscr{G}}(d_i) = n^a$ *for some* $a > 0$. *Further, assume that the structure (or skeleton) of the network is known, but the network information is obtained by estimating the association weights from an independent sample of size n. Then, the test statistics in* (2.7) *is an asymptotically unbiased most powerful test for* (2.6).

*Proof.* By Theorem 2.1, it suffices to show that $\|\hat{A} - A\| = o_P(1)$. First, assume that $\mathscr{G}$ is a DAG. Then, by the results in Shojaie and Michailidis (2010), to find the association weights one needs to regress each node on the set of the parents of that node. Since $\max_i(d_i) = o(n)$, without loss of generality, we can assume that $\max_i(d_i) < n$, and therefore regular regression can be used to estimate the weights. The asymptotic normality of regression estimators then implies that each non-zero entry of the adjacency matrix converges with an exponential rate to the true value. Bonferroni's inequality and the fact that the total number of edges in the graph is a polynomial function of the sample size imply that $\|\hat{A} - A\| = o_P(1)$.

For undirected graphs, we note that partial correlations between each node $i$ and its neighbors $\mathrm{ne}_i$ can be recursively estimated using the following formula:

$$\rho_{i,j|\mathrm{ne}_i} = \frac{\rho_{i,j|\mathrm{ne}_i\backslash h} - \rho_{i,h|\mathrm{ne}_i\backslash h}\rho_{j,h|\mathrm{ne}_i\backslash h}}{\sqrt{(1-\rho^2_{i,h|\mathrm{ne}_i\backslash h})(1-\rho^2_{j,h|\mathrm{ne}_i\backslash h})}}$$

However, Corollary 1 of Kalisch and Bühlmann (2007) implies that if $\max_i(d_i) < n-4$ estimated partial correlations converge to true values with an exponential rate. An argument similar to the case of DAGs then implies that $\|\hat{A} - A\| = o_P(1)$ and the result follows. □

# 3 Performance Analysis

In this section, we evaluate the small sample properties of the proposed inference procedure, through several simulation studies. In all settings, data are generated from a mixed linear model, where the Gaussian noise has an AR(1) correlation structure. We consider different combinations of mean and network information, and investigate the effects of temporal correlation, as well as noise in the network information.

## 3.1 Multiple Experimental Conditions

The first simulation depicts the real data example of Section 4, which corresponds to analysis of responses of yeast cells to environmental stress factors. The network consists of a directed graph with 7 subnetworks and a total of 220 nodes. Each subnetwork in turn consists of a 4-level binary tree and a "hub" node. There are also 3 gateway genes that connect the subnetworks together. The adjacency matrix of the graph is considered to remain constant in different experimental conditions and different time points. The model includes changes in gene expressions under different experimental conditions and different time points. Specifically,

$$\begin{array}{rcll} \mathbb{E}Y_{11} & = & \Lambda\mu & \\ \mathbb{E}Y_{1k} & = & \Lambda(\mu+\delta_k), & k=2,3 \\ \mathbb{E}Y_{jk} & = & \Lambda(\mu+\alpha_j+\delta_k), & j,k=2,3 \end{array} \tag{3.1}$$

The settings of parameters in the first simulation are given in Table 1. Table 2 includes the estimated powers of the t-tests for different mean parameters, as well as powers of the F-test, for the overall significance of the subnetwork, estimated

Table 1: Parameter settings for the first simulation study.

| Subnetwork | Non-zero Mean Parameters | | | |
|---|---|---|---|---|
| 1 | – | | | |
| 2 | $\alpha_2 = 2$ | | | |
| 3 | $\alpha_2 = 1,$ | $\delta_2 = 1$ | | |
| 4 | $\alpha_2 = 1,$ | $\alpha_3 = 1$ | | |
| 5 | $\alpha_2 = 1,$ | $\delta_3 = 1$ | | |
| 6 | $\alpha_2 = 1,$ | $\alpha_3 = 1,$ | $\delta_2 = 1$ | |
| 7 | $\alpha_2 = 1,$ | $\alpha_3 = 1,$ | $\delta_2 = 1,$ | $\delta_3 = 1$ |

from 100 replications [2] with $n = 1$ observations at each combination of experimental condition and time point. To prevent redundancy, the contrast matrix $L$ (see Section 2.4) consists only of contrast vectors used for the main effects (the parameters in the first 4 columns of the Table 2).

It can be seen from these results that when the model is correctly specified, the proposed inference procedure offers high power for detecting non-zero parameters, while maintaining close to nominal significance levels for non-significant parameters.

Table 2: Estimated powers of t-test and F-test for the first simulation study. The first four columns of the table represent the powers for testing the significance of the mean parameters ($\alpha_2$, $\delta_2$, $\alpha_3$ and $\delta_3$ respectively). The powers for testing equality of main effects ($\alpha_2 = \alpha_3$ and $\delta_2 = \delta_3$) are given in the next two columns of the table. Entries in bold represent result of potential interest.

| | Individual Parameters (t-test) | | | | | | Subnetwork |
|---|---|---|---|---|---|---|---|
| Subnetwork | $\alpha_2$ | $\delta_2$ | $\alpha_3$ | $\delta_3$ | $\alpha_2 - \alpha_3$ | $\delta_2 - \delta_3$ | (F-test) |
| 1 | 0.006 | 0.06 | 0.03 | 0.14 | 0.01 | 0.10 | 0.12 |
| 2 | **1.00** | 0.10 | 0.02 | 0.09 | **1.00** | 0.13 | **1.00** |
| 3 | **0.99** | **1.00** | 0.03 | 0.05 | **0.99** | **1.00** | **1.00** |
| 4 | **0.98** | 0.09 | **1.00** | 0.07 | 0.02 | 0.08 | **1.00** |
| 5 | **0.99** | 0.08 | 0.02 | **1.00** | **0.99** | **1.00** | **1.00** |
| 6 | **0.99** | **1.00** | **1.00** | 0.05 | 0.01 | **1.00** | **1.00** |
| 7 | **1.00** | **0.99** | **1.00** | **1.00** | 0.00 | 0.01 | **1.00** |

[2] Simulation replicates are obtained by generating data sets according to the same model with different realizations of the random vectors $\varepsilon$ and $\gamma$.

## 3.2 Effect of Temporal Correlation

The second simulation setting aims to illustrate the effects of temporal correlation, as well as changes in the network structure, in different experimental conditions. Since gene set enrichment analysis methods do not directly incorporate complex experiments, we consider a simple experimental design, including two experimental conditions. However, to illustrate the effect of temporal correlation, we consider the case where data are generated over 5 time points with no replicates. The temporal correlation among observations is generated using an AR(1) process with autocorrelation parameter $\phi$. We consider a network consisting of 4 non-overlapping subnetworks (as described in the first simulation) regulated by 3 hub genes. The correlation among genes in each subnetwork is controlled by a single parameter $\rho$, with different values in distinct subnetworks and experimental conditions. The parameter settings for this simulation are given in Table 3.

Table 3: Parameter settings for the second simulation study. $\alpha_i$ and $\rho_i, i = 1,2$ correspond to the $i$th experimental condition.

| Subnet | Mean Parameters | Correlation Parameters |
|---|---|---|
| 1 | $\alpha_1 = \alpha_2 = 1$ | $\rho_1 = \rho_2 = 0.2$ |
| 2 | $\alpha_1 = 1, \alpha_2 = 2$ | $\rho_1 = \rho_2 = 0.2$ |
| 3 | $\alpha_1 = \alpha_2 = 1$ | $\rho_1 = 0.2, \rho_2 = 0.7$ |
| 4 | $\alpha_1 = 1, \alpha_2 = 2$ | $\rho_1 = 0.2, \rho_2 = 0.7$ |

Given the true values of the parameters, the test statistic in (2.7) has a normal distribution, with means 0 and $l\beta$ under the null and alternative hypotheses, respectively. Hence, it is possible to calculate the true asymptotic powers of rejecting the null hypotheses for each of the subnetworks in this simple setting. Figure 3 includes the estimated powers of tests using GSEA and the proposed network-based method (NetGSA), based on 100 replications, along with the true asymptotic powers of the corresponding tests. It can be seen that when the parameters are clearly insignificant or demonstrate strong significance (Subnetworks 1 and 4), both methods correctly determine the significance of the test. However, in less extreme scenarios (e.g. Subnetworks 2 and 3), the presence of temporal correlation along with the small sample size ($n = 1$) prevent GSEA from correctly determining the statistical significance of subnetworks. On the other hand, by accounting for the temporal correlation, NetGSA offers considerable improvement over GSEA.

Table 4 includes the details of estimated and true powers of tests of significance of subnetworks considered in Simulation 2. In order to investigate the effect of the sample size $n$ on the power of the tests, we also consider the case of 10 independent samples for each experimental condition ($n = 10$). Powers of the tests with

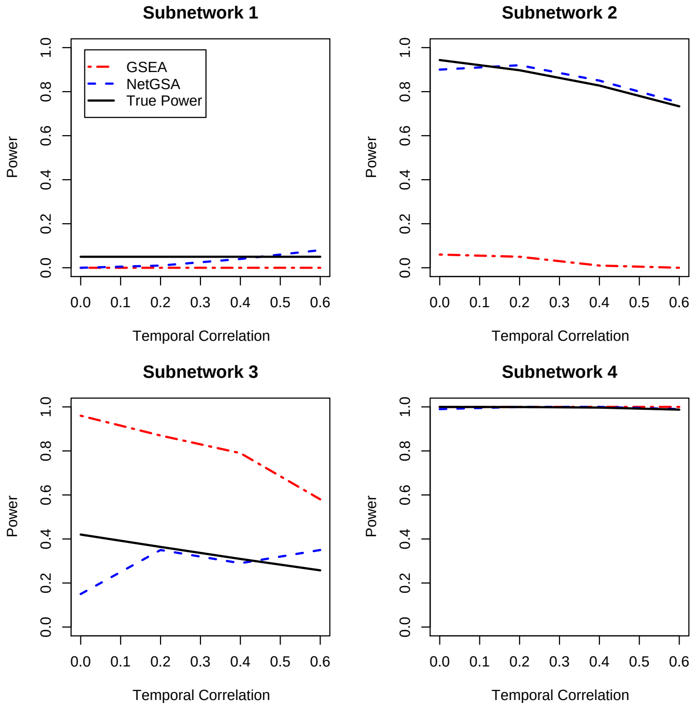

Figure 3: Estimated and true powers for tests of subnetworks in Simulation 2.

$n = 10$ are presented in Table 5. This table indicates that estimated powers of the proposed NetGSA method are more consistent with the values of the true powers for larger sample sizes. In addition, the presence of temporal correlation prevents the GSEA method from distinguishing the significance of subnetwork 2 even with larger sample sizes.

## 3.3 Uncertainty in Network Information

In Section 2.5, we showed that the proposed inference procedure is asymptotically insensitive to small noise in the network information, in case of the simple two-class problems. We also argued that similar results can be expected in more complex experiments. Here we provide empirical evidence for the robustness of the proposed method to noise in the network information in presence of temporal correlation, by considering the simulation settings of Section 3.2, with $n = 1$. The settings of mean and correlation parameters are identical to those in Table 3. In addition, the temporal correlation is fixed at $\phi = 0.4$. In each case, the data are generated according to the mixed linear model with the true network information,

Table 4: Powers of second simulation study, $n = 1$. Entries in bold represent result of potential interest.

| | | $\phi$ | | | |
|---|---|---|---|---|---|
| | | 0 | 0.2 | 0.4 | 0.6 |
| Subnetwork 1 | GSEA | 0.00 | 0.00 | 0.00 | 0.00 |
| | NetGSA | 0.00 | 0.01 | 0.04 | 0.08 |
| | True Power | 0.05 | 0.05 | 0.05 | 0.05 |
| Subnetwork 2 | GSEA | **0.06** | **0.05** | **0.01** | **0.00** |
| | NetGSA | **0.90** | **0.92** | **0.85** | **0.75** |
| | True Power | **0.94** | **0.90** | **0.83** | **0.73** |
| Subnetwork 3 | GSEA | **0.96** | **0.87** | **0.79** | **0.58** |
| | NetGSA | **0.15** | **0.35** | **0.29** | **0.35** |
| | True Power | **0.42** | **0.36** | **0.31** | **0.26** |
| Subnetwork 4 | GSEA | 1.00 | 1.00 | 1.00 | 1.00 |
| | NetGSA | 0.99 | 1.00 | 1.00 | 0.99 |
| | True Power | 1.00 | 1.00 | 0.99 | 0.99 |

Table 5: Powers of second simulation study, $n = 10$. Entries in bold represent result of potential interest.

| | | $\phi$ | | | |
|---|---|---|---|---|---|
| | | 0 | 0.2 | 0.4 | 0.6 |
| Subnetwork 1 | GSEA | 0.00 | 0.00 | 0.00 | 0.00 |
| | NetGSA | 0.00 | 0.02 | 0.00 | 0.06 |
| | True Power | 0.05 | 0.05 | 0.05 | 0.05 |
| Subnetwork 2 | GSEA | **0.06** | **0.04** | **0.03** | **0.05** |
| | NetGSA | **1.00** | **1.00** | **1.00** | **1.00** |
| | True Power | **1.00** | **1.00** | **1.00** | **1.00** |
| Subnetwork 3 | GSEA | 0.95 | 0.96 | 0.87 | **0.61** |
| | NetGSA | 1.00 | 1.00 | 0.98 | **0.96** |
| | True Power | 0.99 | 0.99 | 0.99 | **0.98** |
| Subnetwork 4 | GSEA | 1.00 | 1.00 | 1.00 | 1.00 |
| | NetGSA | 1.00 | 1.00 | 1.00 | 1.00 |
| | True Power | 1.00 | 1.00 | 1.00 | 1.00 |

and estimation and inference is carried out using a perturbed version of the network information. The network information is perturbed by adding an i.i.d. uniform random variable $U \sim \text{Uni}[-\nu, \nu]$ to each non-zero entry of the adjacency matrix; $\nu$ is hence the level of random noise in the network. Figure 4 illustrates the estimated and the true powers for different levels of noise $\nu$. It can be seen that even with small sample sizes ($n = 1$), the estimated powers are similar to the expected ones,

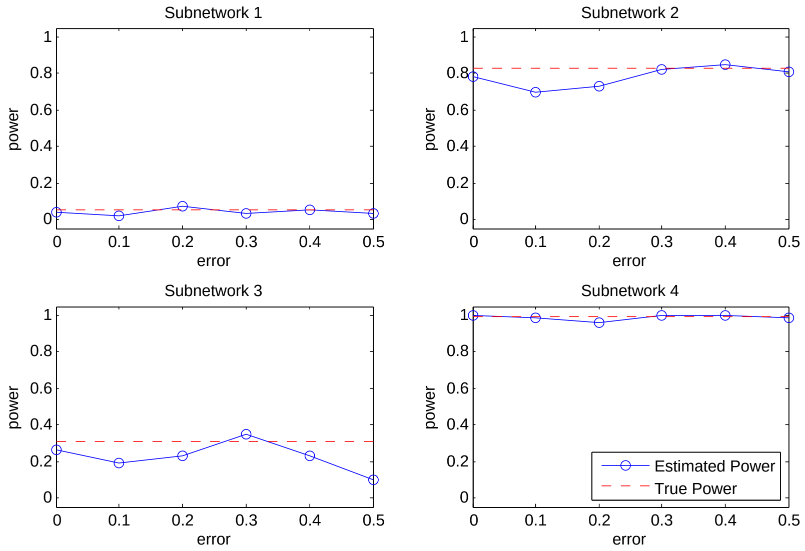

Figure 4: Estimated and true powers for tests of subnetworks in Simulation 3 for different values of random noise.

and the powers with noisy network information are similar to those obtained in the absence of noise ($\nu = 0$). The results of this simulation indicate that the proposed method is robust to small levels of noise (e.g. up to $\sim 30\%$). In addition, the effect of noise is mainly significant in the case of Subnetwork 3, where the difference in the two populations is mainly due to the changes in the network information.

## 3.4 Changes in the Network in Complex Experiments

Our final simulation setting aims to further illustrate the effect of change in the weighted adjacency matrix of the graph in complex experiments. We consider a model with separate intercept and slope parameters, for each of the three treatment conditions. In other words,

$$\mathbb{E}Y_{kj} = \Lambda^{(k)}\alpha_k + \Lambda^{(k)}\delta_k t_j, \quad j,k = 1,2,3, \quad t = (5,15,30).$$

We consider the directed graph of the first simulation setting, but here we allow for changes in both the adjacency matrix of subnetworks, as well as the mean parameters. For illustration purposes, the adjacency matrix of each subnetwork (and hence its influence matrix) is considered to be a function of a single parameter $\rho$ with values in $(-1,1)$, and entries of the adjacency matrix may attain different values in each of the three treatment conditions. Based on the latent variable model, as $\rho$ increases, genes in the network would have higher effects on their neighbors. In this setting, subnetworks 2 and 6 only include changes in the fixed effect parameters. Subnetworks 1 and 7 have moderate changes in the fixed effect parameters coupled with changes in associations among genes. In subnetwork 3, the association among genes is the only source of change. Finally, the parameters of subnetwork 5 are designed so that the individual change in the parameters is not significant; however, the combined effect of changes in intercept and slope parameters is expected to be significant. Table 6 shows the settings of the parameters for this simulation.

Table 6: Significant parameters for the fourth simulation study (unlisted parameters: $\alpha_{null} = 1$, $\delta_{null} = 0.02$ and $\rho_{null} = 0.2$).

| | Significant Parameters | |
|---|---|---|
| Subnetwork | Mean | Influence Matrix |
| 1 | $\alpha_3 = 2$ | $\rho_3 = 0.7$ |
| 2 | $\alpha_3 = 3$ | – |
| 3 | – | $\rho_3 = 0.7$ |
| 4 | – | – |
| 5 | $\alpha_3 = 1.5, \delta_3 = 0.04$ | – |
| 6 | $\delta_3 = 0.10$ | – |
| 7 | $\delta_3 = 0.06$ | $\rho_3 = 0.7$ |

Table 7: Estimated powers of F-test and t-test for the fourth simulation setting. Entries in bold represent results of potential interest.

| | Individual Parameters (t-test) | | | | | | Subnetwork |
|---|---|---|---|---|---|---|---|
| Subnetwork | $\alpha_1 - \alpha_2$ | $\delta_1 - \delta_2$ | $\alpha_1 - \alpha_3$ | $\delta_1 - \delta_3$ | $\alpha_2 - \alpha_3$ | $\delta_2 - \delta_3$ | (F-test) |
| 1 | 0.102 | 0.094 | **0.991** | 0.066 | **0.975** | 0.098 | **0.982** |
| 2 | 0.099 | 0.081 | **0.983** | 0.073 | **0.988** | 0.091 | **0.991** |
| 3 | 0.091 | 0.085 | **0.343** | 0.052 | **0.355** | 0.102 | **0.409** |
| 4 | 0.103 | 0.082 | 0.121 | 0.080 | 0.122 | 0.100 | 0.029 |
| 5 | 0.122 | 0.138 | **0.467** | **0.213** | **0.447** | **0.253** | **0.900** |
| 6 | 0.131 | 0.112 | 0.100 | **0.989** | 0.161 | **0.958** | **0.961** |
| 7 | 0.121 | 0.150 | 0.365 | **0.900** | 0.364 | **0.856** | **0.992** |

Table 7 includes the estimated powers of F and t-tests. It can be seen that powers of tests are higher than the significance level of 0.05 in cases where no changes are present. This may be attributed to the small sample size ($n = 1$). In such cases, family-wise error rates could provide more conservative inference. It can also be seen that the tests are sensitive to changes in the fixed effect parameters, as well as

Table 8: Setting of parameters in the yeast ESR experiment

| Experiment | Time points |
|---|---|
| Mild Heat Shock (29C to 33C) | 5, 15, 30 min after 33C |
| Mild Heat Shock with 1M sorbitol at 29C and 33C | 5, 15, 30 min after 33C |
| Mild Heat Shock with 1M sorbitol at 29C | 5, 15, 30 min after 33C |

associations among genes, and the (positive) change in associations magnifies the change in the fixed effect parameters. Estimated powers for subnetwork 5 indicate that even if the individual effects are not strongly significant (low estimated powers of t-test for intercept and slope), their combined effect, represented by the power of the F-test, can be significant.

# 4 Yeast Environmental Stress Response (ESR)

The ability to respond to environmental changes is important for competitive fitness and survival of living organisms; understanding the response of cells to environmental changes can provide clues to molecular mechanisms that regulate gene expression in response to these changes (Causton et al., 2001). Cells respond to environmental stress factors through a complicated process that is often observable at the expression levels of a large class of genes. Gasch et al. (2000) studied the response of yeast *Saccharomyces cerevisiae* to a wide range of environmental stress factors, and observed the expression levels of genes in the yeast genome over different time intervals. Experimental settings included responses to temperature shocks, toxic chemicals and osmotic changes.

To illustrate the proposed network-based model, we selected a subset of the data available from Gasch et al. (2000). This particular set of experiments studies the response of yeast cells to mild heat shock at different levels of osmolarities (different amounts of sorbitol in the environment). The gene expressions were obtained at three different time points after the cells were resuspended at the final temperature. Table 8 provides the detailed settings of the experiment.

In order to apply our proposed network-based method, we need external information on the weighted adjacency matrix of the underlying gene network. YeastNet is a publicly available database, which includes genes whose functional interactions are verified by integrating a large number of available genomic and proteomic data sets (Lee, Li, and Marcotte, 2007). The result of this integration is a network of $\sim 102,000$ interactions among $\sim 5,900$ genes, covering 95% of known yeast genes. However, YeastNet only provides information on the topology of the network (connections between genes) and does not include the strengths of association of gene interactions. Different methods can be used to efficiently calculate association strengths of gene interactions, when the topology of the network is known (see e.g. Chaudhuri, Drton, and Richardson, 2007). The gene expression data provided in Gasch et al. (2000) includes additional experiments independent of those studied in this section, which can be used to derive association weights. Following connections to graphical models, we estimate the association weight of each edge by the partial correlation coefficient of the corresponding pair of genes (after correcting for time dependence). However, since the additional data do not reflect the same experimental settings, it is not possible to estimate separate influence matrices for different combinations of time and experimental conditions, and hence we ignore this variability. Using additional samples, one could calculate the influence matrix of the network for each of the 9 combinations of experimental conditions and time points, and incorporate these matrices in the design matrices for fixed and random effect parameters.

We are interested in determining pathways that are perturbed in response to the combinations of heat shock and variable osmolarities, as well as those whose expression profiles exhibit significant changes over time. To determine biologically relevant pathways, we use information on gene functions provided in the data set from Gasch et al. (2000), derived from the Gene Ontology (Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, Eppig et al., 2000). We define genetic pathways of interest by combining genes with similar functions into gene sets. Pathways with at least 5 genes are considered, and a total of 73 pathways and 2784 genes ($p = 2784$) with known functions are included in our analysis.

Since there are no replicates available in this data set, it is not possible to include any interaction terms in the model. Hence, we use the model in (4.1) to analyze the variations in gene expressions over time, and in response to different levels of sorbitol in the environment.

$$\begin{aligned} \mathbb{E}Y_{11} &= \Lambda\mu && \\ \mathbb{E}Y_{1k} &= \Lambda(\mu+\delta_k), && k=2,3 \\ \mathbb{E}Y_{jk} &= \Lambda(\mu+\alpha_j+\delta_k), && j,k=2,3 \end{aligned} \tag{4.1}$$

Here $\alpha_j$ and $\delta_k$ represent the change from the baseline condition for $j$th time and $k$th experimental conditions and the temporal correlation among gene expressions is taken into account via an AR(1) model.[3]

The design matrices $\Psi$ and $\Pi$ are $9p \times 5p$ and $9p \times 9p$ matrices and the covariance matrix of $\mathbf{Y}$ is also $9p \times 9p$. In particular, denoting by $\phi$ be the AR(1)

[3] The model in (4.1) is a simplification of $\mathbb{E}Y_{jk} = \Lambda(\alpha_j+\delta_k)$, $j,k = 1,2,3$, where to reduce the number of parameters, the baseline case of $j = k = 1$ is represented with a single parameter $\mu$.

parameter and by $\sigma_\varepsilon^2$ and $\sigma_\gamma^2$ the variance components for $\varepsilon$ and $\gamma$, the vector of variance parameters is $\theta = (\sigma_\gamma^2, \sigma_\varepsilon^2, \phi)$. Then using the notation of Section 2.2, $\mathbf{G} = \sigma_\gamma^2 \mathbf{I}_{9p}$, $\Pi = I_n \otimes \Lambda$, and $\mathbf{R} = \sigma_\varepsilon^2 I_3 \otimes R$, where

$$R = \begin{bmatrix} I & \phi I & \phi^2 I \\ \phi I & I & \phi I \\ \phi^2 I & \phi I & I \end{bmatrix}.$$

Finally, the design matrix for the fixed effect parameters is set up using (2.4) with $\chi$ the design matrix for a single gene according to the model in (4.1). Specifically,

$$\Psi = \chi \otimes \Lambda = \begin{pmatrix} \Lambda & & & & \\ \Lambda & \Lambda & & & \\ \Lambda & & \Lambda & & \\ \Lambda & & & \Lambda & \\ \Lambda & \Lambda & & \Lambda & \\ \Lambda & & \Lambda & \Lambda & \\ \Lambda & & & & \Lambda \\ \Lambda & \Lambda & & & \Lambda \\ \Lambda & & \Lambda & & \Lambda \end{pmatrix}.$$

Using the FDR controlling procedure of Benjamini and Hochberg (1995) with $q^* = 0.05$, 47 pathways show significant changes in response to the experimental conditions and/or over time. Figures 5 and 6 depict the gene network of yeast and some of the significant pathways, respectively. Figure 5 provides a general overview of the whole network where the edges between the nodes are removed and the genes are classified into significant and nonsignificant in order to illustrate the pattern of differential expression throughout the network (clusters of significant and nonsignificant genes point to the corresponding pathways). Figure 6 looks more closely at some of the significant pathways with different degrees of connectivity, and both positive and negative associations among genes. Genes that appear to be isolated are in fact connected to the pathway through other genes that have been omitted when displaying each subnetwork separately.

Gasch et al. (2000) reported that about 900 genes showed significant changes of expression in response to environmental stress factors (over all experimental settings). They also classified the expression levels of these genes into two dominant patterns of expressions. The first set included about 600 genes, which were repressed in ESR, while the rest of genes were induced in ESR. Based on this analysis, genes repressed in ESR are involved in growth-related processes, various aspects of RNA metabolism, nucleotide biosynthesis, secretion, as well as the genes encoding ribosomal proteins. On the other hand, many genes induced in ESR are

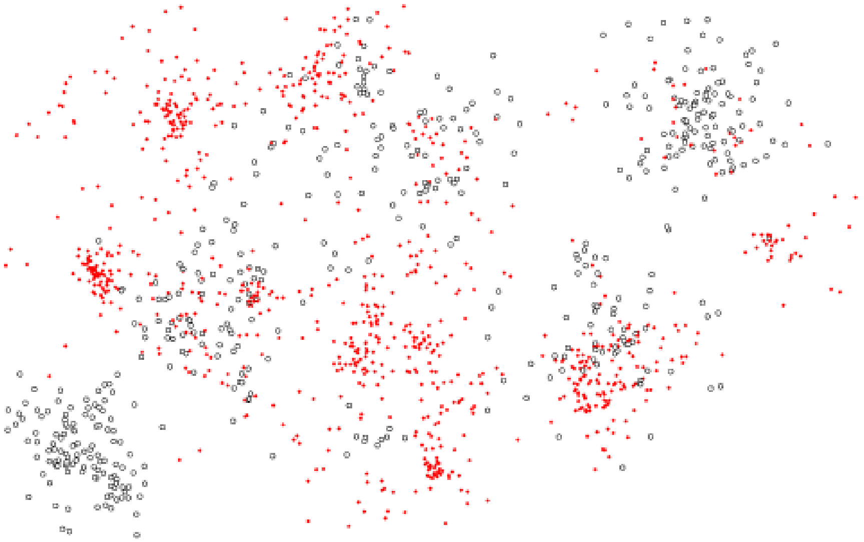

Figure 5: Network of yeast genes considered in the analysis of ESR. Red solid diamonds and empty circles represent genes in significant and nonsignificant pathways, respectively. The plot is drawn using cytoscape 2.6 (`www.cytoscape.org`).

considered to offer cellular protection during stressful conditions, such as heat and osmotic shocks which were considered in our analysis. Some of these processes include Carbohydrate Metabolism, Cell Wall Modifications, Protein Folding And Degradation, DNA Damage Repair, Fatty Acid Metabolism, Metabolite Transport and Intracellular Signalling (see Gasch et al. (2000) for more details on the functions of genes repressed and induced in ESR).

Classification of genes by their functions is facilitated through our network-based enrichment analysis approach, and many of the processes reported in Gasch et al. (2000) are also found significant based on our proposed method. Moreover, examination of the estimated fixed effects allows us to study the pattern of expression of the significant pathways over time and under different levels of sorbitol.

Tables 9 includes the list of significant pathways in analysis of yeast ESR data, along with the p-values from the corresponding F-test. Table 10 provides a list of pathways that show changes of expression over time, as well as pathways that have different expression patterns in different experimental conditions (sorbitol levels). In this table, 24 pathways show significant changes of expression over time, 29 pathways correspond to the change in sorbitol level, and 12 pathways provide evidence for both type of changes. This analysis reveals new features of environmental stress response, by determining which pathways are activated in response to different changes in the cell's environment. Pathways whose expression levels

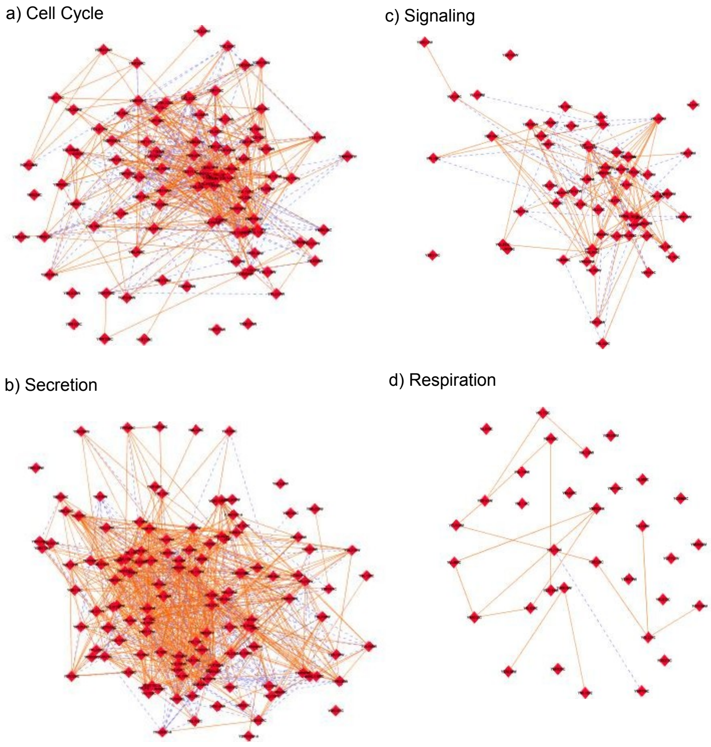

Figure 6: Selected significant pathways considered in analysis of yeast ESR. Solid orange edges indicate positive interactions and dashed blue edges represent negative associations among genes. Plots are drawn using cytoscape 2.6.

do not change in response to sorbitol levels, are only activated in response to heat shock, an obvious example of such pathways being the Heat Shock Response. On the other hand, pathways that only demonstrate significant changes in response to sorbitol level are activated when the osmolarity level of the cell's environment is perturbed. Pathways that demonstrate changes in response to both types of changes include both induced and repressed pathways under ESR. Secretion, DNA Replication, rRNA Processing and Amino Acid Metabolism are examples of pathways that are repressed in the ESR, while different carbohydrate and fatty acid metabolism pathways as well as Oxidative Stress Response are induced under ESR.

Figure 7 provides an alternative view of the changes of expressions in response to environmental stress. In this plot, the average standardized expression levels of pathways, based on the value of the test statistics for each of the significant pathways, is displayed. The pathways are divided into induced and suppressed, based on their value of test statistic at time $t = 5$. As observed in Gasch et al. (2000), it

Table 9: Significant pathways in the analysis of yeast environmental stress response (ESR) data.

| | Pathway Name | P-Value (F-test) | Pathway Size |
|---|---|---|---|
| 1 | PROTEIN SYNTHESIS | 0 | 286 |
| 2 | TRANSPORT | 0 | 143 |
| 3 | SECRETION | 0 | 126 |
| 4 | CELL CYCLE | 0 | 97 |
| 5 | CYTOSKELETON | 0 | 83 |
| 6 | LIPID METABOLISM | 0 | 63 |
| 7 | AMINO ACID BIOSYNTHESIS | 0 | 60 |
| 8 | DNA REPAIR | 0 | 58 |
| 9 | DNA REPLICATION | 0 | 57 |
| 10 | MEIOSIS | 0 | 52 |
| 11 | PROTEIN GLYCOSYLATION | 0 | 51 |
| 12 | PROTEIN FOLDING | 0 | 40 |
| 13 | RRNA PROCESSING | 0 | 38 |
| 14 | VACUOLAR PROTEIN TARGETING | 0 | 38 |
| 15 | GLYCOLYSIS | 0 | 36 |
| 16 | MATING | 0 | 34 |
| 17 | SUGAR METABOLISM | 0 | 27 |
| 18 | SPORULATION | 0 | 22 |
| 19 | AMINO ACID METABOLISM | 0 | 21 |
| 20 | AMINO ACID BIOSYNTHESIS | 0 | 19 |
| 21 | PYRIMIDINE BIOSYNTHESIS | 0 | 12 |
| 22 | STRESS RESPONSE | 0 | 12 |
| 23 | METHIONINE BIOSYNTHESIS | 0 | 11 |
| 24 | SALT TOLERANCE | 0 | 8 |
| 25 | GLYCEROL METABOLISM | 0 | 6 |
| 26 | HEAT SHOCK RESPONSE | 0 | 6 |
| 27 | TREHALOSE METABOLISM | 0 | 6 |
| 28 | AMINO ACID METABOLISM | 0 | 5 |
| 29 | B-VITAMIN BIOSYNTHESIS | 0 | 5 |
| 30 | HIGH OSMOLARITY | 0 | 5 |
| 31 | RESPIRATION | 0.0001 | 30 |
| 32 | PHOSPHOLIPID METABOLISM | 0.0001 | 22 |
| 33 | SPHINGOLIPID METABOLISM | 0.0001 | 9 |
| 34 | CHROMATIN STRUCTURE | 0.0003 | 47 |
| 35 | OXIDATIVE STRESS RESPONSE | 0.0003 | 14 |
| 36 | PURINE BIOSYNTHESIS | 0.0006 | 18 |
| 37 | CELL ORGANIZATION | 0.0016 | 76 |
| 38 | MRNA EXPORT | 0.0028 | 9 |
| 39 | RNA PROCESSING | 0.0035 | 9 |
| 40 | TRNA PROCESSING | 0.0042 | 35 |
| 41 | PYRIMIDINE METABOLISM | 0.005 | 8 |
| 42 | SIGNALING | 0.0075 | 58 |
| 43 | DRUG RESISTANCE | 0.0078 | 11 |
| 44 | TOXIN RESISTANCE | 0.0122 | 26 |
| 45 | ENDOCYTOSIS | 0.0152 | 18 |
| 46 | ATP SYNTHESIS | 0.0163 | 20 |
| 47 | PROTEIN TARGETING | 0.017 | 66 |

can be seen that the change in the expression levels in response to environmental stress factors is transient. The average expression levels of experiments that include change in sorbitol level ($k = 2,3$) are similar. However, these levels are different from the first experimental setting, where no sorbitol is present. Repressed path-

Table 10: Analysis of ESR data: Pathways with significant changes over time and in response to sorbitol

| | Change over **time** | | | Change in response to **sorbitol** | |
|---|---|---|---|---|---|
| | Pathway Name | Pathway Size | | Pathway Name | Pathway Size |
| 1 | PROTEIN SYNTHESIS | 286 | 1 | TRANSPORT | 143 |
| 2 | TRANSPORT | 143 | 2 | SECRETION | 126 |
| 3 | SECRETION | 126 | 3 | CELL CYCLE | 97 |
| 4 | LIPID METABOLISM | 63 | 4 | CELL ORGANIZATION | 76 |
| 5 | DNA REPAIR | 58 | 5 | LIPID METABOLISM | 63 |
| 6 | DNA REPLICATION | 57 | 6 | AMINO ACID BIOSYNTHESIS | 60 |
| 7 | RRNA PROCESSING | 38 | 7 | DNA REPLICATION | 57 |
| 8 | GLYCOLYSIS | 36 | 8 | MEIOSIS | 52 |
| 9 | MATING | 34 | 9 | PROTEIN GLYCOSYLATION | 51 |
| 10 | SUGAR METABOLISM | 27 | 10 | PROTEIN FOLDING | 40 |
| 11 | PHOSPHOLIPID METABOLISM | 22 | 11 | RRNA PROCESSING | 38 |
| 12 | AMINO ACID METABOLISM | 21 | 12 | GLYCOLYSIS | 36 |
| 13 | ATP SYNTHESIS | 20 | 13 | TRNA PROCESSING | 35 |
| 14 | ENDOCYTOSIS | 18 | 14 | RESPIRATION | 30 |
| 15 | PURINE BIOSYNTHESIS | 18 | 15 | SUGAR METABOLISM | 27 |
| 16 | OXIDATIVE STRESS RESPONSE | 14 | 16 | TOXIN RESISTANCE | 26 |
| 17 | STRESS RESPONSE | 12 | 17 | PHOSPHOLIPID METABOLISM | 22 |
| 18 | METHIONINE BIOSYNTHESIS | 11 | 18 | SPORULATION | 22 |
| 19 | PYRIMIDINE METABOLISM | 8 | 19 | AMINO ACID BIOSYNTHESIS | 19 |
| 20 | SALT TOLERANCE | 8 | 20 | PURINE BIOSYNTHESIS | 18 |
| 21 | GLYCEROL METABOLISM | 6 | 21 | OXIDATIVE STRESS RESPONSE | 14 |
| 22 | HEAT SHOCK RESPONSE | 6 | 22 | DRUG RESISTANCE | 11 |
| 23 | TREHALOSE METABOLISM | 6 | 23 | MRNA EXPORT | 9 |
| 24 | AMINO ACID METABOLISM | 5 | 24 | RNA PROCESSING | 9 |
| | | | 25 | SPHINGOLIPID METABOLISM | 9 |
| | | | 26 | GLYCEROL METABOLISM | 6 |
| | | | 27 | TREHALOSE METABOLISM | 6 |
| | | | 28 | AMINO ACID METABOLISM | 5 |
| | | | 29 | HIGH OSMOLARITY | 5 |

ways demonstrate a slight delay in the decline in transcription level. Gasch et al. (2000) characterized this as a feature of the second group of genes repressed in the ESR. Figure 7 also reveals that presence of sorbitol further reduces the expression level of genes. This is true for both induced and repressed pathways. It is important to note that, should the experiment included additional samples, more interesting analyses about interactions among heat shock and change of osmolarity would also be possible.

# 5 Conclusion

In this paper, we introduced a modeling framework for incorporating external network information into the analysis of gene sets in complex experiments, including multiple factors and time course data. The framework utilizes mixed linear models and can handle changes in the network structure. Further, it can also be adapted to

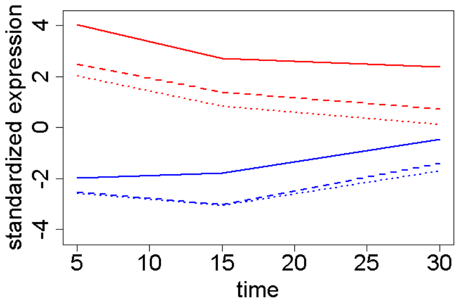

Figure 7: Average expression profile of significant pathways. Red and blue lines represent induced and suppressed pathways, respectively (positive and negative values at the first observation time), and solid, dashed and dotted lines indicate the first, second and third experimental conditions.

handle non-Gaussian data, using the framework of generalized mixed linear models (GLMM).

One of the challenges in analyzing gene expression data using the proposed model is the computational burden of the estimation process. Standard packages for solving mixed linear models cannot handle problems with large vectors/matrices of observations and parameters, without determining a specific independence structure. In this paper, we proposed an iterative algorithm based on block-relaxation for estimating the parameters of the model. This algorithm can be extended to further partition the parameter space and to also partition the set of observations over subnetworks (estimation over subnetworks).

The proposed methodology provides a flexible framework that can be used for studying changes in genetic pathways and allows for systematic inference on such changes as the experimental conditions vary. This model requires external information about the underlying gene network, as well as information on the strength of association between genes. An increasing number of publicly available data sets offer information about the structure of the gene network (the 0-1 adjacency matrix) with different degrees of reliability. However, less information is available about the strength and direction of these connections. An attractive feature of the proposed network-based gene set analysis framework is that it is not sensitive to small noise in the network information. However, bias in the network information can result in significant deviations from the true powers. The problem of estimation of (directed and undirected) gene networks is an important problem in systems biology, and of independent interest. It is however important to note that bias may result from using the same set of gene expression data in order to both estimate the underlying network, and test the significance of pathways.

# Availability

`Matlab` codes for the proposed network-based gene set analysis (NetGSA) in the case of two-class inference problem are available at the first author's website (`http://www.stat.lsa.umich.edu/∼shojaie/`). An R-package (`netGSA`) for the general problem is currently being developed and will be made available through R-CRAN upon completion.

# References

Ashburner, M., C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al. (2000): "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature Genetics*, 25, 25–29.

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

Causton, H., B. Ren, S. Koh, C. Harbison, E. Kanin, E. Jennings, T. Lee, H. True, E. Lander, and R. Young (2001): "Remodeling of Yeast Genome Expression in Response to Environmental Changes," *Molecular Biology Of The Cell*, 12, 323–337.

Chaudhuri, S., M. Drton, and T. Richardson (2007): "Estimation of a covariance matrix with zeros," *Biometrika*, 94, 199–216.

Chung, F. (1997): *Spectral graph theory*, American Mathematical Society.

Dattorro, J. (2005): *Convex optimization & Euclidean distance geometry*, Meboo Publishing USA.

de Leeuw, J. (1994): "Block-relaxation algorithms in statistics," in *Information System and Data Analysis*, Springer-Verlag, 308–325.

Efron, B. and R. Tibshirani (2007): "On testing the significance of sets of genes," *Annals of Applied Statistics*, 1, 107–129.

Friedberg, S. H., A. J. Insel, and L. E. Spence (1996): *Linear Algebra*, Prentice Hall.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown (2000): "Genomic expression programs in the response of yeast cells to environmental changes," *Mol Bio Cell*, 11, 4241–4257.

Gasch, A. P. and M. Werner-Washburne (2002): "The genomics of yeast responses to environmental stress and starvation," *Funct. Integr. Genomics*, 2, 181–192.

Haberman, S. (1989): "Concavity and Estimation," *Annals of Statistics*, 17, 1631–1661.

Harbison, C., D. Gordon, T. Lee, N. Rinaldi, K. Macisaac, T. Danford, N. Hannett, J. Tagne, D. Reynolds, J. Yoo, et al. (2004): "Transcriptional regulatory code of a eukaryotic genome," *Nature*, 431, 99–104.

Hong, F. and H. Li (2006): "Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles," *Biometrics*, 62, 534–544.

Ideker, T., O. Ozier, B. Schwikowski, and A. Siegel (2002): "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, 18, S233–S240.

Ideker, T., V. Thorsson, J. Ranish, R. Christmas, J. Buhler, J. Eng, R. Bumgarner, D. Goodlett, R. Aebersold, and L. Hood (2001): "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." *Science*, 292, 929–34.

Kalisch, M. and P. Bühlmann (2007): "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *The Journal of Machine Learning Research*, 8, 613–636.

Kerr, M. and A. Churchill (2001a): "Statistical design and the analysis of gene expression microarray data," *Genetics Research*, 77, 123–128.

Kerr, M. and G. Churchill (2001b): "Experimental design for gene expression microarrays," *Biostatistics*, 2, 183–201.

Lauritzen, S. (1996): *Graphical models*, Oxford Univ Press.

Lee, I., Z. Li, and E. Marcotte (2007): "An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae," *PLoS ONE*, 2, e988.

DOI: 10.2202/1544-6115.1483

Lindstrom, M. J. and D. M. Bates (1988): "Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data," *Journal of the American Statistical Association*, 83, 1014–1022.

McLean, R. A. and W. L. Sanders (1988): "Approximating degrees of freedom for standard errors in mixed linear models." *Proceedings of the Statistical Computing Section, American Statistical Association*, 50–59.

Mirsky, L. (1975): "A trace inequality of John von Neumann," *Monatshefte für Mathematik*, 79, 303–306.

Oberhofer, W. and J. Kmenta (1974): "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models." *Econometrica*, 42, 579–90.

Rahnenführer, J., F. S. Domingues, J. Maydt, and T. Lengauer (2004): "Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data," *Statistical Applications in Genetics and Molecular Biology*, 3, 1055.

Rue, H. and L. Held (2005): *Gaussian Markov random fields: theory and applications*, Chapman & Hall.

Sanguinetti, G., J. Noirel, and P. Wright (2008): "MMG: a probabilistic tool to identify submodules of metabolic pathways," *Bioinformatics*, 24, 1078–1084.

Searle, S. R. (1971): *Linear Models*, New York: John Wiley & Sons, Inc.

Shojaie, A. and G. Michailidis (2009): "Analysis of Gene Sets Based on the Underlying Regulatory Network," *Journal of Computational Biology*, 16, 407–426.

Shojaie, A. and G. Michailidis (2010): "Penalized Likelihood Methods for Estimation of sparse high dimensional directed acyclic graphs," *Biometrika* (to appear).

Subramanian, A., P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. (2005): "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550.

Tian, L., S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park (2005): "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences*, 102, 13544–13549.

Wei, P. and W. Pan (2008): "Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model," *Bioinformatics*, 24, 404–411.

Wei, Z. and H. Li (2007): "A Markov random field model for network-based analysis of genomic data," *Bioinformatics*, 23, 1537–1544.

Wei, Z. and H. Li (2008): "A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data," *Annals of Applied Statistics*, 2, 408–429.

Yang, Y. and T. Speed (2002): "Design issues for cDNA microarray experiments," *Nature reviews. Genetics(Print)*, 3, 579–588.

Yoneya, T. and H. Mamitsuka (2007): "A hidden Markov model-based approach for identifying timing differences in gene expression under different experimental factors," *Bioinformatics*, 23, 842–849.

DOI: 10.2202/1544-6115.1483