Network-Based Pathway Enrichment Analysis with Incomplete Network Information

Jing Ma\textsuperscript{1}, Ali Shojaie\textsuperscript{2}, and George Michailidis\textsuperscript{1}

\textsuperscript{1}Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
\textsuperscript{2}Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

December 1, 2014

Abstract

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 systems-level view of changes in cellular activity in response to treatments and/or progression of disease states. Methods that use 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 new knowledge data bases containing information about interactions of biomolecules, the existing network information may be incomplete/inaccurate, or is not cell-type or disease condition-specific. To address some of these issues, 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 pathway differences in mean expression levels, as well as interaction mechanisms. We study the asymptotic properties of the proposed network estimator and the test for pathway enrichment, and investigate its small sample performance in simulated experiments and illustrate it on two real cancer data sets.

Keywords: Pathway enrichment analysis; Differential network analysis; High-dimensional data; Graphical lasso.

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 phenotypes. However, in many instances the behavior of biological systems can not be understood based on the activities of their individual members. Thus, to maximize the potential of these profiling technologies, the focus has recently shifted from single entity analysis for identifying differential biomolecules (genes, proteins or metabolites) (Tusher et al., 2001) to techniques that interrogate sets of biomolecules. In particular, analysis of pathways, that comprise functionally related sets of biomolecules, and are believed to be building blocks of cellular functions has received a
lot of attention in recent years. Such methods analyze the coordinated activity of hundreds or thousands of biomolecules at the pathway level, thus addressing the problem at the system level and reducing its complexity. Further, delineation of active pathways enhances explanatory power in detecting subtle, but orchestrated changes in activities of pathway members, which can result in new insights into underlying biological mechanisms of complex diseases.

Hence, pathway enrichment analysis has become an integral part of profiling studies involving high-throughput Omics data. Commonly used variants include over-representation analysis (Al-Shahrour et al., 2005; Beißbarth & Speed, 2004), and gene set enrichment analysis (Subramanian et al., 2005; Efron & Tibshirani, 2007). In both approaches, the pathways are treated as sets of biomolecules and the goal is to assess whether members of a given pathway exhibit higher than expected levels of activity. In short, the idea of over-representation is to determine whether members of a given pathway are over-represented in the set of selected genes, often using a variant of the hypergeometric distribution (Ackermann & Strimmer, 2009). On the other hand, gene set analysis assesses whether members of a given gene set exhibit higher levels of associations with the phenotype of interest (e.g. disease status or different treatment conditions), or in other words, are enriched for activity.

Such methods have been successfully used to both generate new hypotheses, and to provide a framework for determining whether a given gene set is associated with a specific phenotype. Examples include analysis of pathways involved in initiation and progression of cancer and other complex diseases (Cui et al., 2006; Wilson et al., 2010), discovering novel transcriptional effects and co-regulated genes (Palomero et al., 2006; Huarte et al., 2010; Green et al., 2011), and understanding the basic biological processes in model organisms (Gottwein et al., 2007; Baur et al., 2006; Houstis et al., 2006). See Huang et al. (2008) for additional examples of applications. Increased availability of high-throughput proteomics and metabolomics data, has also led researchers to develop variants of over-representation and gene set analysis for analysis of proteomic (e.g. Zhang et al., 2009; Hwang et al., 2010; Isserlin et al., 2010; Cha et al., 2010) and metabolomic data (Lewis et al., 2008; Xu et al., 2009).

The above mentioned methods only consider lists of functionally related biomolecules, but do not explicitly account for interactions amongst them. Nevertheless, an increasing number of public databases contain such information, including Kyoto Encyclopedia of Genes and Genomes (Kanehisa & Goto, 2000), Reactome (Joshi-Tope et al., 2003), RegulonDB (Huerta et al., 1998) and BioCarta (http://www.biocarta.com/). To address this need, Shojaie & Michailidis (2009, 2010) proposed a method, called Network-based Gene Set Analysis, that when testing for pathway activity incorporates the network structure of such interactions, as well as potential changes in the network structure for different experimental conditions. As pointed out in Khatri et al. (2012) topology-based methods such as Network-based Gene Set Analysis, exhibit superior statistical power in identifying differential activity of pathways. Their main limitation is that despite significant progress in identifying and curating biomolecular interactions in databases, the available knowledge is still highly incomplete and occasionally unreliable (see e.g. Zaki et al. (2013) and
Moreover, existing network information often determines molecular interactions in the normal state of the cell, and does not provide any insight into condition/disease-specific alterations in interactions among components of biological systems.

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

The resulting estimates can be incorporated within the Network-based Gene Set Analysis framework, which provides a rigorous statistical framework for assessing alterations in biological pathways, sometimes referred to as differential network biology (Ideker & Krogan, 2012). Moreover, estimation of high dimensional networks subject to hard (or soft) constraints on conditional dependence relationships among random variables represents a canonical problem in the context of graphical models and efficient methods for addressing this problem are of independent interest. Therefore, the first contribution of this paper is the development of an efficient algorithm for constrained network estimation, together with establishing the consistency of the obtained estimates, as a function of available network information.

A second objective of this study is to scale up the Network-based Gene Set Analysis estimation algorithm to very large size networks. The main bottleneck in application of the Network-based Gene Set Analysis methodology arises from the estimation of mixed effects linear parameters – specifically the variance components – for thousands of variables. We develop efficient 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 an efficient algorithm based on Newton’s method, with backtracking line search (Boyd & Vandenberghe, 2004) for step size selection. To supply reliable starting points for this algorithm, we further develop an approximate method-of-moments-type estimator.

This study is strongly motivated by our work on metabolic profiling of cancer and the identification of enriched pathways, briefly discussed next. Note that unlike gene expression data, identification and measurement of metabolites by mass spectrometry techniques is challenging, resulting in reliable measurements for a few hundred metabolites, and hence incomplete coverage of the underlying biochemical pathways. The small number of metabolites in each pathway, and the incomplete coverage of the metabolites particularly hinders the application of over-representation and gene set analysis methods in this setting. In our experience, only topology-based pathway enrichment techniques such as Network-based Gene Set Analysis are capable of reliably estimating pathway activity, as illustrated in Section 5.
2 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 presence of this externally given network information provides a novel and unexplored modification of the corresponding network estimation problem. Specifically, the partial correlation structure underlying a molecular network can be represented by an undirected graph $G = (V, E)$ with $V$ and $E$ being the set of nodes (biomolecules) and edges (interactions), respectively. The edge set $E$ is represented by a $p \times p$ adjacency matrix $A$, whose nonzero elements $A_{ii'} \neq 0$, refer to edges between nodes $i$ and $i'$, and indicate that $i$ and $i'$ are conditionally dependent given all other nodes in the network. Further, the magnitude of the $A_{ii'}$'s determines the strength (positive or negative) of the conditional association between the respective nodes.

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

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

In words, $E_1$ contains known edges, while $E_0$ contains node pairs where it is known that no interaction exists between them. The external information available in $E_1$ does not imply exact knowledge of the magnitude of the interaction $A_{ii'}$. The objective is to estimate the network structure using the framework of Gaussian graphical models, subject to external information encoded in $E_1$ and $E_0$. When $E_1 = E$ and $E_0 = E^c$, the problem becomes that of covariance selection, which has been studied extensively in the literature. However, to the best of our knowledge, the problem of estimating the interactions in $A$ when $E_1$ and $E_0$ only contain partial information has not been investigated before.

Note that this problem is seemingly similar to the problem of matrix completion, although the two problems are fundamentally different in nature. In particular, for matrix completion problems, one is given a subset of entries of an $m \times p$ matrix $Z$, and the goal is to complete the remaining entries from the partially observed matrix, under some structural assumptions on $Z$, such as low-rankness (Candes & Recht, 2009). On the other hand, in the setting of graphical models, the entries of the adjacency matrix are estimated based on observations on the nodes of the graph.

Suppose we observe an $m \times p$ data matrix $Z = (Z_1, \ldots, Z_p)$, where each row represents one sample from a $p$-variate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. We are interested in learning the underlying network and the strength of associations between nodes, as encoded in the inverse covariance matrix $A = \Sigma^{-1}$. In the absence of any external information, the maximum likelihood estimate of $A$ is

$$\min_{A \succ 0} \left\{ \text{tr}(A\hat{\Sigma}) - \log\det A + \lambda \|A\|_1 \right\},$$

(2.1)
where \( \hat{\Sigma} \) is the empirical covariance matrix of the data, \( \|A\|_1 = \sum_{i \neq j'} |A_{ii'}| \) denotes \( \ell_1 \) norm of the parameters, and \( \lambda \) is the regularization parameter. In the presence of external information, the problem can be cast as the following constrained optimization one

\[
\min_{A > 0} \left\{ \text{tr}(A\hat{\Sigma}) - \log\det A \right\},
\]

subject to

\[
\sum_{i \neq i', (i,i') \notin E_0 \cup E_1} |A_{ii'}| \leq t, \quad A_{ii'} = 0, (i,i') \in E_0, \quad A_{ii'} \neq 0, (i,i') \in E_1.
\]

Meinshausen & Bühlmann (2006) introduced neighborhood selection to estimate a high-dimensional graph. For each node \( i \in V \), they considered the optimal prediction of the random variable \( Z_i \) as a linear combination of the remaining variables. The coefficients for optimal prediction \( \theta_i \) are determined by the inverse covariance matrix. Specifically, it holds that \( \theta_{ii'} = -A_{ii'}/A_{ii}, \) for all \( i' \neq i \). The set of nonzero coefficients of \( \theta_i \) is thus the same as the set of nonzero entries in the row vector of \( A_{ii'} (i' \neq i) \), which defines the set of neighbors of node \( i \). Using an \( \ell_1 \)-penalized regression, they estimated the neighborhood for each node and combined the estimates to select the edge set of the corresponding network.

In the following, we present a two-step procedure to solve the above optimization problem. The proposed approach combines the neighborhood selection with penalized maximum likelihood estimation.

Let \( J^1_i \) and \( J^0_i \) denote the set of (potential) neighbors of node \( i \) for which external information is available: \( J^1_i \) is the set of nodes which are known to be in the neighborhood of \( i \), and \( J^0_i \) is the set of nodes which are known to be not connected to \( i \). Let \( Z_{-i} \) denote the submatrix, by removing the \( i \)th column of \( Z \). Assume all columns of \( Z \) are centered and scaled to have norm 1. Denote by \( S^p \) the set of all positive definite matrices of size \( p \times p \) and \( S^p_E = \{ A \in \mathbb{R}^{p \times p} : A_{ii'} = 0, \text{ for all } (i,i') \notin E \text{ where } i \neq i' \} \). The proposed algorithm proceeds in two steps.

1) Estimate the network structure \( \hat{E} \). For every node \( i \),

\[
\hat{\theta}_i = \arg\min_{\theta \in \mathbb{R}^{p-1}} \frac{1}{m} \|Z_i - Z_{-i}\theta\|^2 + 2\lambda \sum_{i' \neq i} t_{ii'} |\theta_{ii'}|,
\]

where weights of the penalty \( t_{ii'} = 0, i' \in J^1_i; t_{ii'} = \infty, i' \in J^0_i \) and \( t_{ii'} = 1 \) elsewhere. An edge \((i,i')\) is estimated if \( \hat{\theta}_{ii} \neq 0 \) or \( \hat{\theta}_{i'i} \neq 0 \).

2) Given the structure \( \hat{E} \), estimate the adjacency matrix \( \hat{A} \) by

\[
\hat{A} = \arg\min_{A \in S^p \cap S^p_E} \left\{ \text{tr}(\hat{\Sigma}A) - \log\det A \right\},
\]

Remark 1. In this algorithm, the first step estimates the coefficients \( \theta_i \) for optimal prediction, such
that penalization respects the external information constraints. In practice, one can adjust the weights $t_{i'i} (i' \neq i)$ to allow for uncertainty in the amount of information available regarding the network of interest. The second step focuses on estimation of the magnitude of nonzero interactions in the adjacency matrix $A$, conditional on the estimated network topology. The optimization problems in both steps are convex and can be solved efficiently using existing software.

The proposed estimator enjoys nice theoretical properties under certain regulatory conditions. Before presenting the main technical result, we introduce some additional notation. Let $\Sigma_0$ be the covariance matrix in the true model and $A_0 = \Sigma_0^{-1}$. For $i = 1, \ldots, p$, let $s^i = \|\theta^i\|_0 - |\mathcal{J}_i|$, where $\|\theta^i\|_0 = \#\{\theta^i : \theta^i_{i'} \neq 0\}$ is the $l_0$ norm. Hence, $s^i$ represents the number of nonzero coordinates after excluding the known ones in each regression model. Write $s = \max_{i=1,\ldots,p} s^i$ and $S = \sum_{i=1}^p \|\theta^i\|_0$. For a subset $J \subset \{1, \ldots, p\}$, let $Z_J$ be the submatrix by removing the columns whose indices are not in $J$. We make the following assumptions.

**Assumption 2.1.** There exists $\phi_1, \phi_2 > 0$ such that

$$0 < \phi_2 \leq \frac{1}{\phi_{\text{max}}(A_0)} = \phi_{\text{min}}(\Sigma_0) \leq \phi_{\text{max}}(\Sigma_0) = \frac{1}{\phi_{\text{min}}(A_0)} \leq \frac{1}{\phi_1} < \infty.$$

And there exists $\varsigma^2 > 0$ such that for all $i$, $\var{Z_i}{Z_{-i}} = 1/A_{0,ii} \geq \varsigma^2$.

**Assumption 2.2.** Let $J$ and $\tilde{J}$ be subsets of $\{1, \ldots, p\}$. Denote $P_J$ the projection matrix onto the column space of $Z_J$ and $I$ the $p \times p$ identity matrix. There exists $\kappa(s, c_0) > 0$ such that

$$\min_{|J| \leq s} \min_{\|\delta_J\|_2 \leq \varsigma \delta_j 1 \leq c_0 \|\delta_j\|_2} \frac{\|(I - P_J)Z\delta\|_2}{\sqrt{m\|\delta\|_2}} \geq \kappa(s, c_0) > 0. \quad (2.4)$$

Assumption 2.2 says that the eigenvalues of the projected matrix $(I - P_J)Z$ on the specified restricted set $\{\delta \in \mathbb{R}^p : |\tilde{J}| \leq s, \|\delta_J\|_2 \leq c_0 \|\delta_j\|_2\}$ are bounded away from 0. This assumption is adapted from the restricted eigenvalue assumptions in Bickel et al. (2009) to allow for presence of external information on the parameters in the subset $J$.

Let $0 \leq r \leq 1$ represent the percentage of available external information, which is defined as $(|E_0| + |E_1|)/\{p(p - 1)/2\}$. Next, we state our main result.

**Theorem 2.3.** Suppose Assumption 2.1 and Assumption 2.2 with $\kappa(2s, 3)$ are satisfied. For constants $c_1 > 4$ and $0 < k_1 < 1$, assume also that

$$16c_1 \{(1 - r)S \log(p - rp)/m\}^{1/2} \leq k_1 \phi_1 \kappa^2(2s, 3), \quad (2.5)$$

where $S$ is the total number of nonzero parameters excluding the diagonal. Consider $\hat{A}$ defined as in
(2.3). Then, with probability at least \(1 - p^2/c_1^2/8\), under appropriately chosen \(\lambda\), we have

\[
\| \hat{A} - A_0 \|_2 \leq \| \hat{A} - A_0 \|_F = O_P\left(\{ S \log(p - r p)/m \}^{1/2}\right) .
\]

(2.6)

Remark 2. Our result indicates that the convergence rate for estimating the inverse covariance matrix depends on the total number of true edges \(S\), the amount of external information \(r\), the dimension \(p\) and the number of observations \(m\). With external knowledge of the network, we get an improvement \(\{ S \log(1 - r)^{-1}/m \}^{1/2}\) in the rate. The proof utilizes techniques from Bickel et al. (2009) and Zhou et al. (2011) and is given in Appendix 1.

The tuning parameter \(\lambda\) in the first step of the proposed algorithm is important in selecting the correct structure of the network, which will further influence the magnitude of the network interactions in the second step. Accurate estimation of these magnitudes are crucial for topol-\-based pathway enrichment methods. We propose to select \(\lambda\) via cross validation to minimize the squared prediction error from all \(p\) regressions. Specifically, the cross validation score for the \(i\)th regression (2.2) is defined as

\[
CV_i(\lambda) = \sum_{j=1}^{m} \{ Z_{ji} - \hat{Z}_{ji} \hat{\theta}(j) \}^2 ,
\]

where \(\hat{\theta}(j)\) is the estimated regression coefficient after removing the \(j\)th sample of \((Z_i, Z_{-i})\). We minimize \(CV(\lambda) = \sum_{i=1}^{p} CV_i(\lambda)\) to select the optimal \(\lambda\).

3 Network-based Gene Set Analysis Modeling and Algorithm

The method of Network-based Gene Set Analysis was originally introduced in Shojaie & Michailidis (2009) and further extended in Shojaie & Michailidis (2010) to accommodate complex experimental designs, as well as more general network structures. To make the presentation self-sufficient, details of the model formulation, estimation and statistical inference issues are provided in the Supplementary Material. Here we focus on the estimation of variance parameters in the corresponding mixed linear model framework and present an updated algorithm that significantly improves computational speed and stability of the method.

Consider \(p\) biomolecules whose activity levels across \(n\) samples are organized in a \(p \times n\) data matrix \(D\). In practice, this set of data can be of a much smaller size compared to the previous data matrix \(Z\), which comes with \(m\) observations, for learning the networks. The Network-based Gene Set Analysis methodology allows for more complex models, including time course observations. For expositional clarity, we present the methodology in the setting of two experimental conditions. Let \(Y_{j}^{(k)} (j = 1, \ldots, n; k = 1, 2)\) be the \(j\)th sample in the expression data under condition \(k\) (\(j\)th column of data matrix \(D\)), with the \(n_1\) columns of \(D\) corresponding to condition 1 (control) and \(n_2 = n - n_1\).
columns to condition 2 (treatment). Denote by $\Lambda^{(k)}$ the influence matrix for condition $k$, which is calculated based on the corresponding adjacency matrix of the network and captures the propagated effect of variables on each other (see Supplementary Material for more details). Also let $\mu^{(k)}$ represent the mean vector under condition $k$. The Network-based Gene Set Analysis framework considers a latent variable model of the form

$$
Y^{(1)}_j = \Lambda^{(1)}\mu^{(1)} + \Lambda^{(1)}\gamma_j + \varepsilon_j, \quad (j = 1, \ldots, n_1),
$$

$$
Y^{(2)}_j = \Lambda^{(2)}\mu^{(2)} + \Lambda^{(2)}\gamma_j + \varepsilon_j, \quad (j = n_1 + 1, \ldots, n).
$$

Here, $\gamma_j$ is the vector of (unknown) random effects, and $\varepsilon_j$ is the vector of random errors. They are independent and normally distributed with mean 0 and variances $\sigma^2_\gamma I_p$ and $\sigma^2_\varepsilon I_p$, respectively.

It is clear that $E(Y^{(k)}_j) = \Lambda^{(k)}\mu^{(k)}$ and $\text{var}(Y^{(k)}_j) = \sigma^2_\gamma \Lambda^{(k)}(\Lambda^{(k)})^T + \sigma^2_\varepsilon I_p$, where $\Lambda^{(k)}(\Lambda^{(k)})^T$ denotes the transpose of matrix $\Lambda^{(k)}$. Let $\mathcal{Y}$, $\mathcal{G}$ and $\mathcal{E}$ represent the rearrangement of data matrix $D$, random effects $(\gamma_j)_{j=1}^n$ and random errors $(\varepsilon_j)_{j=1}^n$ into $np \times 1$ column vectors. Also let $\beta$ be the concatenated vector of means $\mu^{(1)}$ and $\mu^{(2)}$. Using the framework of mixed linear models, we can write

$$
\mathcal{Y} = \Psi \beta + \Pi \mathcal{G} + \mathcal{E}, \quad \mathcal{E} \sim \mathcal{N}_{np}(0, \sigma^2_\varepsilon I_{np}), \quad \mathcal{G} \sim \mathcal{N}_{np}(0, \sigma^2_\gamma I_{np}). \quad (3.1)
$$

Here $\beta$ and $\mathcal{G}$ are fixed and random effect parameters, and $\Psi$ and $\Pi$ are the corresponding design matrices of dimensions $np \times 2p$ and $np \times np$, respectively.

Inference in Network-based Gene Set Analysis requires estimation of the mean parameter $\beta$, which depends on estimates of the variance components $\sigma^2_\gamma$ and $\sigma^2_\varepsilon$. In practice, the variance components can be estimated via maximum likelihood or restricted maximum likelihood, which can be computationally demanding when the underlying network is of large scale. The earlier version of the Network-based Gene Set Analysis considered profiling out one of the variance components and implemented an algorithm from Byrd et al. (1995), which uses a limited-memory modification of the Broyden–Fletcher–Goldfarb–Shanno quasi-Newton method to optimize the profile log-likelihood. However, the above implementation has a few issues. The first issue is its high computational cost due to the inefficient evaluation of matrix inverses and determinants. Moreover, the algorithm from Byrd et al. (1995) requires finite values of the objective function within the supplied box constraints, which is often not satisfied, even after the constraints are adjusted to be within a small range of the optimal estimate. This is particularly the case when the underlying networks are large. To extend the applicability of the Network-based Gene Set Analysis, we consider using Newton’s method for estimating the variance parameters based on the profile log-likelihood to improve both the computational efficiency and stability.

In the following, we discuss computational procedures based on profiling out $\sigma_\varepsilon$ with the restricted maximum likelihood method. Analysis for profiling out $\sigma_\gamma$ or with maximum likelihood estimators follows similarly (See Supplementary Material for more details). Let $\tau = \sigma^2_\gamma / \sigma^2_\varepsilon$. For $k = 1, 2$, write the
covariance matrix of \( Y_j^{(k)} \) as \( \sigma^2 \Sigma^{(k)} = \sigma^2 \{ I_p + \tau \Lambda^{(k)} (\Lambda^{(k)})^T \} \). Denote the residuals \( R_j = Y_j^{(k)} - \Lambda^{(k)} \hat{\mu}_j^{(k)} \) for \( j = 1, \ldots, n \), where \( \hat{\mu}_j^{(k)} \) is the estimate for \( \mu_j^{(k)} \), derived from the estimate of \( \beta \). Let \( N = np \) be the total number of observations for all genes. The nonconstant part of the profile log-likelihood is

\[
p_{R}(\tau \mid Y_1, \ldots, Y_n) = -\frac{1}{2} (n_1 \log \det (\Sigma^{(1)} + n_2 \log \det (\Sigma^{(2)}))
- \frac{1}{2} (N - 2p) \log \left\{ \sum_{j=1}^{n_1} R_j^T (\Sigma^{(1)})^{-1} R_j + \sum_{j=n_1+1}^{n} R_j^T (\Sigma^{(2)})^{-1} R_j \right\}
- \frac{1}{2} \log \det \left\{ n_1 (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)} \right\} - \frac{1}{2} \log \det \left\{ n_2 (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)} \right\}.
\]

(3.2)

To apply Newton’s method, we need the gradient and Hessian matrix of the profile log-likelihood with respect to \( \tau \). Since both the gradient and Hessian matrices depend on \( \Sigma^{(k)} \) and \( \Lambda^{(k)} \), efficient evaluation of matrix inverses and determinants will play an important role in speeding up the estimation of variance parameters. In particular, we made the following two key improvements for implementation of Newton’s method.

First, since \( \Sigma^{(k)} \) \( (k = 1, 2) \) is symmetric and positive definite, we chose to invert from their Cholesky decompositions. For instance, consider the Cholesky decomposition of \( \Sigma^{(k)} = U^T U \), where \( U \) is an upper triangular matrix. The inversion of the triangular matrices results in significant speedup and the inverses of the original matrices can then be computed as \( (\Sigma^{(k)})^{-1} = (U^{-1})(U^{-1})^T \). In the meantime, we also simplified the calculation of determinant of \( \Sigma^{(k)} \) since \( \det (\Sigma^{(k)}) = \det (U)^2 \), which is necessary for evaluating the profile log-likelihood.

Second, quality of the starting point as well as step sizes will both affect convergence of Newton’s method. To select a good starting point, we first used a naive approximation to estimate the variance components. Specifically, consider the residual vectors \( R_j \) \( (j = 1, \ldots, n) \) and assume that there is a single variance \( \sigma^2 \) that applies to all \( \varepsilon_j \) \( (j = 1, \ldots, n) \) and variances of \( \gamma_j \) are different. The variance of \( R_j \) can be decomposed as \( (\sigma^2_{\gamma})_j + \sigma^2_{\varepsilon} \). We then take the minimum of \( \text{var}(R_j) \) as the estimate of \( \sigma^2_{\gamma} \) and average of the remaining variances as the estimate of \( \sigma^2_{\varepsilon} \). Their ratio is used as the initial value for \( \tau \). The approximation runs very fast and does not add much computational cost to the method. To find the appropriate step sizes, we used backtracking line search as described in Boyd & Vandenberghe (2004).

With the above two modifications, Newton’s method is then implemented to optimize the profile log-likelihood and returns an estimate for \( \tau \). As seen in Lindstrom & Bates (1988), the estimate of \( \sigma^2_{\varepsilon} \) can be solved via restricted maximum likelihood as a function of the residuals and \( \Sigma^{(k)} \), which gives

\[
\sigma^2_{\varepsilon} = \frac{1}{N - 2p} \left\{ \sum_{j=1}^{n_1} R_j^T (\Sigma^{(1)})^{-1} R_j + \sum_{j=n_1+1}^{n} R_j^T (\Sigma^{(2)})^{-1} R_j \right\}.
\]

(3.3)
Combining the two estimates, we can derive the estimate of \( \hat{\sigma}_z^2 = \hat{\sigma}_\epsilon^2 \hat{\tau} \).

Finally, to test for pathway enrichment with Network-based Gene Set Analysis, let \( b \) be a row binary vector determining the membership of genes in a pre-specified pathway \( P \). Then Shojaie & Michailidis (2009) show that the contrast vector (Searle, 1971) \( l = b \Lambda \cdot b \) – with \( \cdot \) denoting the Hadamard product – satisfies the constrain \( l^T l = 0 \) and tests the enrichment of pathway \( P \). The advantage of this contrast vector is that it isolates influences from nodes outside the pathways of interest. The null hypothesis of no pathway activity vs the alternative of pathway activation then becomes

\[
H_0 : l \beta = 0, \quad H_1 : l \beta \neq 0. \tag{3.4}
\]

This 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 Network-based Gene Set Analysis is currently the only method that allows for systematic evaluation of effects of these structural changes.

The significance of individual contrast vectors in (3.4) can be tested using the following Wald test statistic

\[
T S = -\frac{l \hat{\beta}}{SE(l \hat{\beta})}, \tag{3.5}
\]

where \( SE(l \hat{\beta}) \) represents the standard error of \( l \hat{\beta} \). Both \( l \) and \( SE(l \hat{\beta}) \) depend on the underlying networks, which are estimated from the data matrix \( Z \). Under the null hypothesis, \( TS \) follows approximately a \( t \)-distribution whose degrees of freedom can be estimated using the Satterthwaite approximation method (see Supplementary Material for details).

As stated in Theorem 2.1 of Shojaie & Michailidis (2010), Network-based Gene Set Analysis is also robust to uncertainty in the network information. Combining their result with the consistency property of our proposed network estimation procedure in Theorem 2.3 of Section 2, we obtain the following corollary.

**Corollary 3.1.** Assume that \( S = o(m/ \log p) \) and consider the adjacency matrices estimated from using equation (2.2) and (2.3) of Section 2. Then the test statistic in (3.5) based on the estimated networks is an asymptotically most powerful unbiased test for (3.4).

## 4 Simulation Results

We present two experiments to demonstrate the performance of the proposed network estimation procedure, as well as its impact on Network-based Gene Set Analysis. We refer readers to the Supplementary
Material for additional simulation scenarios - in particular settings with large number of variable $p$- and discussions.

Our first experiment is based on a undirected network of size $p = 64$. There are 8 subnetworks, each corresponding to a subgraph/pathway of 8 members. Under the null, all subnetworks have the same topology, which was generated from a scale-free random graph, and all nodes have mean expression values 1. To allow for interactions between subnetworks, there is 20% probability for subnetworks to connect to each other. Under the alternative, the proportion of nodes that have mean changes of magnitude 1 is 0%, 40%, 40% and 50% for subnetwork 1–4. The same applies to subnetworks 5–8.

Our second experiment considers a network of size $p = 160$ with a similar design, except that there are 20 members in each subnetwork. Mean expression values for all nodes are the same under the null. Under the alternative, we allow 0%, 40%, 60% and 80% of the nodes to have mean changes of magnitude 0.3 for subnetworks 1–4. Subnetworks 5–8 follow the same pattern. Here an important comparison is to see whether Network-based Gene Set Analysis is able to detect small but coordinated changes in mean expression levels.

In both experiments, we also allowed the structures in subnetworks 5–8 under the alternative to differ from their null equivalent by 10% to simultaneously test pathway enrichment and differential network structure. Fig. 1 shows the slight modification in the topology for subnetworks 5–8, from the null to the alternative hypothesis in the second experiment.

Figure 1: A graph showing the varying structure of pathways 5–8 from null (left) to alternative (right) in Experiment 2. Dashed lines represent edges that are present in only one condition.

To illustrate how external information about the network structure facilitates the estimation, we let the percentage of information $r$ vary from 0 to 1. When $r$ is less than 1, we estimated the adjacency matrices using the proposed two-step procedure and filled in the nonzero edges with the estimated weights. When full knowledge of the network topology is given ($r = 1$), one only needs to apply the second step to estimate the edge weights. Table 1 compares the estimated networks with the true model under several deviance measures based on 200 replications, with a sample size $m = 40$ for experiment 1 and
$m = 100$ for experiment 2. The Matthews correlation coefficient exhibits a clear increasing trend, while the Frobenius norm loss a clear decreasing trend, both indicating the improvement in estimation when the percentage of external information $r$ increases.

Table 1: Deviance measures for network estimation in experiment 1 and 2

| $r$   | FPR(%)  | FNR(%) | MCC  | Fnorm | $p = 64$ | FPR(%)  | FNR(%) | MCC  | Fnorm | $p = 160$ |
|-------|---------|--------|------|-------|----------|---------|--------|------|-------|------------|
| 0.0   | 7.99    | 9.04   | 0.48 | 0.57  | 2.90     | 1.16    | 0.54   | 0.30 |      |            |
| Null  |         |        |      |       |          |         |        |      |       |            |
| 0.2   | 6.68    | 9.76   | 0.52 | 0.55  | 2.29     | 1.53    | 0.59   | 0.28 |      |            |
| 0.8   | 1.74    | 3.91   | 0.79 | 0.38  | 0.55     | 1.07    | 0.83   | 0.19 |      |            |
|       | 0.0     | 8.27   | 8.48 | 0.48  | 0.54     | 2.44    | 0.72   | 0.58 | 0.31 |            |
|       | 0.2     | 6.87   | 8.93 | 0.51  | 0.51     | 1.96    | 0.93   | 0.62 | 0.29 |            |
|       | 0.8     | 1.82   | 0.91 | 0.80  | 0.35     | 0.44    | 1.43   | 0.85 | 0.18 |            |
| Alternative |       |        |      |       |          |         |        |      |       |            |
| 0.2   | 6.87    | 8.93   | 0.51 | 0.51  | 1.96     | 0.93    | 0.62   | 0.29 |      |            |
| 0.8   | 1.82    | 0.91   | 0.80 | 0.35  | 0.44     | 1.43    | 0.85   | 0.18 |      |            |

♯ FPR(%), false positive rate in percentage;
♭ FNR(%), false negative rate in percentage;
† MCC, Matthews correlation coefficient;
‡ Fnorm, Frobenius norm loss.

Next, we evaluated the performance of Network-based Gene Set Analysis in detecting pathway enrichment by comparing it with Gene Set Analysis (Efron & Tibshirani, 2007), which tests either a competitive or self-contained null hypothesis. While a self-contained null hypothesis permutes the samples and compares the gene set in the pathway with itself, a competitive null hypothesis permutes the genes and compares the set of genes in the pathway with a set of genes not in the pathway. Gene Set Analysis recommends using the competitive null approach to take into consideration the distribution of individual gene set scores, which are used to determine the test statistics.

Table 2 presents the estimated powers for each pathway in the two experiments from 200 replicates, given the differences in mean expression levels and/or subnetwork structures described above. Here we used 16 samples for each condition in experiment 1 and 40 in experiment 2, which are different from the datasets used for network estimation. The powers were calculated as the proportion of replicates that show differential changes, based on the false discovery rate controlling procedure in Benjamini & Hochberg (1995) with a $q$-value of 0.05. For Network-based Gene Set Analysis, we looked at scenarios when there is 20% and 80% external structural information, and used the estimated networks to detect enrichment for each pathway. We also included the scenario when 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 pathway were calculated when all unknown parameters were substituted with their corresponding known values. As shown in Table 2, results from Network-based Gene Set Analysis with the exact networks agree with the true powers in both experiments, reflecting low powers for pathways 1 and 2, slightly higher powers for 5 and 6 due to change of pathway topology, high powers for 3 and 4 due to change of mean expression levels and highest powers for pathways 7 and 8 for both changes in mean and structure. When the exact networks
are unknown, we can still see improvement in estimating powers for pathways 3, 7 and 8 in experiment 1, and for pathways 4, 5, and 6 in experiment 2 as the percentage of external information increases from 20% to 80%. Comparing the estimated powers for the same pathway in the two experiments also confirms our hypothesis that Network-based Gene Set Analysis is able to identify large changes in only a few genes of the pathway, as well as weak but coordinated changes in the pathway. In contrast, Gene Set Analysis with competitive null approach fails to identify most of the differentially expressed pathways. Gene Set Analysis with self-contained null hypothesis recognizes mostly correctly the pathways that are significantly differentially expressed, although with lower powers for pathways 3, 4, 7 and 8 in experiment 1 and 6 in experiment 2.

Table 2: Powers based on false discovery rate with $q^* = 0.05$ in experiment 1 and 2

| Pathway | 0.2 | 0.8 | $p = 64$ | 0.2 | 0.8 | $p = 160$ |
|---------|-----|-----|----------|-----|-----|----------|
| 1       | 0.07| 0.09| 0.02     | 0.05| 0.06| 0.00     |
| 2       | 0.14| 0.13| 0.09     | 0.14| 0.18| 0.00     |
| 3       | 0.47| 0.59| 0.66     | 0.62| 0.62| 0.02     |
| 4       | 0.93| 0.92| 0.94     | 0.98| 0.68| 0.22     |
| 5       | 0.26| 0.18| 0.09     | 0.13| 0.10| 0.00     |
| 6       | 0.39| 0.28| 0.35     | 0.47| 0.26| 0.00     |
| 7       | 0.61| 0.65| 0.82     | 0.92| 0.63| 0.02     |
| 8       | 0.78| 0.79| 0.82     | 0.91| 0.51| 0.04     |

0.2/0.8 refer to Network-based Gene Set Analysis with 20%/80% external information;  
$E$ refers to Network-based Gene Set Analysis with the exact networks;  
$T$ refers to the true power;  
GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively.

Finally, to evaluate the computational efficiency of Network-based Gene Set Analysis with the updated algorithm based on Newton’s method, we compared it with the earlier version of Network-based Gene Set Analysis implemented with an algorithm from Byrd et al. (1995). Four different scenarios were considered, including the two experiments described above and another two from the Supplementary Material. The comparison was based on the average elapsed time of Network-based Gene Set Analysis in 100 replicates. All timings were carried out under R version 3.0.2 on a Intel Xeon 2.00 GHz processor. Table 3 presents the results. In general, we see Network-based Gene Set Analysis with the updated algorithm runs significantly faster (two times or more) than the previous implementation. The updated implementation is also more stable in terms of evaluating the profile log-likelihood and its gradient, which is especially important when the underlying network is large. In contrast, the earlier version with the method from Byrd et al. (1995) failed to run successfully for large $p$ because the gradient of the profile log-likelihood was evaluated to be infinite within the supplied box constraints.
Table 3: Timings (in seconds) for Network-based Gene Set Analysis

| p  | Density (%) | (1) Newton’s method | (2) L-BFGS-B | Ratio of (2) to (1) |
|----|-------------|---------------------|--------------|-------------------|
| 64 | 3.42        | 0.221               | 0.424        | 1.919             |
| 160| 1.29        | 1.801               | 7.005        | 3.890             |
| 160| 5.16        | 1.490               | 6.468        | 4.341             |
| 400| 1.13        | 17.102              | NA           | NA                |

\* Density (%) refers to density of studied networks in percentage;  
\* Newton’s method refers to Network-based Gene Set Analysis implemented with Newton’s method;  
\* L-BFGS-B refers to Network-based Gene Set Analysis implemented with the method of Byrd et al. (1995).

5 Applications to Genomics and Metabolomics

In this section, we discuss applications of the proposed Network-based Gene Set Analysis to genomic and metabolomic data to demonstrate its potential in revealing biological insights. The metabolomics data set (Putluri et al., 2011) examines changes in the metabolic profile between 58 cancer and adjacent benign tissue specimens through an untargeted mass spectrometry data acquisition strategy. There are two groups of tissue specimens, with 31 samples from the cancer class and 28 from a benign class. The total number of metabolites detected is 63. Here we focused on estimating the network of metabolic interactions, enhanced by information gleaned from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa & Goto, 2000). To select the estimated networks for both conditions, we performed 5-fold cross validation. We also tested for differential activity of biochemical pathways extracted from the Kyoto Encyclopedia of Genes and Genomes using the same set of data. Shown in Table 4 are estimated \( p \)-values after false discovery rate correction with a \( q \)-value of 0.001 for the significant pathways selected from all methods. Network-based Gene Set Analysis identified the top eight pathways as being significantly differentially active, among which fatty acid biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, and neuroactive ligand-receptor interaction were not identified by Gene Set Analysis with the self-contained null hypothesis. 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 (TCA) which were followed up for biological insights in the original study Putluri et al. (2011). In contrast, the remaining four pathways (pyrimidine metabolism, ABC transporters, histidine metabolism and hsa00220) were only reported as differentially active by Gene Set Analysis with self-contained null. On the other hand, Gene Set Analysis with the competitive null (the recommended setting) failed 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.

For the second application, we consider data from Subramanian et al. (2005), which consists of gene expression profiles of 5217 genes for 62 normal and 24 lung cancer patients. We excluded genes that are
Table 4: p-values for the pathways in the metabolomics data, with false discovery rate correction at $q^* = 0.001$

| Pathway                                      | NetGSA‡ | GSA-s † | GSA-c † |
|----------------------------------------------|---------|---------|---------|
| Fatty acid biosynthesis                      | < 0.001 | 1.000   | 1.000   |
| Aminoacyl-tRNA biosynthesis                  | < 0.001 | < 0.001 | 0.458   |
| Tryptophan metabolism                        | < 0.001 | < 0.001 | 0.338   |
| Pantothenate and CoA biosynthesis            | < 0.001 | < 0.001 | 0.395   |
| Phenylalanine, tyrosine and tryptophan biosynthesis | < 0.001 | 1.000   | 1.000   |
| beta-Alanine metabolism                      | < 0.001 | < 0.001 | 0.338   |
| Neuroactive ligand-receptor interaction       | < 0.001 | 0.006   | 0.338   |
| Phenylalanine metabolism                     | < 0.001 | < 0.001 | 0.542   |
| Pyrimidine metabolism                        | 0.003   | < 0.001 | 0.395   |
| ABC transporters                             | 0.005   | < 0.001 | 0.624   |
| Histidine metabolism                         | 0.029   | < 0.001 | 0.111   |
| hsa00220                                     | 0.146   | < 0.001 | 0.672   |

‡ NetGSA refers to Network-based Gene Set Analysis;
† GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

not present in the 186 pathways from the Kyoto Encyclopedia of Genes and Genomes data base as well as those which do not have recorded network information, which leaves us with 1416 genes. We then performed 5-fold cross validation to estimate the underlying interaction networks for both normal and lung cancer conditions based on the external topology information from the BioGRID Database.

To test for pathway enrichment, we considered a subset of pathways from the Kyoto Encyclopedia data base that describe signalling and biochemical mechanisms and restricted their membership to be at least 5, so that Gene Set Analysis could be applicable. This reduces the number of pathways tested to 61. Table 5 presents the p-values for the significant pathways identified from all three methods based on false discovery rate correction at 0.001, sorted with respect to results from using Network-based Gene Set Analysis. It turns out that Gene Set Analysis does not consider any of the pathways as differentially active, whichever null hypothesis is used. In comparison, the small p-values from Network-based Gene Set Analysis suggest these 11 pathways could be of interest for further investigation. Of particular biological interest is the identification of the TGF-beta signaling pathway, that has been linked to biological mechanisms for onset and progression of lung cancer (see e.g. Ischenko et al. (2014) and references therein).

A Appendix

Theoretical Analysis and Proofs

To prove our main result, we need some additional notations. Define $\tilde{A}_0 = \text{diag}(A_0) + A_{0,E \cap \hat{E}}$, where $E$ and $\hat{E}$ are the true and the estimated edge set, respectively. By definition, $\tilde{A}_0$ and $A_0$ will be different at position $(i, i')$ only when the edge $(i, i')$ is falsely rejected. We first derive an upper bound for the
Table 5: *p*-values for the pathways in the microarray data, with false discovery rate correction at *q*∗ = 0.001

| Pathway                               | NetGSA  |
|---------------------------------------|---------|
| Glycerophospholipid metabolism        | < 0.001 |
| PPAR signaling pathway                | < 0.001 |
| Glycine, serine and threonine metabolism | < 0.001 |
| Cysteine and methionine metabolism    | < 0.001 |
| Glycerolipid metabolism               | < 0.001 |
| TGF-beta signaling pathway            | < 0.001 |
| Fructose and mannose metabolism       | < 0.001 |
| Neurotrophin signaling pathway        | < 0.001 |
| Phosphatidylinositol signaling system | < 0.001 |
| ErbB signaling pathway                | < 0.001 |

| Pathway                               | GSA-s   | GSA-c   |
|---------------------------------------|---------|---------|
| Glycerophospholipid metabolism        | 0.338   | 0.389   |
| PPAR signaling pathway                | 1.000   |         |
| Glycine, serine and threonine metabolism | 0.137   | 0.357   |
| Cysteine and methionine metabolism    | 0.229   | 0.357   |
| Glycerolipid metabolism               | 0.404   | 0.520   |
| TGF-beta signaling pathway            | 0.404   | 0.518   |
| Fructose and mannose metabolism       | 0.308   | 0.408   |
| Neurotrophin signaling pathway        | 0.308   | 0.588   |
| Phosphatidylinositol signaling system | 0.404   | 0.379   |
| ErbB signaling pathway                | 0.035   | 0.249   |
| mTOR signaling pathway                | 0.138   | 0.357   |

NS NetGSA refers to Network-based Gene Set Analysis; † GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

size of *E* and ∥*A*0 − *A*0∥F. To do this, we show that the regression problem (2.2) is essentially a lasso problem, and then invoke the oracle inequalities from Theorem 7.2 of Bickel et al. (2009). To simplify the notation, we drop the superscript *i* for sets *J*0, *J*1 in the *i*th regression, but they should be understood as *J*0,*i*, *J*1,*i*, respectively.

Let *J* = *V* \ {1} ∪ *J*0 ∪ *J*1 represent the set of indices for which there is no information available. Denote *P* *J*1 = *Z* *J*1(*I* − *P* *J*1)−1 *Z* *J*1, the projection onto the column space of *Z* *J*1. The following lemma is needed in the proof of Theorem A.2 below.

**Lemma A.1.** For *i* = 1, ..., *p*, denote ξ̂ = *Z* *i* − ∑ *i′* ≠ *i* θ̂ *i*′ *Z* *i*′, where θ̂ *i* is the optimal prediction coefficient vector in the *i*th regression. Consider the event

\[ A_i = \left\{ \| Z_j^T (I - P_{J\hat{i}}) \xi / m \|_\infty \leq c_1 \frac{\log(p - r p)}{mA_{0,ii}} \right\}^{1/2} \]

with a constant *c*1 > 4, where *A*0,ii is the *i*th diagonal element of the true inverse covariance matrix *A*0. Define the event *A* = ∩ *i=1*^*p* *A* *i*. Then *pr*(*A*) > 1 − *p*2−*c*_1^2/8.

The proof of Lemma A.1 will be provided shortly. Denote *Λ*max the maximal eigenvalue of *Z* *j* *Z* / *m*. Conditional on event *A*, we have the following results on controlling the size of *E* and the Frobenious norm of the deviance, ∥*A*0 − *A*0∥F.

**Theorem A.2.** Suppose all conditions in Theorem 2.3 are satisfied. Then on event *A*, for appropriately chosen *λ*, we have

\[ |E| \leq \frac{64\Lambda_{\text{max}}}{\kappa^2(s, 3)} (1 - r)S + rS, \] (A.1)

and

\[ \| \hat{A}_0 - A_0 \|_F \leq c_3 \left\{ S \log(p - r p) / m \right\}^{1/2} \leq k_1 \phi_1, \] (A.2)
where \( c_3 = 16c_1 \sqrt{(1 - r)/\kappa^2(2s, 3)} \).

**Remark 3.** The result indicates that the cardinality of the estimated edge set is upper bounded by a function of \( r \), the percentage of the external information. The bound for \( |\hat{E}| \) also depends on the restricted eigenvalue \( \kappa(s, 3) \), which is necessarily positive by the assumption that \( \kappa(2s, 3) > 0 \). Two extreme cases occur when (i) \( r = 0 \), i.e. we do not observe any information, thus reducing problem (2.2) to the original neighborhood selection in Meinshausen & Buhlmann (2006); (ii) \( r = 1 \), i.e. the exact network topology is known and hence \( \hat{E} = E \). On the other hand, the upper bound for \( \| \tilde{A}_0 - A_0 \|_F \) decreases as \( r \) increases, i.e. when more external information becomes available. However, since the coefficients also need to be estimated, this deviance always stays positive, even when \( r = 1 \).

**Proof (of Theorem A.2).** Recall that \( P_{J_i} \) is the projection matrix onto the column space of \( Z_{J_i} \). Let \( \tilde{Y} = (I - P_{J_i})Z_i \) be the projection of \( Z_i \) onto the orthogonal space of \( Z_{J_i} \) and \( \tilde{Z} = (I - P_{J_i})Z _J \). With some algebra, the problem (2.2) is equivalent to solving

\[
\min_{\theta_{J_i}} \frac{1}{m} \| \tilde{Y} - \tilde{Z}\theta_{J_i} \|^2 + 2\lambda \| \theta_{J_i} \|_1, \tag{A.3}
\]

which is a lasso problem. It suffices to focus mainly on the set \( \tilde{J} \), as false positive and negative errors will only occur on this set.

To apply Theorem 7.2 of Bickel et al. (2009), we also need to bound the maximum eigenvalue of the matrix \( \tilde{Z}^T \tilde{Z} / m \). Consider the eigendecomposition of the projection \( I - P_{J_i} = U D U^T \), where \( D \) is the diagonal matrix composed of eigenvalues and \( U \) is orthogonal. As \( I - P_{J_i} \) is also a projection matrix, the diagonals of \( D \) are either 0 or 1. It then follows that

\[
\phi_{\text{max}}(\tilde{Z}^T \tilde{Z} / m) = \phi_{\text{max}}(Z_{J_i}^T U D U^T Z_{J_i} / m) \leq \phi_{\text{max}}(Z_{J_i}^T U U^T Z_{J_i} / m) \leq \phi_{\text{max}}(Z_{J_i}^T Z_{J_i} / m) \leq \Lambda_{\text{max}}.
\]

Recall \( s^i \) is the number of nonzero coordinates after excluding the known ones in each regression and \( s = \max_i s^i \). Under the assumption that \( \kappa(2s, 3) > 0 \), we also have that \( \kappa(s^i, 3) \geq \kappa(s, 3) > 0 \) for \( s^i \leq s \). Let \( \hat{\theta}_{J_i} \) be the lasso estimator in (A.3) with

\[
\lambda = c_1 \left\{ \frac{\log(p - rp)}{mA_{0,ii}} \right\}^{1/2}, \tag{A.4}
\]

for \( c_1 > 4 \). Conditioned on event \( \mathcal{A} \), we can invoke Theorem 7.2 of Bickel et al. (2009) and obtain simultaneously for all \( i \),

\[
\| \hat{\theta}^i \|_0 \leq \frac{64\Lambda_{\text{max}}}{\kappa^2(s, 3)} s^i, \tag{A.5}
\]
and

\[
\|\hat{\theta}_j - \theta_j\|_2 \leq \frac{16c_1}{A_{0,ii}^2(2s,3)} \{s^t \log(p - rp)/m\}^{1/2}. \tag{A.6}
\]

Combining (A.5) with the number of known edges \(s_1^i\) as given in \(J_1^i\), we get

\[
|\tilde{E}| \leq \sum_{i=1}^p \{\|\hat{\theta}_j\|_0 + |J_1^i|\} \leq \frac{64A_{\max}^2}{\kappa^2(s,3)} \sum_{i=1}^p s_1^i + \sum_{i=1}^p s_1^i.
\]

The upper bound in (A.1) follows immediately, since by definition the number of known and unknown edges are \(\sum_{i=1}^p s_1^i = rS\) and \(\sum_{i=1}^p s^i = (1 - r)S\), respectively.

To control \(\|\hat{A}_0 - A_0\|_F\), recall that for every \(i\), \(A_{0,ii'} = -\theta_{i'} A_{0,ii}\). Using the bound in (A.6), we have

\[
\|\hat{A}_0 - A_0\|_F^2 = \sum_{i=1}^p \sum_{i' \in J(\hat{\theta}) \cap J(\theta)} (\theta_{i'}^i A_{0,ii})^2 = \sum_{i=1}^p A_{0,ii}^2 \sum_{i' \in J(\hat{\theta}) \cap J(\theta)} |\theta_{i'}^i - \hat{\theta}_{i'}^i|^2
\]

\[
\leq \sum_{i=1}^p A_{0,ii}^2 \|\theta_{ii} - \hat{\theta}_{ii}\|_2^2 \leq \left\{\frac{16c_1}{\kappa^2(2s,3)}\right\}^2 \frac{(1 - r)S \log(p - rp)}{m}.
\]

The last inequality in (A.2) follows from condition (2.5) in Theorem 2.3. \(\square\)

**Proof (of Lemma A.1).** For every \(i\), since \(Z_i\) is normally distributed as \(\mathcal{N}(0, \Sigma_0)\), it is easy to verify that \(\xi^i\) is also a normal random variable with mean 0 and variance \(1/A_{0,ii}\). Define random variables

\[
Y_{ii'} = (A_{0,ii}/m)^{1/2} Z_{i'}^T \xi^i \text{ for } i \neq i'.
\]

Then, \(Z_{i'}^T Z_{i'}/m = 1\) implies that \(Y_{ii'} \sim \mathcal{N}(0, 1)\). Let \(\lambda\) be defined as in (A.4). Using the fact that \(|Z_{i'}^T(I - P_\lambda)\xi^i/m|\) is stochastically smaller than \(|Z_{i'}^T \xi^i/m|\) for all \(i' \in \tilde{J}\) and an elementary bound on the tails of Gaussian distributions

\[
\text{pr}(\mathcal{A}) \leq \sum_{i=1}^p \sum_{i' \in J} \text{pr}\left(\{|Z_{i'}^T(I - P_\lambda)\xi^i/m| > \lambda/2\}\right)
\]

\[
\leq \sum_{i=1}^p \sum_{i' \in J} \text{pr}\left(|Y_{ii'}| > (mA_{0,ii})^{1/2} \lambda/2\right) \leq \sum_{i=1}^p \sum_{i' \in J} \exp\left\{-m A_{0,ii} \lambda^2 / 8\right\}
\]

\[
\leq p(p - rp) \exp\left\{-c_1^2 \log(p - rp)/8\right\} \leq p^2 - c_1^2/8.
\]

Therefore, \(\text{pr}(\mathcal{A}) > 1 - p^2 - c_1^2/8\). \(\square\)

With Lemma A.1 and Theorem A.2, we are ready to prove our main results in Theorem 2.3. The following proof is adapted from Zhou et al. (2011).

**Proof (of Theorem 2.3).** Consider \(\hat{A}\) defined in (2.3). It suffices to show that

\[
\|\hat{A} - \hat{A}_0\|_F = \mathcal{O}_p\left(\left\{S \log(p - rp)/m\right\}^{1/2}\right),
\]

18
since by triangle inequality and Theorem A.2, we can conclude
\[
\|\hat{A} - A_0\|_F \leq \|\hat{A} - \hat{A}_0\|_F + \|\hat{A}_0 - A_0\|_F \leq O_P \left( \{S \log(p - rp)/m\}^{1/2} \right).
\]

Denote \( \Sigma_0 = \hat{A}_0^{-1} \), which is positive definite since by Theorem A.2,
\[
\phi_{\min}(\hat{A}_0) \geq \phi_{\min}(A_0) - \|\hat{A}_0 - A_0\|_2 \geq \phi_{\min}(A_0) - \|\hat{A}_0 - A_0\|_F \geq \phi_1 - k_1 \phi_1 > 0.
\]  (A.7)

Given \( \hat{A}_0 \in S^p_+ \cap S^p_\hat{E} \), define a new convex set:
\[
U_m(\hat{A}_0) = \{B - \hat{A}_0 \mid B \in S^p_+ \cap S^p_\hat{E}\} \subset S^p_\hat{E}.
\]

Let
\[
Q(A) = \trace(A\hat{\Sigma}) - \trace(\hat{A}_0\hat{\Sigma}) - \log\det A + \log\det \hat{A}_0.
\]

Since the estimate \( \hat{A} \) minimizes \( Q(A) \), \( \hat{A} = \hat{A} - \hat{A}_0 \) minimizes \( G(\Delta) = Q(\Delta + \hat{A}_0) \).

The main idea of this proof is as follows. For a sufficiently large \( M > 0 \), consider sets
\[
T_1 = \{\Delta \in U_m(\hat{A}_0), \|\Delta\|_F = Mr_m\}, \quad T_2 = \{\Delta \in U_m(\hat{A}_0), \|\Delta\|_F \leq Mr_m\},
\]
where
\[
r_m = \{S \log(p - rp)/m\}^{1/2}.
\]

Note that \( T_1 \) is non-empty. Indeed, consider \( B_\epsilon = \epsilon \hat{A}_0 \) for \( \epsilon = Mr_m/\|\hat{A}_0\|_F \). Then \( B_\epsilon = (1 + \epsilon)\hat{A}_0 - \hat{A}_0 \in U_m(\hat{A}_0) \), hence \( B_\epsilon \in T_1 \). Denote by \( \bar{0} \) the matrix of all zero entries. It is clear that \( G(\Delta) \) is convex, and \( G(\hat{\Delta}) \leq G(\bar{0}) = Q(\hat{A}_0) = 0 \). Thus if we can show that \( G(\Delta) > 0 \) for all \( \Delta \in T_1 \), the minimizer \( \hat{\Delta} \) must be inside \( T_2 \) and hence \( \|\hat{\Delta}\|_F \leq Mr_m \). To see this, note that the convexity of \( Q(A) \) implies that
\[
\inf_{\|\Delta\|_F = Mr_m} Q(\hat{A}_0 + \Delta) > Q(\hat{A}_0) = 0.
\]

There exists therefore a local minimizer in the ball \( \{\hat{A}_0 + \Delta : \|\Delta\|_F \leq Mr_m\} \), or equivalently, for \( \hat{\Delta} \in T_2 \), i.e. \( \|\hat{\Delta}\|_F \leq Mr_m \).

In the remainder of the proof, we focus on
\[
G(\Delta) = Q(\Delta + \hat{A}_0) = \trace(\Delta\hat{\Sigma}) - \log\det(\Delta + \hat{A}_0) + \log\det \hat{A}_0.
\]  (A.8)
Applying a Taylor expansion to $\log\det(\tilde{A}_0 + \Delta)$ in (A.8) gives

\[
\log\det(\tilde{A}_0 + \Delta) - \log\det \tilde{A}_0 = \frac{d}{dt}\log\det(\tilde{A}_0 + t\Delta)|_{t=0} + \int_0^1 (1-t)\frac{d^2}{dt^2}\log\det(\tilde{A}_0 + t\Delta)dt
\]

\[
= \text{tr}(\Delta \tilde{\Sigma}_0) - \text{vec}(\Delta)^T \left\{ \int_0^1 (1-t)(\tilde{A}_0 + t\Delta)^{-1} \otimes (\tilde{A}_0 + t\Delta)^{-1} dt \right\} \text{vec}(\Delta), \quad (A.9)
\]

where $\text{vec}(\Delta)$ denotes the vectorized $\Delta$, and $\otimes$ is the Kronecker product. For $\Delta \in T_1$, let $K_1$ be the integral term in (A.9), and define

\[
K_2 = \text{tr}\left\{ \Delta(\tilde{\Sigma} - \Sigma_0) \right\}, \quad K_3 = \text{tr}\left\{ \Delta(\tilde{\Sigma}_0 - \Sigma_0) \right\}.
\]

We can then write

\[
G(\Delta) = K_1 + \text{tr}(\Delta \tilde{\Sigma}) - \text{tr}(\Delta \tilde{\Sigma}_0) = K_1 + K_2 - K_3.
\]

Next, we bound each of the terms $K_1, K_2$ and $K_3$ to find a lower bound for $G(\Delta)$.

First consider $K_2$. Since the diagonal elements of $\tilde{\Sigma}$ and $\Sigma_0$ are the same after scaling,

\[
|K_2| \leq \left| \sum_{i \neq i'} (\tilde{\Sigma}_{ii'} - \Sigma_0,ii') \Delta_{ii'} \right|.
\]

By Lemma A.3 of Bickel & Levina (2008), there exists a positive constant $c_2$ depending on $\phi_{\max}(\Sigma_0)$ such that

\[
\max_{i \neq i'} |\tilde{\Sigma}_{ii'} - \Sigma_0,ii'| \leq c_2 \{\log(p - rp)/m\}^{1/2},
\]

with probability tending to 1. Let $\Delta^+ = \text{diag}(\Delta)$ be the matrix of diagonal elements of $\Delta$, and write $\Delta^- = \Delta - \Delta^+$. Then, $K_2$ is bounded by

\[
|K_2| \leq c_2 \{\log(p - rp)/m\}^{1/2} \|\Delta^-\|_1. \quad (A.10)
\]

For $K_3$, we can use the upper bound for $\|\tilde{A}_0 - A_0\|_F$ in (A.2), and the lower bound for $\phi_{\min}(\tilde{A}_0)$ in (A.7), to write,

\[
|K_3| \leq \|\Delta\|_F \|\tilde{\Sigma}_0 - \Sigma_0\|_F \leq \|\Delta\|_F \frac{\|\tilde{A}_0 - A_0\|_F}{\phi_{\min}(\tilde{A}_0)\phi_{\min}(A_0)} \quad (A.11)
\]

\[
\leq \|\Delta\|_F \frac{c_3 \{S \log(p - rp)/m\}^{1/2}}{(1 - k_1)\phi_1^2}. \quad (A.12)
\]
The second inequality in (A.11) comes from the rotation invariant property of Frobenius norm, i.e.

\[ \| \tilde{\Sigma}_0 - \Sigma_0 \|_F = \| \Sigma_0 (A_0 - \tilde{A}_0) \tilde{\Sigma}_0 \|_F \leq \phi_{\text{max}}(\Sigma_0) \| A_0 - \tilde{A}_0 \|_F \phi_{\text{max}}(\tilde{\Sigma}_0). \]

Using (A.2), we can also obtain an upper bound for the maximum eigenvalue of \( \tilde{A}_0 \):

\[ \phi_{\text{max}}(\tilde{A}_0) \leq \phi_{\text{max}}(A_0) + \| \tilde{A}_0 - A_0 \|_2 \leq \phi_{\text{max}}(A_0) + \| \tilde{A}_0 - A_0 \|_F \leq \frac{1}{\phi_2} + k_1 \phi_1. \]

Since \( r_m \rightarrow 0 \), there exists a sufficiently large \( k_2 > 0 \) such that for \( \Delta \in T_1 \),

\[ \| \Delta \|_2 \leq \| \Delta \|_F = M r_m < \frac{1}{\phi_2} k_2. \]

Following Rothman et al. (2008) [see Page 502, proof of Theorem 1], a lower bound for \( K_1 \) can be found as

\[ K_1 \geq \frac{\phi_2^2}{2(1 + k_1 + k_2)^2} \| \Delta \|_F^2 / \{ 2(\phi_{\text{max}}(\tilde{A}_0) + \| \Delta \|_2)^2 \} \]

\[ \geq \frac{\phi_2^2}{2(1 + k_1 + k_2)^2} \| \Delta \|_F^2 / \{ 2(1/\phi_2 + k_1 \phi_1 + k_2/\phi_2)^2 \} = \frac{\phi_2^2}{2(1 + k_1 + k_2)^2} \Delta F. \]

Combining (A.10), (A.12) and (A.13),

\[ G(\Delta) \geq \frac{\phi_2^2}{2(1 + k_1 + k_2)^2} \| \Delta \|_F^2 - c_2 \{ \log(p - r p) / m \}^{1/2} \| \Delta^- \|_1 - c_3 \{ S \log(p - r p) / m \}^{1/2} (1 - k_1) \phi_1^2 \| \Delta \|_F. \]

For \( \Delta \in T_1 \), applying Cauchy-Schwarz inequality yields

\[ \| \Delta^- \|_1 \leq (|E|)^{1/2} \| \Delta^- \|_F. \]

We thus have

\[ G(\Delta) \geq \frac{\phi_2^2}{2(1 + k_1 + k_2)^2} \| \Delta \|_F^2 - c_2 \{ \log(p - r p) / m \}^{1/2} \| \Delta^- \|_F - \frac{c_3}{(1 - k_1) \phi_1^2} \{ S \log(p - r p) / m \}^{1/2} \| \Delta \|_F \]

\[ \geq \| \Delta \|_F^2 \left\{ \frac{\phi_2^2}{2(1 + k_1 + k_2)^2 - \frac{c_2}{M} \{ |E| / S \}^{1/2} - \frac{c_3}{M(1 - k_1) \phi_1^2} } \right\} > 0, \]

for \( M \) sufficiently large. \( \square \)
References

ACKERMANN, M. & STRIMMER, K. (2009). A general modular framework for gene set enrichment analysis. *BMC bioinformatics* **10**, 47.

AL-SHAHROUR, F., DÍAUS-URIARTE, R. & DOPAZO, J. (2005). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* **21**, 2988–2993.

BAUR, J. A., PEARSON, K. J., PRICE, N. L., JAMIESON, H. A., LERIN, C., KALRA, A., PRABHU, V. V., ALLARD, J. S., LOPEZ-LLUCH, G., LEWIS, K. et al. (2006). Resveratrol improves health and survival of mice on a high-calorie diet. *Nature* **444**, 337–342.

BEISSBARTH, T. & SPEED, T. P. (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

BICKEL, P. J. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

BOYD, S. & VANDENBERGHE, L. (2004). *Convex optimization*. Cambridge University Press.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*. Springer.

BYRD, R. H., LU, P., NOCEDAL, J. & ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Candes, E. J. & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9**, 717–772.

CHA, S., IMIELINSKI, M. B., REJTA, T., RICHARDSON, E. A., THAKUR, D., SGROI, D. C. & KARGER, B. L. (2010). In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Molecular & Cellular Proteomics* **9**, 2529–2544.

CHUANG, H.-Y., RASSENTI, L., SALCEDO, M., LICON, K., KOHLMANN, A., HAERLACH, T., FOÀ, R., IDEKER, T. & KIPPS, T. J. (2012). Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood* **120**, 2639–2649.
Cui, L., Jeong, H., Borovecki, F., Parkhurst, C. N., Tanese, N. & Kainc, D. (2006). Transcriptional repression of pgc-1α by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration. *Cell* **127**, 59–69.

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

Gottwein, E., Mukherjee, N., Sachse, C., Frenzel, C., Majoros, W. H., Chi, J.-T. A., Braich, R., Manoharan, M., Soutschek, J., Ohler, U. et al. (2007). A viral microRNA functions as an orthologue of cellular mir-155. *Nature* **450**, 1096–1099.

Green, M. R., Monti, S., Dalla-Favera, R., Pasqualucci, L., Walsh, N. C., Schmidt-Supprian, M., Kutok, J. L., Rodig, S. J., Neuberg, D. S., Rajewsky, K. et al. (2011). Signatures of murine B-cell development implicate YY1 as a regulator of the germinal center-specific program. *Proceedings of the National Academy of Sciences* **108**, 2873–2878.

Harpison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J. et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104.

Houstis, N., Rosen, E. D. & Lander, E. S. (2006). Reactive oxygen species have a causal role in multiple forms of insulin resistance. *Nature* **440**, 944–948.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols* **4**, 44–57.

Huart, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M. et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409–419.

Huerta, A. M., Salgado, H., Thieffry, D. & Collado-Vides, J. (1998). Regulodb: A database on transcriptional regulation in escherichia coli. *Nucleic Acids Research* **26**, 55–59.

Hwang, H., Bowen, B. P., Lefort, N., Flynn, C. R., De Filippis, E. A., Roberts, C., Smoke, C. C., Meyer, C., Højlund, K., Yi, Z. et al. (2010). Proteomics analysis of human skeletal muscle reveals novel abnormalities in obesity and type 2 diabetes. *Diabetes* **59**, 33–42.

Ideker, T., Dutkowski, J. & Hood, L. (2011). Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**, 860–863.

Ideker, T. & Krogan, N. J. (2012). Differential network biology. *Molecular systems biology* **8**.
ISCHENKO, I., LIU, J., PETRENKO, O. & HAYMAN, M. (2014). Transforming growth factor-beta signaling network regulates plasticity and lineage commitment of lung cancer cells. *Cell Death & Differentiation*.

ISSERLIN, R., MERICO, D., ALIKHANI-KOUPAEI, R., GRAMOLINI, A., BADER, G. D. & EMILI, A. (2010). Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* **10**, 1316–1327.

JOSHI-TOPE, G., VASTRIK, I., GOPINATH, G., MATTHEWS, L., SCHMIDT, E., GILLESPIE, M., D’EUSTACHIO, P., JASSAL, B., LEWIS, S., WU, G., BIRNEY, E. & STEIN, L. (2003). The genome knowledgebase: A resource for biologists and bioinformaticists. *Cold Spring Harbor Symposia on Quantitative Biology* **68**, 237–244.

KANEHISA, M. & GOTO, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30.

KHATRI, P., SIROTA, M. & BUTTE, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* **8**, e1002375.

LEWIS, G. D., ASNANI, A. & GERSZTEN, R. E. (2008). Application of metabolomics to cardiovascular biomarker and pathway discovery. *Journal of the American College of Cardiology* **52**, 117–123.

LINDSTROM, M. J. & BATES, D. M. (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. & SANDERS, W. L. (1988). Approximating degrees of freedom for standard errors in mixed linear models. In *Proceedings of the Statistical Computing Section, American Statistical Association*.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

PALOMERO, T., LIM, W. K., ODOM, D. T., SULIS, M. L., REAL, P. J., MARGOLIN, A., BARNES, K. C., O’NEIL, J., NEUBERG, D., WENG, A. P. et al. (2006). Notch1 directly regulates c-myc and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proceedings of the National Academy of Sciences* **103**, 18261–18266.

PUTLURI, N., SHOJAIE, A., VASU, V. T., VAREED, S. K., NALLURI, S., PUTLURI, V., THANGJAM, G. S., PANZITTI, K., TALLMAN, C. T., BUTLER, C. et al. (2011). Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression. *Cancer Research* **71**, 7376–7386.
ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

SEARLE, S. (1971). Linear models.

SHOJAIE, A. & MICHAILEDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology* **16**, 407–426.

SHOJAIE, A. & MICHAILEDIS, G. (2010). Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology* **9**.

SUBRAMANIAN, A., TAMAYO, P., Mootha, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550.

TUSHER, V. G., TIBSHIRANI, R. & CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

WILSON, B. G., WANG, X., SHEN, X., MCKENNA, E. S., LEMIEUX, M. E., CHO, Y.-J., KOELLHOFER, E. C., POMEROY, S. L., ORKIN, S. H. & ROBERTS, C. W. (2010). Epigenetic antagonism between polycomb and swi/snf complexes during oncogenic transformation. *Cancer cell* **18**, 316–328.

XU, E. Y., SCHAEFER, W. H. & XU, Q. (2009). Metabolomics in pharmaceutical research and development: metabolites, mechanisms and pathways. *Current opinion in drug discovery & development* **12**, 40–52.

ZAKI, N., EFIMOV, D. & BERENGUERES, J. (2013). Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinformatics* **14**, 163.

ZHANG, D. Y., YE, F., GAO, L., LIU, X., ZHAO, X., CHE, Y., WANG, H., WANG, L., WU, J., SONG, D. et al. (2009). Proteomics, pathway array and signaling network-based medicine in cancer. *Cell Div* **4**, 20.

Zhou, S., Rutimann, P., Xu, M. & Bühlmann, P. (2011). High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research* **12**, 2975–3026.
B Supplementary Material

Network-based Gene Set Analysis Modeling and Inference

To make the presentation self-sufficient, we provide next a short overview of the Network-based Gene Set Analysis methodology, originally introduced in Shojaie & Michailidis (2009) and further extended in Shojaie & Michailidis (2010). In particular, we review model formulation, estimation and statistical inference issues. Consider \( p \) genes (proteins/metabolites) whose activity levels across \( n \) samples are organized in a \( p \times n \) matrix \( D \). Let \( n_1 \) and \( n_2 \) be the number of samples in each experimental condition and denote \( n = n_1 + n_2 \).

In the framework of Network-based Gene Set Analysis, the effect of genes (proteins/metabolites) in the network are captured using a latent variable model (Shojaie & Michailidis, 2010). Denote by \( y \) an arbitrary column of the data matrix, and decompose the observed data into signal, \( Y \), plus noise, \( \varepsilon \), i.e. \( Y = X + \varepsilon \). The latent variable model assumes that the signal for each node is its baseline activity plus a linear combination of signals from its immediate neighbors in the network. In other words, for each node \( i \),

\[
X_i = \gamma_i + \sum_{i' \in \text{ne}_i} \rho_{i'i} X_{i'},
\]

where \( \text{ne}_i \) denotes the neighbors of \( i \) in the network, and \( \rho_{i'i} \) is the magnitude of effect of \( i' \) on \( i \). In case of undirected networks, \( i' \in \text{ne}_i \) if and only if \( i \in \text{ne}_{i'} \), and moreover, \( \rho_{i'i} = \rho_{i'i'} \). However, the Network-based Gene Set Analysis framework in general works with both directed and undirected graphs, and also if \( \rho_{i'i} \neq \rho_{i'i'} \). We collect the interaction weights \( \rho_{i'i}^{(k)} \) for condition \( k \) (\( k = 1, 2 \)) into the weighted adjacency matrix of the network \( A^{(k)} \), and consider the general setting where \( A^{(k)} \neq A^{(k')} \).

Throughout the paper, we assume that \( \rho_{ii} = 0 \), and will drop the the superscript of \( A^{(k)} \) when not necessary.

Assume that \( \gamma \) and \( \varepsilon \) are independent and normally distributed; specifically, \( \gamma \sim \mathcal{N}_p(\mu, \sigma^2_\gamma I_p) \) and \( \varepsilon \sim \mathcal{N}_p(0, \sigma^2_\varepsilon I_p) \), where \( I_p \) denotes the \( p \)-identity matrix. The Network-based Gene Set Analysis model can be summarized in vector notation as

\[
Y = \Lambda \gamma + \varepsilon. \tag{B.1}
\]

Here, \( \Lambda \) is the influence matrix of the graph, which is calculated from the adjacency matrix \( A \), and captures the propagated effect of each node on all other nodes in the network, beyond their immediate neighbors. Shojaie & Michailidis (2009) proposed a general characterization of \( \Lambda \) based on a normalized version of \( A \), motivated by the relationship between \( \Lambda \) and \( A \) in the special cases of directed acyclic or sub-stochastic graphs in Shojaie & Michailidis (2009). It turns out that, in the case of undirected Gaussian graphical models, \( \Lambda \) is equivalent to the Cholesky factor of the correlation matrix \( \Sigma \), and is
\[ \Lambda = \lim_{\zeta \to 0} (I - A(\zeta))^{-1} = (I - A)^{+}, \] where \( M^{+} \) denotes the Moore-Penrose pseudo-inverse of matrix \( M \), and \( A(\zeta) \) is found by normalizing the entries in each row of the adjacency matrix

\[ A(\zeta)_{ii'} = \frac{A_{ii'}}{\left( \sum_{i' = 1}^{p} |A_{ii'}| \right) + \zeta}, \] (B.2)

for some \( \zeta > 0 \).

Consider the gene expression matrix abovementioned with two experimental conditions and \( n_k \) observations at each condition. As described in Section 3, the problem can be formulated using mixed linear models as

\[ \mathcal{Y} = \Psi \beta + \Pi \mathcal{G} + \mathcal{E}, \quad \mathcal{E} \sim \mathcal{N}_{np}(0, \sigma_{\epsilon}^2 I_{np}), \quad \mathcal{G} \sim \mathcal{N}_{np}(0, \sigma_{\gamma}^2 I_{np}), \] (B.3)

where \( \beta \) and \( \mathcal{G} \) are fixed and random effect parameters.

The precise form of the design matrices \( \Psi \) and \( \Pi \) depends on whether the influence matrix \( \Lambda \) can change over time or over different experimental conditions (see Harbison et al. (2004) for examples of changes in regulatory networks in different experimental conditions). We refer readers to Shojaie & Michailidis (2010) for the general Network-based Gene Set Analysis under complex experimental designs and consider here the special case where \( \Lambda^{(k)} (k = 1, 2) \) is constant over time. Hence, the matrices \( \Psi \) and \( \Pi \) are defined as

\[ \Psi = \begin{pmatrix} \Lambda^{(1)} & \ldots & \Lambda^{(1)} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \Lambda^{(2)} & \ldots & \Lambda^{(2)} \end{pmatrix}^T, \quad \Pi = \text{diag}(\Lambda^{(1)} \ldots \Lambda^{(1)}, \Lambda^{(2)} \ldots \Lambda^{(2)})^T, \]

which are of dimensions \( np \times 2p \) and \( np \times np \), respectively.

For the mixed linear model of equation (B.3), write \( W \equiv \text{var}(\mathcal{Y}) = \sigma_{\gamma}^2 \Pi \Pi^T + \sigma_{\epsilon}^2 I_{np} \) and the maximum likelihood estimate of \( \beta \) is (Searle, 1971)

\[ \hat{\beta} = (\Psi^T \hat{W}^{-1} \Psi)^{-1} \Psi^T \hat{W}^{-1} \mathcal{Y}. \] (B.4)

The estimate depends on estimates of the variance components, \( \sigma_{\gamma}^2 \) and \( \sigma_{\epsilon}^2 \), which are usually estimated via optimization of profile log-likelihood, as discussed in Section 3.

By accounting for correlations through the network as well as other sources of correlation in the data, the Network-based Gene Set Analysis model provides a general and flexible framework for evaluating changes in molecular expression in complex experimental settings (see Shojaie & Michailidis, 2010, for examples of applications). Let \( C = (\Psi^T \hat{W}^{-1} \Psi)^{-1} \). As introduced in Section 3, with the contrast vector (Searle, 1971) \( l = b \Lambda \cdot b \), one can test the null hypothesis of no pathway activity vs the alternative of pathway activation

\[ H_0 : l \beta = 0, \quad H_1 : l \beta \neq 0, \]
using the following Wald test statistic

$$TS = \frac{l^\hat{\beta}}{\sqrt{l^\hat{ClT}}}.$$ 

Under the null hypothesis, the degrees of freedom $\nu$ of the above test statistic can be estimated using the Satterthwaite approximation method (McLean & Sanders, 1988)

$$\nu = \frac{2(l^\hat{ClT})^2}{\omega^T \hat{V} \omega},$$

where $\omega = \partial lCl^T / \partial \eta$, and $\hat{V}$ is the empirical covariance matrix of the parameter $\eta = (\sigma^2_y, \sigma^2_\varepsilon)^T$. The inferential framework of Network-based Gene Set Analysis also allows for simultaneous test of significance of multiple contrast vectors, formed into a contrast matrix $L$, where each row of $L$ includes one of the contrast vectors. In this case, the test statistic follows an $F$ distribution, whose degrees of freedom can be estimated from data. Due to space considerations, we omit the discussion of this latter test, and refer the interested reader to Shojaie & Michailidis (2010) for more details.

**Additional Simulation Results**

To benchmark the performance of the proposed network estimation procedure as well as Network-based Gene Set Analysis, we carried out another two experiments, which we describe as the third and fourth following the earlier two in the simulation section of the main text. These are also the two experiments we mentioned when comparing the running time of Network-based Gene Set Analysis with different variance estimation algorithms.

Our third experiment considers a undirected network with $p = 160$. The simulation design is similar to that in the second experiment, except that each of the 8 subnetworks has a denser structure. Specifically, there are 80 edges connecting the 20 genes in each subnetwork under the null. The probability of an interaction between subnetworks is 0.3. Under the alternative, there is an increase of 0.6 in mean expression values for varying proportions of genes (0%, 30%, 50% and 90%) for pathways 1–4 and 5–8. Moreover, half of the interactions in the latter four subnetworks disappear.

The fourth experiment is about networks of size $p = 400$, which also illustrates that the proposed method scales well with the size of the networks based on implementation of the updated optimization algorithm. Again the topology is similar to previous scenarios so that it consists of 20 subnetworks, each corresponding to a pathway with 20 genes. The probability of an interaction between pathways is also 0.3. All subnetworks have the same topology and were generated as scale-free random graphs such that there are 40 edges linking the 20 genes. We then divided the 20 subnetworks into two groups, with the first 10 in the first group, and the last 10 in the second group. Under the null, mean expression values for all subnetworks were set to be 1. Under the alternative, the first 6 subnetworks in each group remained to have the same mean expression values, but 20%, 30%, 30% and 40% of genes in the last four
subnetworks had 0.5 unit higher expression values, respectively. In addition, subnetwork structure for the second group under the alternative differed from their null equivalent by 22.5%. This experiment is also of interest because we created a setting where there are enough pathways in order for the permutation based Gene Set Analysis to calibrate the number of permutations required.

Table 6 presents the deviance measures for estimating the networks with 200 replicates and sample sizes of 300 for both $p = 160$ and $p = 400$, when varying levels of external information are available. In both experiments, we see performance improvement in Matthews correlation coefficient and Frobenius norm loss as the structural information of the networks $r$ increases. Under the alternative of $p = 160$, the slightly better performance in terms of Frobenius norm loss is due to the sparser network structure relative to the null.

Table 6: Deviance measures for network estimation in experiment 3 and 4

| $r$  | $p = 160$ | $p = 400$ |
|------|-----------|-----------|
|      | FPR(%)‡   | FNR(%) b  | MCC †  | Fnorm ‡ | FPR(%) | FNR(%) | MCC   | Fnorm |
| Null |           |           |        |         |        |        |       |       |
| 0.0  | 3.22      | 4.43      | 0.75   | 0.50    | 2.00   | 2.96   | 0.58  | 0.35  |
| 0.2  | 2.38      | 4.86      | 0.79   | 0.48    | 1.60   | 3.16   | 0.62  | 0.33  |
| 0.8  | 0.52      | 3.63      | 0.93   | 0.37    | 0.47   | 1.97   | 0.83  | 0.22  |
| Alternative |        |           |        |         |        |        |       |       |
| 0.0  | 2.98      | 2.94      | 0.73   | 0.40    | 1.72   | 2.33   | 0.61  | 0.35  |
| 0.2  | 2.20      | 3.36      | 0.78   | 0.38    | 1.35   | 2.58   | 0.66  | 0.32  |
| 0.8  | 0.41      | 3.42      | 0.93   | 0.27    | 0.37   | 1.83   | 0.86  | 0.22  |

‡ FPR(%), false positive rate in percentage;
 b FNR(%), false negative rate in percentage;
 † MCC, Matthews correlation coefficient;
 ‡ Fnorm, Frobenius norm loss.

Table 7 shows the estimated powers after correcting for false discovery rate in the third experiment with $p = 160$. While Gene Set Analysis with the competitive null hypothesis tends to suggest that none of the pathways is significantly differentially expressed under the alternative, its equivalent with the self-contained null overestimates powers for most pathways, particularly for pathway 5. In comparison, Network-based Gene Set Analysis slightly underestimates, but mostly correctly the significance of each subnetwork. Moreover, the differences in powers between each pair of pathways (1 and 5, 2 and 6, 3 and 7, as well as 4 and 8) indicate that the topologies for each pair are different, since both had the same amount of changes in mean expression values. When the exact networks are unknown, we see improvement in detected powers as the structural information increases from 20% to 80%, which suggests that a small amount of external knowledge is beneficial for making reliable inference using the network based method.

The estimated powers after correcting for false discovery rate in experiment 4 are shown separately in Table 8, as there are 20 pathways with varying parameters. When the exact networks with the correct edge weights are known, Network-based Gene Set Analysis returns estimated powers that match the true powers very well, with high powers for pathways 8, 9 and 10 which have significant changes.
Table 7: Powers based on false discovery rate with $q^* = 0.05$ in experiment 3

| Pathway | 0.2$^z$ | 0.8$^z$ | E$^z$ | T$^\dagger$ | GSA-s$^\ddagger$ | GSA-c$^\ddagger$ |
|---------|---------|---------|-------|----------|----------------|-----------------|
| 1       | 0.04    | 0.06    | 0.01  | 0.05     | 0.07           | 0.00            |
| 2       | 0.46    | 0.54    | 0.39  | 0.54     | 0.63           | 0.00            |
| 3       | 0.80    | 0.86    | 0.78  | 0.87     | 0.94           | 0.00            |
| 4       | 0.91    | 0.94    | 0.97  | 0.99     | 0.99           | 0.01            |
| 5       | 0.33    | 0.27    | 0.14  | 0.17     | 0.53           | 0.01            |
| 6       | 0.29    | 0.30    | 0.15  | 0.25     | 0.41           | 0.00            |
| 7       | 0.49    | 0.56    | 0.52  | 0.64     | 0.74           | 0.00            |
| 8       | 0.68    | 0.76    | 0.72  | 0.88     | 0.85           | 0.00            |

$^z$ 0.2/0.8 refer to Network-based Gene Set Analysis with 20%/80% external information;
$^z$ E refers to Network-based Gene Set Analysis with the exact networks;
$^\dagger$ T refers to the true power;
$^\ddagger$ GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively.

in mean expression values, high powers for pathways 11–16 that have significant changes in structures and high powers for pathways 17–20 with both changes. When there is 20% external information on the underlying pathway topology, Network-based Gene Set Analysis is able to identify mostly correctly the powers for all pathways, with slight overestimation for pathways 4, 5 and 6. There is also improvement when 80% structural information is known, although the improvement is minor compared to the amount of structural information required. In comparison, Gene Set Analysis with the self-contained null hypothesis also performs well in recognizing correctly the differentially expressed pathways. However, the equally high powers for pathways 11–20 fail to reveal the actual differences among these pathways. On the other hand, Gene Set Analysis with the competitive null is still not able to identify any differential expression among all 20 pathways. The conflicting results from Gene Set Analysis with different null hypotheses also raise concerns as to which version to choose in practice.

Derivation for Newton’s Method

The implementation of Newton’s method requires the gradient and the Hessian of the objective function. Here we provide details about how to calculate the two important derivatives based on the profile log-likelihood when profiling out $\sigma_e$ in Section 3. The derivation follows similarly when profiling out $\sigma_{\gamma}$.

Given the observations $Y_1, \ldots, Y_n$ (with the first $n_1$ samples from condition 1 and the remaining
Table 8: Powers based on false discovery rate with $q^* = 0.05$ in experiment 4

| Pathway | $0.2^\sharp$ | $0.8^\sharp$ | $E^\flat$ | $T^\dagger$ | GSA-s$\ddagger$ | GSA-c$\ddagger$ |
|---------|-----------|-----------|------|------|------------|------------|
| 1       | 0.09      | 0.10      | 0.06 | 0.05 | 0.11       | 0.00       |
| 2       | 0.12      | 0.11      | 0.04 | 0.05 | 0.10       | 0.00       |
| 3       | 0.26      | 0.17      | 0.04 | 0.05 | 0.10       | 0.00       |
| 4       | 0.36      | 0.26      | 0.04 | 0.05 | 0.08       | 0.01       |
| 5       | 0.42      | 0.34      | 0.03 | 0.05 | 0.10       | 0.00       |
| 6       | 0.48      | 0.38      | 0.01 | 0.05 | 0.07       | 0.00       |
| 7       | 0.47      | 0.38      | 0.17 | 0.26 | 0.42       | 0.00       |
| 8       | 0.92      | 0.95      | 1.00 | 1.00 | 1.00       | 0.00       |
| 9       | 0.75      | 0.81      | 0.92 | 0.94 | 0.99       | 0.00       |
| 10      | 0.89      | 0.95      | 1.00 | 1.00 | 1.00       | 0.00       |
| 11      | 0.73      | 0.76      | 0.80 | 0.82 | 0.97       | 0.00       |
| 12      | 0.74      | 0.81      | 0.83 | 0.86 | 0.98       | 0.00       |
| 13      | 0.72      | 0.71      | 0.79 | 0.80 | 0.94       | 0.01       |
| 14      | 0.74      | 0.77      | 0.69 | 0.76 | 0.95       | 0.00       |
| 15      | 0.82      | 0.78      | 0.83 | 0.85 | 0.96       | 0.00       |
| 16      | 0.81      | 0.82      | 0.80 | 0.85 | 0.97       | 0.00       |
| 17      | 0.81      | 0.86      | 0.96 | 0.98 | 0.99       | 0.00       |
| 18      | 0.99      | 1.00      | 1.00 | 1.00 | 1.00       | 0.00       |
| 19      | 0.95      | 0.97      | 1.00 | 1.00 | 1.00       | 0.00       |
| 20      | 0.98      | 1.00      | 1.00 | 1.00 | 1.00       | 0.00       |

$\sharp$ 0.2/0.8 refer to Network-based Gene Set Analysis with 20%/80% external information;
\flat E refers to Network-based Gene Set Analysis with the exact networks;
\dagger T refers to the true power;
\ddagger GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively.

\[ n_2 = n - n_1 \text{ samples from condition 2}, \] the nonconstant part of the “full” log-likelihood \( l_F \) is

\[
l_F(\sigma, \tau \mid Y_1, \ldots, Y_n) = -\frac{1}{2} \left\{ n_1 \logdet(\sigma^2 \Sigma^{(1)}) + n_2 \logdet(\sigma^2 \Sigma^{(2)}) \right\} - \frac{1}{2} \sigma^{-2} \left\{ \sum_{j=1}^{n_1} R_j^T (\Sigma^{(1)})^{-1} R_j + \sum_{j=n_1+1}^{n} R_j^T (\Sigma^{(2)})^{-1} R_j \right\},
\]

where \((R_j)_{j=1}^{n} \) are the vector of residuals and \( \Sigma^{(k)} (k = 1, 2) \) are the covariance matrices as defined in section 3. Similarly, the nonconstant part of the log-likelihood using restricted maximum likelihood is

\[
l_R(\sigma, \tau \mid Y_1, \ldots, Y_n) = l_F(\sigma, \tau \mid Y_1, \ldots, Y_n)
- \frac{1}{2} \logdet \left\{ n_1 \sigma^{-2} (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)} \right\} - \frac{1}{2} \logdet \left\{ n_2 \sigma^{-2} (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)} \right\}.
\]

To simplify the computation, we solve for \( \sigma^2 \) as a function of \( \tau \). The maximum likelihood estimate of
\[ \sigma^2 \text{ is} \]
\[
\hat{\sigma}^2 = \frac{1}{N} \left\{ \sum_{j=1}^{n_1} R_j^T (\Sigma^{(1)})^{-1} R_j + \sum_{j=n_1+1}^{n} R_j^T (\Sigma^{(2)})^{-1} R_j \right\},
\]

whereas its restricted maximum likelihood estimate is defined in equation (3.3) of Section 3. Substituting \( \sigma^2 \) with the corresponding estimate, we obtain the profile log-likelihood

\[
p_F(\tau \mid Y_1, \ldots, Y_n) = -\frac{1}{2} \left( n_1 \log \det \Sigma^{(1)} + n_2 \log \det \Sigma^{(2)} \right)
- \frac{1}{2} N \log \left\{ \sum_{j=1}^{n_1} R_j^T (\Sigma^{(1)})^{-1} R_j + \sum_{j=n_1+1}^{n} R_j^T (\Sigma^{(2)})^{-1} R_j \right\},
\]

for maximum likelihood, and equation (3.2) for restricted maximum likelihood.

As \( \Sigma^{(k)} \ (k = 1, 2) \) are the only terms that depend on \( \tau \), we first look at the derivatives of \( \log \det \Sigma^{(k)}, R_j^T (\Sigma^{(k)})^{-1} R_j \) and \( \log \det \{ (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \Lambda^{(k)} \} \) with respect to \( \tau \). The derivatives of the profile log-likelihood can be calculated easily from these quantities. Denote

\[
B^{(k)} = (\Sigma^{(k)})^{-1} \frac{d\Sigma^{(k)}}{d\tau} (\Sigma^{(k)})^{-1}, \quad H^{(k)} = (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \Lambda^{(k)}.
\]

Then

\[
\frac{d \log \det (\Sigma^{(k)})}{d\tau} = \text{tr} \left\{ (\Sigma^{(k)})^{-1} d\Sigma^{(k)} \right\},
\]

\[
\frac{d^2 \log \det (\Sigma^{(k)})}{d\tau^2} = \text{tr} \left\{ - (B^{(k)})^T d\Sigma^{(k)} + (\Sigma^{(k)})^{-1} d^2 \Sigma^{(k)} \right\},
\]

\[
\frac{d R_j^T (\Sigma^{(k)})^{-1} R_j}{d\tau} = -R_j^T B^{(k)} R_j, \quad \frac{d^2 R_j^T (\Sigma^{(k)})^{-1} R_j}{d\tau^2} = -R_j^T \frac{dB^{(k)}}{d\tau} R_j,
\]

\[
\frac{d \log \det H^{(k)}}{d\tau} = -\text{tr} \left\{ (H^{(k)})^{-1} (\Lambda^{(k)})^T B^{(k)} \Lambda^{(k)} \right\},
\]

\[
\frac{d^2 \log \det H^{(k)}}{d\tau^2} = -\text{tr} \left\{ (H^{(k)})^{-1} (\Lambda^{(k)})^T \frac{dB^{(k)}}{d\tau} \Lambda^{(k)} \right\}
- \text{tr} \left\{ (H^{(k)})^{-1} (\Lambda^{(k)})^T \frac{d^2 \Sigma^{(k)}}{d\tau^2} \right\},
\]

where

\[
\frac{dB^{(k)}}{d\tau} = - (\Sigma^{(k)})^{-1} \left\{ \frac{d\Sigma^{(k)}}{d\tau} (\Sigma^{(k)})^{-1} \frac{d\Sigma^{(k)}}{d\tau} - \frac{d^2 \Sigma^{(k)}}{d\tau^2} \right\} (\Sigma^{(k)})^{-1}.
\]

Given the covariance \( \Sigma^{(k)} \ (k = 1, 2) \) defined in Section 3, we can further simplify the above derivatives.
and obtain
\[
\frac{d \log \det \Sigma^{(k)}}{d \tau} = \text{tr} \left\{ H^{(k)} \right\}, \quad \frac{d^2 \log \det \Sigma^{(k)}}{d \tau^2} = -\text{tr} \left\{ H^{(k)} H^{(k)} \right\},
\]
\[
\frac{d R_j^T (\Sigma^{(k)})^{-1} R_j}{d \tau} = -R_j^T (\Sigma^{(k)})^{-1} \Lambda^{(k)} (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} R_j,
\]
\[
\frac{d^2 R_j^T (\Sigma^{(k)})^{-1} R_j}{d \tau^2} = 2R_j^T (\Sigma^{(k)})^{-1} \Lambda^{(k)} H^{(k)} (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} R_j,
\]
\[
\frac{d \log \det H^{(k)}}{d \tau} = -\text{tr} \left\{ H^{(k)} \right\}, \quad \frac{d^2 \log \det H^{(k)}}{d \tau^2} = \text{tr} \left\{ H^{(k)} H^{(k)} \right\}.
\]