Supplementary materials to "Network-based pathway enrichment analysis with incomplete network information"

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

[1]*Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine*
[2]*Department of Biostatistics, University of Washington*
[3]*Department of Statistics, University of Florida*

A Theoretical Analysis and Proofs

We first introduce additional notation needed in the remainder. Define $\tilde{\Omega}_0 = \mathrm{diag}(\Omega_0) + \Omega_{0,E\cap\hat{E}}$, where E and $\hat{E}$ are the true and the estimated edge set, respectively. By definition, $\tilde{\Omega}_0$ and Ω_0 will be different at position (i, i') only when the edge (i, i') is falsely rejected. In the following, we first derive an upper bound for the size of $\hat{E}$ and $\|\tilde{\Omega}_0 - \Omega_0\|_F$. For the ease of presentation, we drop the superscript i for sets J_0 and J_1 in the ith regression, but they should be understood as J_0^i and J_1^i, respectively.

The following lemma is needed in the proof of Theorem 1 below.

Lemma 1. *For $i = 1, \ldots, p$, denote by $\boldsymbol{\xi}^i = \mathbf{Z}_i - \sum_{i'\neq i}\theta^i_{i'}\mathbf{Z}_{i'}$, where $\boldsymbol{\theta}^i$ is the optimal prediction coefficient vector in the ith regression. Consider the event*

$$\mathcal{F}_i := \left\{\mathbf{Z} : \frac{1}{m}\|\mathbf{Z}^T_{-i}\boldsymbol{\xi}^i\|_\infty \leq \frac{c_1}{2}\sqrt{\frac{\log(p-rp)}{m\omega_{0,ii}}}\right\}$$

with a constant $c_1 > 4$, where $\omega_{0,ii}$ is the ith diagonal element of the true inverse covariance matrix Ω_0. Define the event $\mathcal{F} = \bigcap_{i=1}^p \mathcal{F}_i$. Then $\mathbb{P}(\mathcal{F}) \geq 1 - 2p^{2-c_1^2/8}$.

The proof of Lemma 1 will be provided shortly. Denote by $\Lambda_{\max}$ the maximal eigenvalue of $\mathbf{Z}^T\mathbf{Z}/m$. Conditioning on the event $\mathcal{F}$, we have the following results on controlling the size of $\hat{E}$ and the Frobenius norm of the deviance, $\|\tilde{\Omega}_0 - \Omega_0\|_F$.

*To whom correspondence should be addressed: *jinma@upenn.edu*.

Theorem 1. *Suppose the conditions in Theorem 2.2 are satisfied. Then on event $\mathcal{F}$, for appropriately chosen λ, we have*

$$|\hat{E}| \leq \frac{64\Lambda_{\max}}{\kappa^2(s)}(1-r)S_0 + rS_0, \tag{A.1}$$

and

$$\|\tilde{\Omega}_0 - \Omega_0\|_F \leq c_3\sqrt{\frac{S_0\log(p-rp)}{m}} \leq k_1\phi_1, \tag{A.2}$$

where $c_3 = 16c_1\sqrt{(1-r)}/\kappa^2(2s)$.

Remark 1. *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 $\kappa(s)$, which is necessarily positive by the assumption that $\kappa(2s) > 0$. Two extreme cases occur when (i) $r = 0$, i.e. we do not observe anyexternal information, thus reducing problem (2.4) to the original neighborhood selection in [8]; (ii) $r = 1$, i.e. the exact network topology is known and hence $\hat{E} = E$. On the other hand, the upper bound for $\|\tilde{\Omega}_0 - \Omega_0\|_F$ 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 1. Recall $\tilde{J} = V\backslash\{J_1 \cup J_0 \cup \{i\}\}$ is the set of indices for which there is no information available. Denote by $\mathbf{P}_{J_1} = \mathbf{Z}_{J_1}(\mathbf{Z}_{J_1}^T\mathbf{Z}_{J_1})^{-1}\mathbf{Z}_{J_1}^T$ the projection onto the column space of $\mathbf{Z}_{J_1}$. It is easy to see that the problem (2.4) is equivalent to solving

$$\min_{\boldsymbol{\theta}_{\tilde{J}}} \frac{1}{m}\|(\mathbf{I}_p - \mathbf{P}_{J_1})\mathbf{Z}_i - \mathbf{Z}_{\tilde{J}}\boldsymbol{\theta}_{\tilde{J}}\|_2^2 + 2\lambda\|\boldsymbol{\theta}_{\tilde{J}}\|_1. \tag{A.3}$$

To bound $\hat{E}$ and $\|\tilde{\Omega}_0 - \Omega_0\|_F$, it suffices to focus mainly on the set $\tilde{J}$, as false positive and negative errors will only occur on this set.

Denote by s_1^i and s^i, respectively, the number of known ones and the number of nonzero coordinates after excluding the known ones in the ith regression, and $s = \max_{i=1,\ldots,p}(s_1^i + s^i)$. If $\mathbf{Z}$ satisfies the restricted eigenvalue condition in Assumption 2 with $\kappa(2s) > 0$, then $\mathbf{Z}_{\tilde{J}}$ satisfies the same assumption with $\kappa(2s^i) \geq \kappa(2s) > 0$ for $s^i \leq s$. Moreover, $\kappa(s^i) \geq \kappa(s) \geq \kappa(2s) > 0$. Let $\hat{\boldsymbol{\theta}}_{\tilde{J}}^i$ be the lasso estimator in (A.3) with

$$\lambda = c_1\sqrt{\frac{\log(p-rp)}{m\omega_{0,ii}}} \tag{A.4}$$

for $c_1 > 4$. Conditioning on the event $\mathcal{F}$, we can invoke Theorem 7.2 of [2] and obtain simultaneously for

all i,

$$\|\hat{\boldsymbol{\theta}}^i_{\tilde{J}}\|_0 \leq \frac{64\Lambda_{\max}}{\kappa^2(s^i)} s^i, \tag{A.5}$$

and

$$\|\hat{\boldsymbol{\theta}}^i_{\tilde{J}} - \boldsymbol{\theta}^i_{\tilde{J}}\|_2 \leq \frac{16c_1}{\omega_{0,ii}\kappa^2(2s^i)} \sqrt{\frac{s^i \log(p-rp)}{m}}. \tag{A.6}$$

Combining (A.5) with the number of known edges s^i_1 as given in J^i_1, we get

$$|\hat{E}| \leq \sum_{i=1}^p \{\|\hat{\boldsymbol{\theta}}^i_{\tilde{J}}\|_0 + |J^i_1|\} \leq \frac{64\Lambda_{\max}}{\kappa^2(s)} \sum_{i=1}^p s^i + \sum_{i=1}^p s^i_1.$$

The upper bound in (A.1) follows immediately, since by definition the number of known and unknown edges are $\sum_{i=1}^p s^i_1 = rS_0$ and $\sum_{i=1}^p s^i = (1-r)S_0$, respectively.

To bound $\|\tilde{\Omega}_0 - \Omega_0\|_F$, recall that for every $i' \neq i$, $\omega_{0,ii'} = -\theta^i_{i'}\omega_{0,ii}$. Using the bound in (A.6), we have

$$\begin{aligned}
\|\tilde{\Omega}_0 - \Omega_0\|^2_F &= \sum_{i=1}^p \sum_{i' \in J(\theta^i) \cap J(\hat{\theta}^i)^c} (\theta^i_{i'}\omega_{0,ii})^2 = \sum_{i=1}^p \omega^2_{0,ii} \sum_{i' \in J(\theta^i) \cap J(\hat{\theta}^i)^c} |\theta^i_{i'} - \hat{\theta}^i_{i'}|^2 \\
&\leq \sum_{i=1}^p \omega^2_{0,ii} \|\boldsymbol{\theta}^i_{\tilde{J}} - \hat{\boldsymbol{\theta}}^i_{\tilde{J}}\|^2_2 \leq \left\{\frac{16c_1}{\kappa^2(2s)}\right\}^2 \frac{(1-r)S_0 \log(p-rp)}{m}.
\end{aligned}$$

The last inequality in (A.2) follows from condition (2.7) in Theorem 2.2. □

Proof of Lemma 1. For every i, it is easy to verify that $\boldsymbol{\xi}^i$ is normally distributed with mean $\mathbf{0}$ and variance $1/\omega_{0,ii}\mathbf{I}_m$. Define random variables $\Upsilon_{ii'} = (\omega_{0,ii}/m)^{1/2}\mathbf{Z}^T_{i'}\boldsymbol{\xi}^i$ for $i' \neq i$. Then, $\mathbf{Z}^T_{i'}\mathbf{Z}_{i'}/m = 1$ implies that $\Upsilon_{ii'} \sim \mathcal{N}(0,1)$. Let λ be defined as in (A.4). Using an elementary bound on the tails of Gaussian distributions,

$$\begin{aligned}
\mathbb{P}(\mathcal{F}^c) &\leq \sum_{i=1}^p \sum_{i' \neq i} \mathbb{P}\left(\{|\mathbf{Z}^T_{i'}\boldsymbol{\xi}^i|/m > \lambda/2\}\right) \\
&\leq \sum_{i=1}^p \sum_{i' \neq i} \mathbb{P}\left(|\Upsilon_{ii'}| > (m\omega_{0,ii})^{1/2}\lambda/2\right) \leq \sum_{i=1}^p \sum_{i' \neq i} 2\exp\left\{-m\omega_{0,ii}\lambda^2/8\right\} \\
&\leq 2p(p-1)\exp\left\{-c_1^2\log(p-rp)/8\right\} \leq 2p^{2-c_1^2/8}.
\end{aligned}$$

Therefore, $\mathbb{P}(\mathcal{F}) \geq 1 - 2p^{2-c_1^2/8}$. □

With Lemma 1 and Theorem 1, we are ready to prove our main result in Theorem 2.2. The following proof is adapted from [12].

Proof of Theorem 2.2. Consider $\hat{\Omega}$ defined in (2.5). It suffices to show that on the event $\mathcal{F}$

$$\|\hat{\Omega} - \tilde{\Omega}_0\|_F = O\left(\sqrt{\frac{S_0 \log(p - rp)}{m}}\right),$$

since by triangle inequality and Theorem 1, we can conclude

$$\|\hat{\Omega} - \Omega_0\|_F \leq \|\hat{\Omega} - \tilde{\Omega}_0\|_F + \|\tilde{\Omega}_0 - \Omega_0\|_F \leq O\left(\sqrt{\frac{S_0 \log(p - rp)}{m}}\right).$$

Denote $\tilde{\Sigma}_0 = \tilde{\Omega}_0^{-1}$, which is positive definite since by Theorem 1,

$$\phi_{\min}(\tilde{\Omega}_0) \geq \phi_{\min}(\Omega_0) - \|\tilde{\Omega}_0 - \Omega_0\|_2 \geq \phi_{\min}(\Omega_0) - \|\tilde{\Omega}_0 - \Omega_0\|_F \geq \phi_1 - k_1\phi_1 > 0. \tag{A.7}$$

The first inequality in (A.7) comes from the fact that for any nonzero vector $\boldsymbol{\delta} \in \mathbb{R}^p$, $\boldsymbol{\delta}^T\tilde{\Omega}_0\boldsymbol{\delta} = \boldsymbol{\delta}^T\Omega_0\boldsymbol{\delta} + \boldsymbol{\delta}^T(\tilde{\Omega}_0 - \Omega_0)\boldsymbol{\delta} \geq \phi_{\min}(\Omega_0) - \phi_{\max}(\tilde{\Omega}_0 - \Omega_0)$.

Given $\tilde{\Omega}_0 \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}}^p$, define a new convex set:

$$\mathcal{U}_m(\tilde{\Omega}_0) = \{\mathbf{B} - \tilde{\Omega}_0 \mid \mathbf{B} \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}}^p\} \subset \mathcal{S}_{\hat{E}}^p.$$

Let

$$Q(\Omega) = \operatorname{trace}(\Omega\hat{\Sigma}) - \operatorname{trace}(\tilde{\Omega}_0\hat{\Sigma}) - \operatorname{logdet}\Omega + \operatorname{logdet}\tilde{\Omega}_0.$$

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

The main idea of this proof is as follows. For a sufficiently large $M > 0$, consider sets

$$\mathcal{T}_1 = \{\Delta \in \mathcal{U}_m(\tilde{\Omega}_0), \|\Delta\|_F = Mr_m\}, \quad \mathcal{T}_2 = \{\Delta \in \mathcal{U}_m(\tilde{\Omega}_0), \|\Delta\|_F \leq Mr_m\},$$

where

$$r_m = \sqrt{\frac{S_0 \log(p - rp)}{m}}.$$

Note that $\mathcal{T}_1$ is non-empty. Indeed, consider $\mathbf{B}_\epsilon = \epsilon\tilde{\Omega}_0$ for $\epsilon = Mr_m/\|\tilde{\Omega}_0\|_F$. Then $\mathbf{B}_\epsilon = (1+\epsilon)\tilde{\Omega}_0 - \tilde{\Omega}_0 \in \mathcal{U}_m(\tilde{\Omega}_0)$, hence $\mathbf{B}_\epsilon \in \mathcal{T}_1$. Denote by $\bar{0}$ the matrix of all zero entries. It is clear that $G(\Delta)$ is convex, and $G(\hat{\Delta}) \leq G(\bar{0}) = Q(\tilde{\Omega}_0) = 0$. Thus if we can show that $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_1$, the minimizer $\hat{\Delta}$ must

be inside $\mathcal{T}_2$ and hence $\|\hat{\Delta}\|_F \leq Mr_m$. To see this, note that the convexity of $Q(\Omega)$ implies that

$$\inf_{\|\Delta\|_F = Mr_m} Q(\tilde{\Omega}_0 + \Delta) > Q(\tilde{\Omega}_0) = 0.$$

There exists therefore a local minimizer in the ball $\{\tilde{\Omega}_0 + \Delta : \|\Delta\|_F \leq Mr_m\}$, or equivalently, for $\hat{\Delta} \in \mathcal{T}_2$, i.e. $\|\hat{\Delta}\|_F \leq Mr_m$.

In the remainder of the proof, we focus on

$$G(\Delta) = Q(\Delta + \tilde{\Omega}_0) = \operatorname{trace}(\Delta\hat{\Sigma}) - \operatorname{logdet}(\Delta + \tilde{\Omega}_0) + \operatorname{logdet}\tilde{\Omega}_0. \tag{A.8}$$

Applying a Taylor expansion to $\operatorname{logdet}(\tilde{\Omega}_0 + \Delta)$ in (A.8) gives

$$\begin{aligned}
&\operatorname{logdet}(\tilde{\Omega}_0 + \Delta) - \operatorname{logdet}\tilde{\Omega}_0 \\
=&\frac{d}{dt}\operatorname{logdet}(\tilde{\Omega}_0 + t\Delta)\big|_{t=0}\Delta + \int_0^1 (1-t)\frac{d^2}{dt^2}\operatorname{logdet}(\tilde{\Omega}_0 + t\Delta)dt \\
=&\operatorname{trace}(\Delta\tilde{\Sigma}_0) - \operatorname{vec}(\Delta)^T\left\{\int_0^1 (1-t)(\tilde{\Omega}_0 + t\Delta)^{-1} \otimes (\tilde{\Omega}_0 + t\Delta)^{-1}dt\right\}\operatorname{vec}(\Delta),
\end{aligned} \tag{A.9}$$

where $\operatorname{vec}(\Delta)$ denotes the vectorized Δ, and $\otimes$ is the Kronecker product. For $\Delta \in \mathcal{T}_1$, let K_1 be the integral term in (A.9), and define

$$K_2 = \operatorname{trace}\left\{\Delta(\hat{\Sigma} - \Sigma_0)\right\}, \quad K_3 = \operatorname{trace}\left\{\Delta(\tilde{\Sigma}_0 - \Sigma_0)\right\}.$$

We can then write

$$G(\Delta) = K_1 + \operatorname{trace}(\Delta\hat{\Sigma}) - \operatorname{trace}(\Delta\tilde{\Sigma}_0) = K_1 + K_2 - K_3.$$

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 $\hat{\Sigma}$ and Σ_0 are the same after scaling,

$$|K_2| \leq |\sum_{i \neq i'}(\hat{\Sigma}_{ii'} - \Sigma_{0,ii'})\Delta_{ii'}|.$$

By Lemma A.3 of [1], there exists a positive constant c_2 depending on $\phi_{\max}(\Sigma_0)$ such that

$$\max_{i \neq i'} |\hat{\Sigma}_{ii'} - \Sigma_{0,ii'}| \leq c_2\sqrt{\frac{\log(p - rp)}{m}},$$

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

$$|K_2| \leq c_2 \sqrt{\frac{\log(p-rp)}{m}} \|\Delta^-\|_1. \tag{A.10}$$

For K_3, we can use the upper bound for $\|\tilde{\Omega}_0 - \Omega_0\|_F$ in (A.2), and the lower bound for $\phi_{\min}(\tilde{\Omega}_0)$ in (A.7), to write,

$$|K_3| \leq \|\Delta\|_F \|\tilde{\Sigma}_0 - \Sigma_0\|_F \leq \|\Delta\|_F \frac{\|\tilde{\Omega}_0 - \Omega_0\|_F}{\phi_{\min}(\tilde{\Omega}_0)\phi_{\min}(\Omega_0)} \tag{A.11}$$

$$\leq \|\Delta\|_F \frac{c_3 \{S_0 \log(p-rp)/m\}^{1/2}}{(1-k_1)\phi_1^2}. \tag{A.12}$$

The second inequality in (A.11) comes from the rotation invariant property of Frobenius norm, i.e.

$$\|\tilde{\Sigma}_0 - \Sigma_0\|_F = \|\Sigma_0(\Omega_0 - \tilde{\Omega}_0)\tilde{\Sigma}_0\|_F \leq \phi_{\max}(\Sigma_0)\|\Omega_0 - \tilde{\Omega}_0\|_F \phi_{\max}(\tilde{\Sigma}_0).$$

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

$$\phi_{\max}(\tilde{\Omega}_0) \leq \phi_{\max}(\Omega_0) + \|\tilde{\Omega}_0 - \Omega_0\|_2 \leq \phi_{\max}(\Omega_0) + \|\tilde{\Omega}_0 - \Omega_0\|_F \leq \frac{1}{\phi_2} + k_1\phi_1.$$

Since $r_m \to 0$, there exists a sufficiently large $k_2 > 0$ such that for $\Delta \in \mathcal{T}_1$,

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

Following [10, Page 502, proof of Theorem 1], a lower bound for K_1 can be found as

$$\begin{aligned} K_1 &\geq \|\Delta\|_F^2 / \{2(\phi_{\max}(\tilde{\Omega}_0) + \|\Delta\|_2)^2\} \\ &\geq \|\Delta\|_F^2 / \{2\left(1/\phi_2 + k_1\phi_1 + k_2/\phi_2\right)^2\} = \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} \|\Delta\|_F^2. \end{aligned} \tag{A.13}$$

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

$$\begin{aligned} G(\Delta) \geq\ & \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} \|\Delta\|_F^2 - c_2 \sqrt{\frac{\log(p-rp)}{m}} \|\Delta^-\|_1 \\ & - \frac{c_3}{(1-k_1)\phi_1^2} \sqrt{\frac{S_0 \log(p-rp)}{m}} \|\Delta\|_F. \end{aligned}$$

For $\Delta \in \mathcal{T}_1$, applying Cauchy-Schwarz inequality yields

$$\|\Delta^-\|_1 \leq \sqrt{|\hat{E}|} \cdot \|\Delta^-\|_F.$$

We thus have

$$\begin{aligned} G(\Delta) &\geq \frac{\phi_2^2}{2(1+k_1\phi_1\phi_2+k_2)^2}\|\Delta\|_F^2 - c_2\sqrt{\frac{|\hat{E}|\log(p-rp)}{m}}\|\Delta^-\|_F \\ &\quad - \frac{c_3}{(1-k_1)\phi_1^2}\sqrt{\frac{S_0\log(p-rp)}{m}}\|\Delta\|_F \\ &\geq \|\Delta\|_F^2\left\{\frac{\phi_2^2}{2(1+k_1\phi_1\phi_2+k_2)^2} - \frac{c_2}{M}\sqrt{\frac{|\hat{E}|}{S_0}} - \frac{c_3}{M(1-k_1)\phi_1^2}\right\} > 0, \end{aligned}$$

for M sufficiently large. □

Proof of Corollary 1. Under the assumptions in Theorem 2.2, we have

$$\|\Delta_{\Omega_0}\|_2 = \|\hat{\Omega} - \Omega_0\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{S_0\log(p-rp)}{m}}\right) = o_{\mathbb{P}}(1).$$

The partial correlation matrix corresponding to $\hat{\Omega}$ can be written as

$$\hat{\mathbf{A}} = \mathbf{I}_p - \hat{\mathbf{D}}^{-1/2}\hat{\Omega}\hat{\mathbf{D}}^{-1/2} = \mathbf{A}_0 + \mathbf{D}_0^{-1/2}\Omega_0\mathbf{D}_0^{-1/2} - (\hat{\mathbf{D}})^{-1/2}\hat{\Omega}\hat{\mathbf{D}}^{-1/2} = \mathbf{A}_0 + \Delta_{\mathbf{A}_0},$$

where

$$\begin{aligned} \Delta_{\mathbf{A}_0} &= \mathbf{D}_0^{-1/2}\Omega_0\mathbf{D}_0^{-1/2} - (\hat{\mathbf{D}})^{-1/2}\hat{\Omega}\hat{\mathbf{D}}^{-1/2} \\ &= \mathbf{D}_0^{-1/2}(\Omega_0 - \hat{\Omega})\mathbf{D}_0^{-1/2} + \mathbf{D}_0^{-1/2}\hat{\Omega}\big(\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\big) + \big(\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\big)\hat{\Omega}\hat{\mathbf{D}}^{-1/2}. \end{aligned} \tag{A.14}$$

Next we show that each of the summands on the right hand side of (A.14) has ℓ_2 norm $o_{\mathbb{P}}(1)$ and conclude thus $\|\Delta_{\mathbf{A}_0}\|_2 = o_{\mathbb{P}}(1)$.

By Assumption 1, the diagonal entries of Ω_0 satisfy $\omega_{0,ii} \geq \phi_{\min}(\Omega_0) \geq \phi_1$ for all $i = 1, \ldots, p$. Thus, $\|\mathbf{D}_0^{-1/2}\|_2 = \max_i \omega_{0,ii}^{-1/2} \leq \phi_1^{-1/2}$. It follows that

$$\|\mathbf{D}_0^{-1/2}(\Omega_0 - \hat{\Omega})\mathbf{D}_0^{-1/2}\|_2 \leq \|\mathbf{D}_0^{-1/2}\|_2^2\|\Omega_0 - \hat{\Omega}\|_2 = o_{\mathbb{P}}(1).$$

For the remaining two terms, first notice that $\|\mathbf{D}_0 - \hat{\mathbf{D}}\|_2 \leq \|\mathbf{D}_0 - \hat{\mathbf{D}}\|_F \leq \|\Omega_0 - \hat{\Omega}\|_F = o_{\mathbb{P}}(1)$. Therefore,

$$\begin{aligned}\|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 &= \max_{i=1,\ldots,p} |\omega_{0,ii}^{-1/2} - \hat{\omega}_{ii}^{-1/2}| = \max_{i=1,\ldots,p} \left| \frac{\omega_{0,ii}^{1/2} - \hat{\omega}_{ii}^{1/2}}{\omega_{0,ii}^{1/2}\hat{\omega}_{ii}^{1/2}} \right| \\ &= \max_{i=1,\ldots,p} \left| \frac{\omega_{0,ii} - \hat{\omega}_{ii}}{\omega_{0,ii}^{1/2}\hat{\omega}_{ii}^{1/2}(\omega_{0,ii}^{1/2} + \hat{\omega}_{ii}^{1/2})} \right| \leq \phi_1^{-1}(\phi_1 - o_{\mathbb{P}}(1))^{-1/2}\|\mathbf{D}_0 - \hat{\mathbf{D}}\|_2,\end{aligned}$$

where the last inequality comes from that fact that

$$\min_i |\hat{\omega}_{ii}| = \min_i |\hat{\omega}_{ii} - \omega_{0,ii} + \omega_{0,ii}| \geq \min_i |\omega_{0,ii}| - \max_i |\hat{\omega}_{ii} - \omega_{0,ii}| \geq \phi_1 - o_{\mathbb{P}}(1).$$

Hence, $\|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1)$. Note further,

$$\|\hat{\Omega}\|_2 = \|\hat{\Omega} - \Omega_0 + \Omega_0\|_2 \leq \|\Omega_0\|_2 + \|\hat{\Omega} - \Omega_0\|_2 = \|\Omega_0\|_2 + o_{\mathbb{P}}(1)$$

is bounded above. It follows thus,

$$\begin{aligned}\|\mathbf{D}_0^{-1/2}\hat{\Omega}\big(\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\big)\|_2 &\leq \|\mathbf{D}_0^{-1/2}\|_2\|\hat{\Omega}\|_2\|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1), \\ \|\big(\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\big)\hat{\Omega}\hat{\mathbf{D}}^{-1/2}\|_2 &\leq \|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2\|\hat{\Omega}\|_2\|\hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1).\end{aligned}$$

This completes the proof.

□

The following proof of Theorem 3.1 adapts from that of Theorem 2.1 in [11].

Proof of Theorem 3.1. Consider the special case where the row vector $\mathbf{b} = \mathbf{1}^T$, i.e. the whole network is tested as one pathway. The general case when $\mathbf{b} \neq \mathbf{1}^T$ follows from a similar argument.

For the partial correlation $\mathbf{A}_0^{(k)}$ $(k = 1, 2)$ defined in Section 3.2, it holds that $\Lambda^{(k)}(\Lambda^{(k)})^T = (\mathbf{I}_p - \mathbf{A}_0^{(k)})^{-1} = \sum_{t=0}^{\infty}(\mathbf{A}_0^{(k)})^t$. Hence

$$\begin{aligned}\hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T = \sum_{t=0}^{\infty}(\hat{\mathbf{A}}^{(k)})^t &= \sum_{t=0}^{\infty}(\mathbf{A}_0^{(k)})^t + \sum_{t=1}^{\infty}\sum_{u=1}^{t}\binom{t}{u}(\mathbf{A}_0^{(k)})^{t-u}(\Delta_{\mathbf{A}_0^{(k)}})^u \\ &= \Lambda^{(k)}(\Lambda^{(k)})^T + \Delta_{\Lambda^{(k)}}.\end{aligned}$$

For $\hat{\mathbf{A}}^{(k)}$ defined under the assumptions in Theorem 2.2 and 3.1, we have $\|\Delta_{\mathbf{A}_0^{(k)}}\|_2 = o_{\mathbb{P}}(1)$ by Corollary 1. Thus, $\|\Delta_{\Lambda^{(k)}}\|_2 = o_{\mathbb{P}}(1)$.

Using results from [11], the test statistic in (3.11) can be written as

$$TS = \frac{\mathbf{b}(\bar{\mathbf{Y}}^{(2)} - \bar{\mathbf{Y}}^{(1)})}{\sqrt{\hat{\sigma}^2_{\boldsymbol{\gamma}} \left[\mathbf{b}\left\{\frac{1}{n_1}\hat{\Lambda}^{(1)}(\hat{\Lambda}^{(1)})^T + \frac{1}{n_2}\hat{\Lambda}^{(2)}(\hat{\Lambda}^{(2)})^T\right\}\mathbf{b}^T\right] + \hat{\sigma}^2_{\boldsymbol{\varepsilon}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{b}\mathbf{b}^T}},$$

where $\bar{\mathbf{Y}}^{(k)}$ is the mean expression of genes in the experimental condition k. [11] show that TS is an asymptotically most powerful unbiased test for (3.10) when the correct network information is provided. Therefore, to establish the result in Theorem 3.1, it suffices to show that the denominator of TS is a consistent estimator.

In the following, we first consider the log-likelihood $l_F(\vartheta; \hat{\Lambda})$ based on the estimated networks $\hat{\Lambda} = (\hat{\Lambda}^{(1)}, \hat{\Lambda}^{(2)})$ and correct variance components $\vartheta = (\sigma^2_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\varepsilon}})$. We then establish that the maximum likelihood estimator $\hat{\vartheta}_{\hat{\Lambda}} \to_{\mathbb{P}} \vartheta$ as $\hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T \to_{\mathbb{P}} \Lambda^{(k)}(\Lambda^{(k)})^T$ for both k. Hence the denominator of TS is consistent and TS is an asymptotically most powerful unbiased test for (3.10).

Let $\hat{\mathbf{W}}^{(k)} = \sigma^2_{\boldsymbol{\gamma}}\hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_p$ for $k = 1, 2$. Up to a constant, the negative log-likelihood

$$l_F(\vartheta; \hat{\Lambda}) = \frac{n_1}{2n} l(\vartheta; \hat{\Lambda}^{(1)}) + \frac{n_2}{2n} l(\vartheta; \hat{\Lambda}^{(2)})$$

with

$$l(\vartheta; \hat{\Lambda}^{(1)}) = \text{logdet}(\hat{\mathbf{W}}^{(1)}) + \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{R}_j^T (\hat{\mathbf{W}}^{(1)})^{-1}\mathbf{R}_j,$$
$$l(\vartheta; \hat{\Lambda}^{(2)}) = \text{logdet}(\hat{\mathbf{W}}^{(2)}) + \frac{1}{n_2}\sum_{j=1+n_1}^{n} \mathbf{R}_j^T (\hat{\mathbf{W}}^{(2)})^{-1}\mathbf{R}_j,$$

where $\mathbf{R}_j = \mathbf{Y}_j^{(1)} - \bar{\mathbf{Y}}^{(1)}$ $(j = 1, \ldots, n_1)$ and $\mathbf{R}_j = \mathbf{Y}_j^{(2)} - \bar{\mathbf{Y}}^{(2)}$ $(j = 1 + n_1, \ldots, n)$. We treat $l(\vartheta; \hat{\Lambda}^{(1)})$ first. In particular, we can approximate $l(\vartheta; \hat{\Lambda}^{(1)})$ using its one-term Taylor expansion around $\mathbf{W}^{(1)}$

$$l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + \text{trace}\left\{\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})^T \Delta_{\mathbf{W}^{(1)}}\right\} + o(\|\Delta_{\mathbf{W}^{(1)}}\|_2^2),$$

where $\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})$ is the gradient of $l(\vartheta; \Lambda^{(1)})$ with respect to $\mathbf{W}^{(1)}$ and

$$\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)}) = (\mathbf{W}^{(1)})^{-1} - n_1^{-1}\sum_{j=1}^{n_1} (\mathbf{W}^{(1)})^{-1}\mathbf{R}_j\mathbf{R}_j^T(\mathbf{W}^{(1)})^{-1}.$$

Let $\Gamma = \Delta_{\mathbf{W}^{(1)}}/\|\Delta_{\mathbf{W}^{(1)}}\|_2$ and denote

$$g(\vartheta) = \operatorname{trace}\left\{\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})^T \Gamma\right\} = \operatorname{trace}\left\{(\mathbf{W}^{(1)})^{-1}\Gamma\right\} - n_1^{-1}\sum_{j=1}^{n_1} \mathbf{R}_j^T (\mathbf{W}^{(1)})^{-1}\Gamma(\mathbf{W}^{(1)})^{-1}\mathbf{R}_j.$$

then

$$l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + g(\vartheta)\|\Delta_{\mathbf{W}^{(1)}}\|_2 + o(\|\Delta_{\mathbf{W}^{(1)}}\|_2^2).$$

Using von Neumann's trace inequality [9], we can bound the first term in $g(\vartheta)$ by

$$\begin{aligned}
\left|\operatorname{trace}\left\{(\mathbf{W}^{(1)})^{-1}\Gamma\right\}\right| &\le \sum_{i=1}^{p} \varsigma_{[i]}((\mathbf{W}^{(1)})^{-1})\varsigma_{[i]}(\Gamma) \\
&\le p\varsigma_{[1]}\big((\sigma_{\boldsymbol{\gamma}}^2 \Lambda^{(1)}(\Lambda^{(1)})^T + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_p)^{-1}\big)\varsigma_{[1]}(\Gamma) \\
&= p\frac{1}{\phi_{\min}(\sigma_{\boldsymbol{\gamma}}^2 \Lambda^{(1)}(\Lambda^{(1)})^T + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_p)}\varsigma_{[1]}(\Gamma),
\end{aligned}$$

where $\varsigma_{[i]}(\mathbf{A})$ denotes the ith largest singular value of $\mathbf{A}$. By construction, $\varsigma_{[1]}(\Gamma) = 1$ and $\phi_{\min}(\sigma_{\boldsymbol{\gamma}}^2 \Lambda^{(1)}(\Lambda^{(1)})^T + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_p) \ge \sigma_{\boldsymbol{\varepsilon}}^2$. Hence $|\operatorname{trace}\{(\mathbf{W}^{(1)})^{-1}\Gamma\}| \le p/\sigma_{\boldsymbol{\varepsilon}}^2$. On the other hand, with probability tending to 1,

$$\begin{aligned}
n_1^{-1}\sum_{j=1}^{n_1} \mathbf{R}_j^T (\mathbf{W}^{(1)})^{-1}\Gamma(\mathbf{W}^{(1)})^{-1}\mathbf{R}_j &\le \|(\mathbf{W}^{(1)})^{-1}\Gamma(\mathbf{W}^{(1)})^{-1}\|_2 n_1^{-1}\sum_{j=1}^{n_1} \mathbf{R}_j^T\mathbf{R}_j \\
&\le \|(\mathbf{W}^{(1)})^{-1}\|_2^2\|\Gamma\|_2 n_1^{-1}\sum_{j=1}^{n_1} \mathbf{R}_j^T\mathbf{R}_j = \sigma_{\boldsymbol{\varepsilon}}^{-4}\mathbb{E}(\|\mathbf{R}_j\|_2^2),
\end{aligned}$$

where the last step follows from the strong law of large numbers. This implies that $g(\vartheta)$ is bounded for nontrivial $\sigma_{\boldsymbol{\varepsilon}}^2$. Note also $\Delta_{\mathbf{W}^{(1)}} = \hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)} = \sigma_{\boldsymbol{\gamma}}^2\{\hat{\Lambda}^{(1)}(\hat{\Lambda}^{(1)})^T - \Lambda^{(1)}(\Lambda^{(1)})^T\} = \sigma_{\boldsymbol{\gamma}}^2\Delta_{\Lambda^{(k)}}$. Hence $g(\vartheta)\|\Delta_{\mathbf{W}^{(1)}}\|_2 = g(\vartheta)\sigma_{\boldsymbol{\gamma}}^2\|\Delta_{\Lambda^{(k)}}\|_2 = o_{\mathbb{P}}(1)$. Therefore $l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + o_{\mathbb{P}}(1)$, and similarly one can show that $l(\vartheta; \hat{\Lambda}^{(2)}) = l(\vartheta; \Lambda^{(2)}) + o_{\mathbb{P}}(1)$. They together imply that

$$l_F(\vartheta; \hat{\Lambda}) = l_F(\vartheta; \Lambda) + o_{\mathbb{P}}(1).$$

Now conditioning on the event $\{l_F(\vartheta; \hat{\Lambda}) = l_F(\vartheta; \Lambda)\}$, the estimate of the variance components is $\hat{\vartheta} = \operatorname{argmin}_{\vartheta} l_F(\vartheta; \Lambda)$. Since $l_F(\vartheta; \Lambda)$ is convex with respect to ϑ, M-estimation results in [6] imply that $\mathbb{P}(\hat{\vartheta} = \vartheta) = 1$ and hence $\hat{\vartheta} \to_{\mathbb{P}} \vartheta$ as $\hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T \to_{\mathbb{P}} \Lambda^{(k)}(\Lambda^{(k)})^T$ for both k. It follows immediately that the denominator of the test statistic TS is a consistent estimator as $\hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T \to_{\mathbb{P}} \Lambda^{(k)}(\Lambda^{(k)})^T$ for both

k. This concludes the proof. □

B Efficient Estimation of Model Parameters

In this section, we present in details the strategy used to scale up the NetGSA algorithm for large scale networks as well as necessary derivations.

As pointed out in Section 2.2.1 of the main text, inference in NetGSA requires estimation of the mean parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and variance components $\sigma^2_{\boldsymbol{\gamma}}$ and $\sigma^2_{\boldsymbol{\varepsilon}}$. After rearranging the data $\mathcal{D}$ to be a $N \times 1$ vector $\mathbf{Y}$, we can write the model using the matrix notation as

$$\mathbf{Y} = \Psi\boldsymbol{\beta} + \Pi\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{B.1}$$

where the design matrix

$$\Pi = \mathrm{bdiag}(\Lambda^{(1)}, \ldots, \Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(2)}) \in \mathbf{R}^{N\times N},$$

and

$$\Psi = \begin{pmatrix} \Lambda^{(1)} & \\ \vdots & \\ \Lambda^{(1)} & \\ & \Lambda^{(2)} \\ & \vdots \\ & \Lambda^{(2)} \end{pmatrix} \in \mathbf{R}^{N\times 2p}.$$

The variance of $\mathbf{Y}$, i.e. $\mathbf{W} = \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N + \sigma^2_{\boldsymbol{\gamma}}\Pi\Pi'$. The mean $\boldsymbol{\beta}$ can be estimated via the maximum likelihood as

$$\hat{\boldsymbol{\beta}} = (\Psi'\hat{\mathbf{W}}^{-1}\Psi)^{-1}\Psi'\hat{\mathbf{W}}^{-1}\mathbf{Y},$$

where $\hat{\mathbf{W}}$ is defined using the estimated variances. The variances are often estimated via the maximum likelihood or restricted maximum likelihood using the profile likelihood. Thus, one can use an iterative algorithm to jointly estimate $\boldsymbol{\beta}$ and the variance components.

However, estimation of the variance components is computationally demanding for large networks. To ensure stability, the earlier version of the NetGSA considered profiling out one of the variance components and implemented an algorithm from [4], 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 [4] 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 NetGSA, we consider using Newton's method for estimating the variance parameters based on the profile log-likelihood to improve the computational stability. In particular, we make the following two key improvements for implementation of Newton's method.

First, it is clear that $\mathrm{Var}(\mathbf{Y}_j^{(k)}) = \sigma_\varepsilon^2 \left\{\mathbf{I}_p + \tau\Lambda^{(k)}(\Lambda^{(k)})^T\right\} = \sigma_\varepsilon^2\Sigma^{(k)}$, where $\tau = \sigma_\gamma^2/\sigma_\varepsilon^2$. Since the profile log-likelihood as well as its gradient and Hessian matrix with respect to τ all depend on $\Sigma^{(k)}$ ($k = 1, 2$) and their inverses, we choose to invert from their Cholesky decompositions $\Sigma^{(k)} = \mathbf{U}^T\mathbf{U}$, where $\mathbf{U}$ 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 $(\Sigma^{(k)})^{-1} = (\mathbf{U}^{-1})(\mathbf{U}^{-1})^T$. In the meantime, we also simplify the calculation of the determinant of $\Sigma^{(k)}$ since $\det(\Sigma^{(k)}) = \det(\mathbf{U})^2$, which is necessary for evaluating the profile log-likelihood.

Second, the 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 use a method-of-moment-type estimate of the variance components. Specifically, denote the residuals $\mathbf{R}_j = \mathbf{Y}_j^{(k)} - \Lambda^{(k)}\hat{\boldsymbol{\mu}}^{(k)}$ for $j = 1, \ldots, n$, where $\hat{\boldsymbol{\mu}}^{(k)}$ is the estimate of $\boldsymbol{\mu}^{(k)}$. Assume that there is a single variance σ_ε^2 that applies to all $\boldsymbol{\varepsilon}_j$ ($j = 1, \ldots, n$) and variances of $\boldsymbol{\gamma}_j$ are different. The variance of $\mathbf{R}_j$ can be decomposed as $(\sigma_\gamma^2)_j + \sigma_\varepsilon^2$. We then take the minimum of $\mathrm{Var}(\mathbf{R}_j)$ as the estimate of σ_ε^2 and average of the remaining variances as the estimate of σ_γ^2. Their ratio is used as the initial value for τ. The approximation runs very fast and does not add much computational cost to the method. To find the appropriate step sizes, we use backtracking line search as described in [3, page 464].

With the above two modifications, Newton's method can then be implemented to optimize the profile log-likelihood and returns an estimate of τ. Estimates of $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\varepsilon^2$ follow immediately. The implementation of Newton's method requires the gradient and the Hessian of the objective function, i.e., the profile log-likelihood. Next we provide details on how to calculate these quantities from the profile log-likelihood when profiling out σ_ε, based on the general framework introduced in [7]. The derivation follows similarly when profiling out σ_γ.

Let $N = np$ be the total number of observations for all genes. Recall that for $k = 1, 2$, $\Sigma^{(k)} = \mathbf{I}_p + \tau\Lambda^{(k)}(\Lambda^{(k)})^T$ with $\tau = \sigma_\gamma^2/\sigma_\varepsilon^2$. The residuals $\mathbf{R}_j = \mathbf{Y}_j^{(k)} - \Lambda^{(k)}\hat{\boldsymbol{\mu}}^{(k)}$ for $j = 1, \ldots, n$, where $\hat{\boldsymbol{\mu}}^{(k)}$ is the estimate of $\boldsymbol{\mu}^{(k)}$. Given the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ (with the first n_1 samples from condition 1 and

the remaining $n_2 = n - n_1$ samples from condition 2), the nonconstant part of the "full" log-likelihood l_F is

$$\begin{aligned} l_F(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = &- \frac{1}{2}\left\{ n_1 \operatorname{logdet}(\sigma_\varepsilon^2 \Sigma^{(1)}) + n_2 \operatorname{logdet}(\sigma_\varepsilon^2 \Sigma^{(2)}) \right\} \\ &- \frac{1}{2}\sigma_\varepsilon^{-2} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^{n} \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}. \end{aligned}$$

Similarly, the nonconstant part of the log-likelihood using the restricted maximum likelihood is

$$\begin{aligned} l_R(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = &l_F(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \\ &- \frac{1}{2} \operatorname{logdet}\left\{ n_1 \sigma_\varepsilon^{-2} (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)} + n_2 \sigma_\varepsilon^{-2} (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)} \right\}. \end{aligned}$$

We first solve for σ_ε^2 as a function of τ. The maximum likelihood estimate of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^{n} \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}, \tag{B.2}$$

whereas its restricted maximum likelihood estimate is

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N - 2p} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^{n} \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}. \tag{B.3}$$

Substituting σ_ε^2 with its corresponding estimate, we obtain the profile log-likelihood

$$\begin{aligned} p_F(\tau \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = &-\frac{1}{2}(n_1 \operatorname{logdet} \Sigma^{(1)} + n_2 \operatorname{logdet} \Sigma^{(2)}) \\ &- \frac{1}{2} N \log \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^{n} \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}, \end{aligned} \tag{B.4}$$

for maximum likelihood and

$$\begin{aligned} p_R(\tau \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = &-\frac{1}{2}(n_1 \operatorname{logdet} \Sigma^{(1)} + n_2 \operatorname{logdet} \Sigma^{(2)}) \\ &- \frac{1}{2}(N - 2p) \log \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^{n} \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\} \\ &- \frac{1}{2} \operatorname{logdet}\left\{ n_1 (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)} + n_2 (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)} \right\}, \end{aligned} \tag{B.5}$$

for restricted maximum likelihood.

As $\Sigma^{(k)}$ $(k = 1, 2)$ are the only terms that depend on τ, we first look at the derivatives of $\text{logdet}\,\Sigma^{(k)}$, $\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\mathbf{R}_j$, and $\text{logdet}(\mathbf{H})$ with respect to τ, where $\mathbf{H} = n_1\mathbf{H}^{(1)}+n_2\mathbf{H}^{(2)}$ and $\mathbf{H}^{(k)} = (\Lambda^{(k)})^T(\Sigma^{(k)})^{-1}\Lambda^{(k)}$ for $k = 1, 2$. Let

$$\mathbf{B}^{(k)} = (\Sigma^{(k)})^{-1}\frac{d\Sigma^{(k)}}{d\tau}(\Sigma^{(k)})^{-1}.$$

Then

$$\frac{d\,\text{logdet}(\Sigma^{(k)})}{d\tau} = \text{trace}\left\{(\Sigma^{(k)})^{-1}\frac{d\Sigma^{(k)}}{d\tau}\right\},$$

$$\frac{d^2\,\text{logdet}(\Sigma^{(k)})}{d\tau^2} = \text{trace}\left\{-(\mathbf{B}^{(k)})^T\frac{d\Sigma^{(k)}}{d\tau} + (\Sigma^{(k)})^{-1}\frac{d^2\Sigma^{(k)}}{d\tau^2}\right\},$$

$$\frac{d\,\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\mathbf{R}_j}{d\tau} = -\mathbf{R}_j^T\mathbf{B}^{(k)}\mathbf{R}_j, \quad \frac{d^2\,\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\mathbf{R}_j}{d\tau^2} = -\mathbf{R}_j^T\frac{d\mathbf{B}^{(k)}}{d\tau}\mathbf{R}_j,$$

$$\frac{d\,\text{logdet}(\mathbf{H})}{d\tau} = -\,\text{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k(\Lambda^{(k)})^T\mathbf{B}^{(k)}\Lambda^{(k)}\right\},$$

and

$$\frac{d^2\,\text{logdet}(\mathbf{H})}{d\tau^2} = -\,\text{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k(\Lambda^{(k)})^T\mathbf{B}^{(k)}\Lambda^{(k)} \times \mathbf{H}^{-1}\sum_{k=1,2} n_k(\Lambda^{(k)})^T\mathbf{B}^{(k)}\Lambda^{(k)}\right\}$$
$$-\,\text{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k(\Lambda^{(k)})^T\frac{d\mathbf{B}^{(k)}}{d\tau}\Lambda^{(k)}\right\},$$

where

$$\frac{d\mathbf{B}^{(k)}}{d\tau} = -(\Sigma^{(k)})^{-1}\left\{2\frac{d\Sigma^{(k)}}{d\tau}(\Sigma^{(k)})^{-1}\frac{d\Sigma^{(k)}}{d\tau} - \frac{d^2\Sigma^{(k)}}{d\tau^2}\right\}(\Sigma^{(k)})^{-1}.$$

Given the covariance $\Sigma^{(k)}$ $(k = 1, 2)$ defined in Section 3.1, we can further simplify the above derivatives and obtain

$$\frac{d\,\text{logdet}\,\Sigma^{(k)}}{d\tau} = \text{trace}\left\{\mathbf{H}^{(k)}\right\}, \quad \frac{d^2\,\text{logdet}\,\Sigma^{(k)}}{d\tau^2} = -\,\text{trace}\left\{\mathbf{H}^{(k)}\mathbf{H}^{(k)}\right\},$$

$$\frac{d\,\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\mathbf{R}_j}{d\tau} = -\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\Lambda^{(k)}(\Lambda^{(k)})^T(\Sigma^{(k)})^{-1}\mathbf{R}_j,$$

$$\frac{d^2\,\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\mathbf{R}_j}{d\tau^2} = 2\mathbf{R}_j^T(\Sigma^{(k)})^{-1}\Lambda^{(k)}\mathbf{H}^{(k)}(\Lambda^{(k)})^T(\Sigma^{(k)})^{-1}\mathbf{R}_j,$$

$$\frac{d\,\mathrm{logdet}(\mathbf{H})}{d\tau} = -\,\mathrm{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k \mathbf{H}^{(k)}\mathbf{H}^{(k)}\right\},$$

$$\frac{d^2\,\mathrm{logdet}(\mathbf{H})}{d\tau^2} = -\,\mathrm{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k \mathbf{H}^{(k)}\mathbf{H}^{(k)}\right\} + 2\,\mathrm{trace}\left\{\mathbf{H}^{-1}\sum_{k=1,2} n_k \mathbf{H}^{(k)}\mathbf{H}^{(k)}\mathbf{H}^{(k)}\right\}.$$

With the above quantities, one can then calculate the gradient and Hessian of the profile log-likelihood p_R for restricted maximum likelihood and use Newton's method to obtain an estimate of τ. Estimate of $\hat{\sigma}^2_{\varepsilon}$ is calculated from (B.3), and $\hat{\sigma}^2_{\gamma} = \hat{\tau}\hat{\sigma}^2_{\varepsilon}$. Estimation with maximum likelihood follows similarly by applying Newton's method to p_F and utilizing (B.2).

C Additional Simulation Results

To benchmark the performance of the proposed network estimation procedure as well as NetGSA, we first revisit the two simulation experiments presented in Section 3 of the main paper and report the Type I error (or the observed false discovery proportion) when the null hypothesis is true. In addition, we consider two other simulation experiments and refer to them as the third and fourth settings, following the earlier two settings in the main paper. The simulations in this section are also discussed when comparing the run time of NetGSA with different variance estimation algorithms in Section 3 of the main paper.

C.1 Simulation Studies 1 and 2

C.1.1 Powers

We have shown the estimated powers in Tables 2 and 3 in the main paper for the two experiments based on adjusted false discovery rate (FDR) cutoffs. For completeness, we present here the estimated powers in Tables A1 and A2 when the FDR cutoff is $q^* = 0.05$. Due to the use of different FDR cutoffs, one expects to see higher powers for the columns corresponding to 0.2, 0.8, 0.2(m) and 0.8(m), and slightly lower powers for E and GSA-c in Tables A1 and A2 compared to, respectively, Tables 2 and 3 in the main paper. In both Table A1 and Table A2, we still observe the following: NetGSA with the exact networks does a very good job in recovering the true powers for each pathway; NetGSA with more external structural information generally reports powers that are closer to the true power; further NetGSA is robust to misspecification in external structural information. Further, for pathway 1 that has neither mean nor structural changes, we note that the powers are sometimes greater than 0.05 when NetGSA with estimated network information is applied. This is partly due to the network estimation error.

Table A1: Powers with false discovery rate cutoff $q^* = 0.05$ in experiment 1. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information.

					$p = 100$			
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.05	0.10	0.03	0.06	0.17	0.01	0.05	0.08
2	0.18	0.13	0.03	0.06	0.09	0.00	0.16	0.16
3	0.50	0.48	0.30	0.46	0.36	0.00	0.50	0.63
4	0.49	0.29	0.02	0.07	0.25	0.04	0.44	0.29
5	0.94	0.98	0.89	0.97	0.97	0.00	0.95	0.95
6	0.46	0.49	0.20	0.26	0.36	0.00	0.49	0.41
7	0.84	0.90	0.94	0.99	0.98	0.04	0.82	0.92
8	0.54	0.68	0.42	0.57	0.87	0.00	0.56	0.61

Table A2: Powers with false discovery rate cutoff $q^* = 0.05$ in experiment 2. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information.

					$p = 160$			
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.10	0.11	0.02	0.05	0.07	0.01	0.10	0.11
2	0.52	0.60	0.15	0.36	0.51	0.00	0.53	0.60
3	0.96	1.00	0.95	0.99	1.00	0.00	0.97	1.00
4	0.98	1.00	1.00	1.00	1.00	0.09	0.98	1.00
5	0.38	0.34	0.02	0.11	0.10	0.03	0.41	0.36
6	0.46	0.35	0.01	0.07	0.24	0.01	0.46	0.34
7	0.78	0.83	0.89	0.92	0.99	0.00	0.78	0.82
8	0.92	0.99	1.00	1.00	1.00	0.02	0.91	0.98

C.1.2 Type I errors

To validate the type I error when the null hypothesis is true, we use the same null setup as presented in Section 3 of the main paper for both experiment 1 and 2. The network structure and node mean expressions under the alternative are set to be the same as in the null case. We use $n_1 = n_2 = 25$ samples for each condition in experiment 1 and $n_1 = n_2 = 40$ in experiment 2 for pathway enrichment analysis. When the underlying networks are not available, we estimate the networks based on external information ranging from 0%, 20%, 80% to 100% and 100 observations generated from the true network. Scenarios with misspecified structural information are also considered. In the following, we present type I errors based on both the adjusted FDR cutoffs and uniform FDR cutoff at $q^* = 0.05$, where the former corresponds to using $q^* = 0.01$ for cases 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

Table A3 and A4 present the type I errors based on adjusted FDR cutoffs evaluated over 100 replications for experiment 1 and 2, respectively. As expected, the type I errors when all true parameters are plugged in the NetGSA model are 0.05 for all subnetworks in both experiments. When the exact networks are known, one only estimates the variance components in the NetGSA model and observes small false discovery proportions. When the exact networks are not available such that one estimates the partial correlations as well as the variance components, the type I errors are generally greater than q^*; in particular, the type I errors get worse as the amount of external information decreases. This is likely due to the small sample sizes for estimating the networks. In general, one benefits from having more external structural information and/or more observations for recovering the underlying networks when using NetGSA. In comparison, both GSA-c and GSA-s have type I errors smaller than 0.05.

Table A3: Type I error when the null hypothesis is true in experiment 1. False discovery rate cutoffs are $q^* = 0.01$ for cases 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.00	0.00	0.00	0.05	0.02	0.03	0.00	0.01
2	0.06	0.06	0.01	0.05	0.03	0.05	0.07	0.02
3	0.29	0.16	0.02	0.05	0.01	0.02	0.28	0.14
4	0.16	0.12	0.00	0.05	0.03	0.05	0.17	0.11
5	0.02	0.01	0.01	0.05	0.00	0.03	0.03	0.01
6	0.24	0.12	0.00	0.05	0.01	0.02	0.23	0.12
7	0.18	0.13	0.03	0.05	0.03	0.05	0.19	0.14
8	0.11	0.10	0.00	0.05	0.02	0.02	0.09	0.09

Table A4: Type I error when the null hypothesis is true in experiment 2. False discovery rate cutoffs are $q^* = 0.01$ for cases 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.03	0.03	0.00	0.05	0.01	0.03	0.03	0.03
2	0.08	0.05	0.00	0.05	0.01	0.05	0.08	0.05
3	0.17	0.09	0.00	0.05	0.01	0.01	0.17	0.08
4	0.34	0.25	0.01	0.05	0.03	0.03	0.35	0.25
5	0.25	0.09	0.00	0.05	0.01	0.02	0.27	0.11
6	0.32	0.18	0.00	0.05	0.02	0.04	0.33	0.18
7	0.23	0.13	0.00	0.05	0.00	0.04	0.28	0.13
8	0.26	0.12	0.00	0.05	0.02	0.01	0.29	0.11

As a comparison, Table A5 and A6 present the type I errors based on the uniform FDR cutoffs 0.05 evaluated over 100 replications for experiment 1 and 2, respectively. The reported false discovery proportions for NetGSA with estimated networks including columns 0.2, 0.8, 0.2(m) and 0.8(m) are generally higher than the corresponding columns in Table A3 and A4, especially pathways 2-4 and 6-8.

Table A5: Type I error when the null hypothesis is true in experiment 1. False discovery rate cutoff is $q^* = 0.05$.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.03	0.04	0.00	0.05	0.01	0.01	0.04	0.03
2	0.18	0.18	0.01	0.05	0.04	0.02	0.18	0.17
3	0.32	0.25	0.00	0.05	0.02	0.00	0.36	0.25
4	0.39	0.29	0.01	0.05	0.03	0.00	0.39	0.32
5	0.09	0.04	0.00	0.05	0.02	0.05	0.11	0.07
6	0.30	0.19	0.00	0.05	0.04	0.03	0.32	0.21
7	0.39	0.29	0.01	0.05	0.04	0.02	0.39	0.32
8	0.29	0.19	0.02	0.05	0.04	0.03	0.27	0.21

Table A6: Type I error when the null hypothesis is true in experiment 2. False discovery rate cutoff is $q^* = 0.05$.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2m	0.8m
1	0.15	0.14	0.00	0.05	0.01	0.03	0.13	0.15
2	0.14	0.12	0.00	0.05	0.00	0.01	0.15	0.12
3	0.25	0.17	0.00	0.05	0.01	0.01	0.26	0.17
4	0.33	0.26	0.00	0.05	0.04	0.01	0.32	0.28
5	0.46	0.32	0.00	0.05	0.01	0.02	0.42	0.32
6	0.43	0.26	0.00	0.05	0.02	0.03	0.44	0.26
7	0.40	0.30	0.00	0.05	0.02	0.02	0.40	0.32
8	0.47	0.29	0.00	0.05	0.01	0.01	0.46	0.29

It is important to make a distinction between the samples used for enrichment analysis and those for network estimation. If one has access to a large number of observations that can only be used for network estimation as well as some external structural information, then NetGSA can leverage both resources to achieve more reliable enrichment testing. However, methods like GSA are unable to take advantage of such rich external information.

C.2 Simulation Studies 3 and 4

C.2.1 The setup

Our third simulation experiment considers an undirected network with $p = 160$ and a design similar to the second experiment. However, in this case, each of the 8 subnetworks has a denser structure and there are more interactions between subnetworks. Specifically, there are 70 edges connecting the 20 nodes in each subnetwork under the null. There is 30% chance of an interaction between four randomly selected nodes from each subnetwork to four randomly selected nodes from every other subnetwork. Under the alternative, there is an increase of 0.5 in mean values for varying proportions of nodes (0%, 30% and 50%)

for subnetworks 1-3 and 5-7. For subnetworks 4 and 8, 70% of the nodes have mean values decreased by 0.5. Moreover, 13% of the edges in subnetworks 5-8 under the alternative are different from their null counterpart.

The fourth experiment uses an undirected network of size $p = 400$ and illustrates the scalability of the proposed method using the new optimization algorithm. The network consists of 20 subnetworks, each corresponding to a pathway with 20 genes. The probability of an interaction between the hub node in one subnetwork and two randomly selected nodes from another subnetwork is 0.4. Under the null, all subnetworks have the same topology generated from a scale-free random graph such that there are 37 edges linking the 20 nodes; all the nodes have mean expression values 1. Under the alternative, subnetworks 1-6 and 11-16 keep the same mean expression values, but 20%, 40%, 60% and 80% of nodes in subnetworks 7-10 and 17-20 have 0.5 unit increase in their mean values, respectively. In addition, subnetworks 11-20 under the alternative all have 39 edges and their structure differs from their null equivalent by 30%. This experiment is also of interest because we created a setting where there are enough subnetworks in order for the permutation based Gene Set Analysis [5] to calibrate the number of permutations required.

In both experiments, we also included scenarios where a proportion of the supplied structural information is incorrectly specified. This is to check whether NetGSA is robust to model misspecification. In particular, about 50% (20%) of the supplied edges are actually not present in the true model for the case $r = 0.2$ ($r = 0.8$).

C.2.2 Network estimation

Table A7 presents the deviance measures for estimating the networks with 100 replicates and sample sizes of $m = 500$ for $p = 160$ and $m = 400$ for $p = 400$, 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 correctly specified structural information of the networks r increases ($r = 0.2, 0.8$ corresponding to 20% and 80% total information, respectively). When there exists misspecified edges in the external information (denoted by 0.2(m) and 0.8(m)), we used two tuning parameters for network estimation, one for controlling the overall sparsity of the network and the other for correcting the misspecified edges. The optimal tuning parameters were selected over a grid of values using BIC. It can be seen that the performance of network estimation is not compromised by much after properly selected tuning parameters.

Table A7: Deviance measures for network estimation in experiment 3 and 4. FPR(%), false positive rate in percentage; FNR(%), false negative rate in percentage; MCC, Matthews correlation coefficient; Fnorm, Frobenius norm loss.

		$p = 160$				$p = 400$			
	r	FPR(%)	FNR(%)	MCC	Fnorm	FPR(%)	FNR(%)	MCC	Fnorm
Null	0.0	8.20	12.38	0.59	0.58	4.19	14.58	0.44	0.50
	0.2	7.21	14.88	0.60	0.58	3.89	12.29	0.46	0.48
	0.8	3.04	5.47	0.80	0.47	1.88	3.87	0.65	0.40
	0.2(m)	7.48	12.83	0.60	0.57	3.89	12.44	0.46	0.49
	0.8(m)	3.07	8.39	0.78	0.49	1.88	4.23	0.64	0.40
Alternative	0.0	8.16	11.62	0.59	0.57	4.25	14.70	0.44	0.50
	0.2	7.15	14.41	0.60	0.57	3.96	12.38	0.46	0.48
	0.8	3.02	5.24	0.80	0.46	1.95	3.75	0.64	0.40
	0.2(m)	7.42	12.60	0.60	0.56	3.97	12.57	0.46	0.48
	0.8(m)	3.03	8.17	0.78	0.48	1.95	4.13	0.64	0.40

C.2.3 Powers

Table A8 shows the estimated powers after correcting for false discovery rate in the third experiment with $p = 160$. When the exact networks are known, NetGSA estimated powers match very well with the true powers. In the case of unknown networks, we see consistent recovery of high powers for subnetworks 3, 4, 6, 7 and 8 using NetGSA even with only 20% external information. This suggests that, with large enough samples for network estimation, a small amount of external knowledge is sufficient for making reliable inference using the network-based method. Interestingly, GSA-c identifies only subnetworks 3 and 4 as significantly differential with high power, whereas GSA-s returns relatively high power for subnetworks 3 and 6 but surprisingly low power for subnetwork 8. One possible reason for this pattern is that the busy interactions between subnetworks and the negative mean changes in subnetworks 4 and 8 affected the ability of GSA to properly recognize the correct differential behavior. The last two columns in Table A8 show the estimated powers from NetGSA when the external information is misspecified. For both cases ($r = 20\%$ and $r = 80\%$), the results bear high similarity to those in the first two columns, which suggests that NetGSA is robust to model misspecification.

The estimated powers after correcting for false discovery rate in the fourth experiment are shown separately in Table A10. When the exact networks with the correct edge weights are known, we again see that NetGSA estimated powers match the true powers closely, with very low powers for subnetworks 1-6 which have no changes in neither mean expressions nor structures, high powers for subnetworks 8-10 which have significant changes in mean expression values, low powers for subnetworks 11-16 that have changes in

Table A8: Powers in experiment 3. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information. False discovery rate cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m), 0.8(m) and GSA-s, 0.10 for E and GSA-c.

				$p = 160$				
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.00	0.00	0.05	0.05	0.06	0.04	0.00	0.00
2	0.49	0.50	0.77	0.71	0.59	0.04	0.49	0.52
3	0.75	0.71	0.91	0.90	0.97	0.76	0.78	0.72
4	0.90	0.88	1.00	0.99	0.59	0.97	0.89	0.87
5	0.18	0.12	0.07	0.05	0.48	0.03	0.17	0.13
6	0.47	0.44	0.66	0.68	0.78	0.27	0.48	0.44
7	0.68	0.62	0.96	0.95	0.64	0.04	0.69	0.60
8	0.76	0.77	0.99	1.00	0.14	0.50	0.77	0.75

Table A9: Powers in experiment 3. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information. False discovery rate cutoffs are $q^* = 0.05$.

				$p = 160$				
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.02	0.02	0.01	0.05	0.02	0.00	0.02	0.02
2	0.70	0.72	0.67	0.71	0.65	0.02	0.72	0.75
3	0.81	0.78	0.87	0.90	0.97	0.39	0.80	0.80
4	0.95	0.92	0.99	0.99	0.53	0.79	0.94	0.92
5	0.30	0.32	0.03	0.05	0.44	0.00	0.27	0.31
6	0.56	0.65	0.63	0.68	0.88	0.11	0.57	0.66
7	0.78	0.74	0.93	0.95	0.72	0.06	0.78	0.76
8	0.90	0.87	1.00	1.00	0.12	0.27	0.88	0.88

structures and very high powers for pathways 17-20 with changes in both. When there is 20% external information on the underlying network topology, NetGSA's powers for subnetworks 8-10 and 18-20 are close to true powers. However, NetGSA overestimates the powers for subnetworks 11-16. This is due to the small sample size ($m = 400$) for estimating the underlying networks. When the external information is slightly misspecified, the last two columns indicate that NetGSA still returns valid powers that are comparable to those obtained with correctly specified structural information. In comparison, GSA-s believes almost all subnetworks except 1-6 are significantly differential. On the other hand, when testing against the competitive null, the results from GSA-c suggest that only subnetworks 10, 19 and 20 are significantly differential. The conflicting results from GSA with or without randomization of the genes also raise concerns as to which version to choose in practice.

Table A10: Powers in experiment 4. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information. False discovery rate cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

				$p = 400$				
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.02	0.02	0.05	0.05	0.09	0.03	0.02	0.02
2	0.14	0.12	0.06	0.05	0.13	0.03	0.15	0.12
3	0.19	0.18	0.05	0.05	0.11	0.01	0.18	0.18
4	0.29	0.24	0.03	0.05	0.15	0.04	0.30	0.25
5	0.32	0.29	0.05	0.05	0.15	0.02	0.32	0.29
6	0.46	0.44	0.05	0.05	0.18	0.03	0.44	0.43
7	0.56	0.50	0.85	0.83	0.93	0.00	0.58	0.52
8	0.75	0.73	1.00	1.00	1.00	0.00	0.75	0.73
9	0.92	0.89	1.00	1.00	1.00	0.19	0.93	0.89
10	0.98	0.99	1.00	1.00	1.00	1.00	0.98	0.99
11	0.49	0.46	0.09	0.06	1.00	0.00	0.47	0.48
12	0.57	0.60	0.07	0.07	1.00	0.00	0.54	0.59
13	0.59	0.57	0.07	0.05	1.00	0.05	0.59	0.57
14	0.58	0.63	0.12	0.07	0.99	0.03	0.59	0.63
15	0.58	0.66	0.07	0.07	0.99	0.02	0.57	0.66
16	0.60	0.50	0.08	0.07	1.00	0.03	0.57	0.51
17	0.65	0.68	0.90	0.89	1.00	0.02	0.63	0.68
18	0.73	0.81	1.00	1.00	1.00	0.34	0.74	0.80
19	0.85	0.85	1.00	1.00	1.00	1.00	0.85	0.86
20	0.86	0.88	1.00	1.00	1.00	1.00	0.86	0.87

C.2.4 Type I errors

Finally, we also look at the scenarios where the null hypothesis is true for experiment 3 and 4. Again we present type I errors obtained based on both the adjusted FDR cutoffs and the uniform FDR cutoffs at 0.05, where the adjusted FDR cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c. The results can be found in Tables A12, A13, A14 and A15.

Since the sample size used for network estimation in experiment 3 is sufficiently large, we observe very good control of type I errors in Table A12 for NetGSA, even with estimated networks. In Table A14, type I errors are high for some pathways, which is again due to the small sample size for estimating 400×400 networks.

Table A11: Powers in experiment 4. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-c/GSA-s refer to Gene Set Analysis with/without randomization of the genes in 1000 permutations, respectively; 0.2(m)/0.8(m) refer to NetGSA with 20%/80% misspecified external information. False discovery rate cutoffs are $q^* = 0.05$.

	$p = 400$							
Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2(m)	0.8(m)
1	0.10	0.11	0.04	0.05	0.10	0.03	0.10	0.11
2	0.25	0.20	0.04	0.05	0.12	0.03	0.25	0.20
3	0.37	0.34	0.02	0.05	0.07	0.00	0.36	0.31
4	0.36	0.38	0.05	0.05	0.19	0.03	0.36	0.38
5	0.57	0.51	0.03	0.05	0.08	0.01	0.58	0.51
6	0.48	0.45	0.05	0.05	0.11	0.01	0.48	0.45
7	0.58	0.58	0.69	0.83	0.83	0.00	0.56	0.59
8	0.84	0.82	1.00	1.00	1.00	0.00	0.85	0.83
9	0.94	0.93	1.00	1.00	1.00	0.11	0.93	0.93
10	0.98	0.97	1.00	1.00	1.00	0.91	0.98	0.97
11	0.65	0.66	0.01	0.06	1.00	0.00	0.67	0.64
12	0.64	0.68	0.06	0.07	1.00	0.01	0.64	0.67
13	0.63	0.58	0.00	0.05	1.00	0.03	0.63	0.58
14	0.71	0.72	0.04	0.07	0.99	0.00	0.69	0.72
15	0.67	0.69	0.03	0.07	1.00	0.02	0.66	0.69
16	0.68	0.61	0.02	0.07	1.00	0.00	0.69	0.62
17	0.68	0.68	0.76	0.89	1.00	0.00	0.68	0.69
18	0.74	0.78	0.99	1.00	1.00	0.08	0.78	0.78
19	0.93	0.94	1.00	1.00	1.00	0.92	0.91	0.94
20	0.90	0.92	1.00	1.00	1.00	0.99	0.92	0.92

D Additional Results on Metabolomics and Genomics

Table A16, A17 and A18 present the full list of pathways used in each of the studies and their corresponding false discovery rate corrected p-values, respectively.

References

[1] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2 2008.

[2] Peter J. Bickel, Ya'Acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Table A12: Type I error when the null hypothesis is true in experiment 3. False discovery rate cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2m	0.8m
1	0.00	0.00	0.01	0.05	0.03	0.02	0.00	0.00
2	0.01	0.00	0.00	0.05	0.02	0.04	0.01	0.00
3	0.04	0.05	0.00	0.05	0.01	0.04	0.04	0.03
4	0.07	0.07	0.01	0.05	0.02	0.04	0.09	0.08
5	0.11	0.11	0.01	0.05	0.02	0.04	0.11	0.11
6	0.09	0.11	0.02	0.05	0.04	0.05	0.07	0.10
7	0.13	0.12	0.00	0.05	0.03	0.03	0.12	0.11
8	0.13	0.10	0.00	0.05	0.02	0.03	0.14	0.11

Table A13: Type I error when the null hypothesis is true in experiment 3. False discovery rate cutoff is $q^* = 0.05$.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2m	0.8m
1	0.05	0.03	0.00	0.05	0.04	0.01	0.05	0.03
2	0.09	0.08	0.00	0.05	0.03	0.04	0.10	0.09
3	0.18	0.17	0.00	0.05	0.05	0.02	0.18	0.17
4	0.16	0.15	0.01	0.05	0.02	0.01	0.15	0.13
5	0.23	0.18	0.00	0.05	0.07	0.02	0.21	0.18
6	0.21	0.17	0.01	0.05	0.05	0.02	0.18	0.19
7	0.18	0.16	0.01	0.05	0.02	0.01	0.17	0.16
8	0.25	0.19	0.01	0.05	0.02	0.05	0.22	0.19

[5] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1):107–129, 2007.

[6] Shelby J Haberman. Concavity and estimation. *Annals of Statistics*, 17(4):1631–1661, 1989.

[7] M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 12 1988.

[8] Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[9] Leon Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975.

[10] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[11] Ali Shojaie and George Michailidis. Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 22, 2010.

[12] Shuheng Zhou, Philipp Rutimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026, 2011.

Table A14: Type I error when the null hypothesis is true in experiment 4. False discovery rate cutoffs are $q^* = 0.01$ for 0.2, 0.8, 0.2(m) and 0.8(m), 0.05 for GSA-s and 0.10 for E and GSA-c.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2m	0.8m
1	0.03	0.03	0.02	0.05	0.03	0.02	0.03	0.03
2	0.21	0.18	0.00	0.05	0.02	0.00	0.21	0.18
3	0.24	0.26	0.00	0.05	0.04	0.01	0.25	0.26
4	0.25	0.29	0.02	0.05	0.03	0.01	0.26	0.28
5	0.34	0.35	0.02	0.05	0.01	0.01	0.35	0.34
6	0.41	0.40	0.01	0.05	0.03	0.02	0.42	0.41
7	0.37	0.40	0.01	0.05	0.02	0.00	0.40	0.39
8	0.43	0.40	0.02	0.05	0.02	0.02	0.41	0.39
9	0.48	0.52	0.01	0.05	0.02	0.01	0.50	0.52
10	0.45	0.52	0.01	0.05	0.02	0.02	0.44	0.51
11	0.45	0.47	0.00	0.05	0.03	0.00	0.45	0.48
12	0.50	0.50	0.03	0.05	0.02	0.06	0.51	0.49
13	0.55	0.51	0.00	0.05	0.04	0.01	0.55	0.52
14	0.63	0.65	0.01	0.05	0.03	0.00	0.60	0.65
15	0.52	0.50	0.02	0.05	0.02	0.02	0.52	0.50
16	0.56	0.54	0.02	0.05	0.02	0.04	0.57	0.52
17	0.63	0.67	0.02	0.05	0.03	0.00	0.64	0.66
18	0.58	0.55	0.01	0.05	0.03	0.02	0.57	0.57
19	0.59	0.59	0.01	0.05	0.02	0.04	0.58	0.60
20	0.53	0.51	0.01	0.05	0.01	0.03	0.55	0.52

Table A15: Type I error when the null hypothesis is true in experiment 4. False discovery rate cutoff is $q^* = 0.05$.

Pathway	0.2	0.8	E	T	GSA-s	GSA-c	0.2m	0.8m
1	0.03	0.04	0.00	0.05	0.02	0.00	0.03	0.03
2	0.27	0.27	0.00	0.05	0.04	0.00	0.27	0.26
3	0.36	0.36	0.00	0.05	0.04	0.00	0.35	0.35
4	0.39	0.40	0.01	0.05	0.04	0.02	0.40	0.41
5	0.51	0.49	0.01	0.05	0.04	0.00	0.52	0.51
6	0.59	0.59	0.00	0.05	0.02	0.00	0.61	0.58
7	0.56	0.54	0.00	0.05	0.02	0.02	0.55	0.53
8	0.61	0.60	0.01	0.05	0.04	0.04	0.58	0.58
9	0.59	0.60	0.00	0.05	0.02	0.02	0.59	0.59
10	0.61	0.59	0.00	0.05	0.04	0.00	0.58	0.62
11	0.60	0.58	0.00	0.05	0.04	0.01	0.61	0.58
12	0.63	0.60	0.00	0.05	0.03	0.00	0.60	0.60
13	0.64	0.64	0.01	0.05	0.02	0.02	0.62	0.67
14	0.56	0.67	0.03	0.05	0.02	0.03	0.56	0.68
15	0.73	0.76	0.00	0.05	0.02	0.00	0.71	0.77
16	0.64	0.60	0.00	0.05	0.02	0.01	0.63	0.63
17	0.68	0.64	0.00	0.05	0.02	0.01	0.64	0.62
18	0.64	0.64	0.00	0.05	0.04	0.00	0.58	0.63
19	0.70	0.60	0.00	0.05	0.02	0.01	0.66	0.60
20	0.72	0.73	0.00	0.05	0.04	0.00	0.71	0.73

Table A16: p-values after false discovery rate correction for all pathways in the metabolomics data

Pathway	NetGSA	GSA-s	GSA-c
Tryptophan metabolism	$3e^{-5}$	0.00	1.00
beta-Alanine metabolism	$3e^{-5}$	0.00	1.00
Aminoacyl-tRNA biosynthesis	$2e^{-4}$	0.00	1.00
ABC transporters	$4e^{-4}$	0.00	1.00
Fatty acid biosynthesis	$2e^{-3}$	1.00	1.00
Pyrimidine metabolism	$2e^{-3}$	0.00	1.00
Phenylalanine metabolism	$4e^{-3}$	0.00	1.00
Pantothenate and CoA biosynthesis	0.01	0.00	1.00
Phenylalanine, tyrosine and tryptophan biosynthesis	0.02	1.00	1.00
Caffeine metabolism	0.04	0.15	1.00
Glycine, serine and threonine metabolism	0.15	$4e^{-3}$	1.00
Lysine biosynthesis	0.19	1.00	1.00
Methionine metabolism	0.20	1.00	1.00
Histidine metabolism	0.26	0.00	0.42
Propanoate metabolism	0.34	0.04	1.00
Arginine and proline metabolism	0.39	0.06	1.00
Glutathione metabolism	0.43	0.12	1.00
Arginine biosynthesis	0.47	0.01	1.00
Alanine and aspartate metabolism	0.57	1.00	1.00
Valine, leucine and isoleucine biosynthesis	0.61	1.00	1.00
Purine metabolism	1.00	0.03	1.00
Glutamate metabolism	1.00	1.00	1.00
Tyrosine metabolism	1.00	1.00	1.00
Cyanoamino acid metabolism	1.00	1.00	1.00
Nitrogen metabolism	1.00	0.43	1.00
Tropane, piperidine and pyridine alkaloid biosynthesis	1.00	0.02	1.00
Neuroactive ligand-receptor interaction	1.00	0.02	1.00

Table A17: p-values after false discovery rate correction for all pathways in the Lung cancer data

Pathway	NetGSA	GSA-s	GSA-c
Jak-STAT signaling pathway	0.18	0.31	1.00
p53 signaling pathway	0.22	0.68	1.00
Wnt signaling pathway	0.28	0.61	1.00
mTOR signaling pathway	0.42	0.46	1.00
Glutathione metabolism	0.42	1.00	1.00
Purine metabolism	0.49	0.46	1.00
Cysteine and methionine metabolism	0.49	0.46	1.00
ErbB signaling pathway	0.74	0.07	1.00
Chemokine signaling pathway	0.74	0.61	1.00
MAPK signaling pathway	0.77	0.61	1.00
Pentose phosphate pathway	0.82	1.00	1.00
Pyrimidine metabolism	0.83	0.46	1.00
Cell cycle	0.87	0.80	1.00
Glycolysis / Gluconeogenesis	1.00	0.98	1.00
Citrate cycle (TCA cycle)	1.00	1.00	1.00
Fructose and mannose metabolism	1.00	1.00	1.00
Galactose metabolism	1.00	1.00	1.00
Fatty acid metabolism	1.00	1.00	1.00
Oxidative phosphorylation	1.00	1.00	1.00
Alanine, aspartate and glutamate metabolism	1.00	1.00	1.00
Valine, leucine and isoleucine degradation	1.00	1.00	1.00
Lysine degradation	1.00	1.00	1.00
Arginine and proline metabolism	1.00	1.00	1.00
Histidine metabolism	1.00	1.00	1.00
Tyrosine metabolism	1.00	1.00	1.00
Tryptophan metabolism	1.00	1.00	1.00
beta-Alanine metabolism	1.00	1.00	1.00
Starch and sucrose metabolism	1.00	0.61	1.00
Amino sugar and nucleotide sugar metabolism	1.00	1.00	1.00
PPAR signaling pathway	1.00	1.00	1.00
Calcium signaling pathway	1.00	1.00	1.00
Phosphatidylinositol signaling system	1.00	1.00	1.00
Notch signaling pathway	1.00	0.68	1.00
Hedgehog signaling pathway	1.00	1.00	1.00
TGF-beta signaling pathway	1.00	1.00	1.00
VEGF signaling pathway	1.00	0.98	1.00
Toll-like receptor signaling pathway	1.00	0.98	1.00
NOD-like receptor signaling pathway	1.00	0.61	1.00
RIG-I-like receptor signaling pathway	1.00	1.00	1.00
T cell receptor signaling pathway	1.00	0.46	1.00
B cell receptor signaling pathway	1.00	0.61	1.00
Fc epsilon RI signaling pathway	1.00	0.98	1.00
Neurotrophin signaling pathway	1.00	0.46	1.00
Insulin signaling pathway	1.00	0.46	1.00
GnRH signaling pathway	1.00	1.00	1.00
Adipocytokine signaling pathway	1.00	1.00	1.00
Epithelial cell signaling in Helicobacter pylori infection	1.00	1.00	1.00

Table A18: p-values after false discovery rate correction for all pathways in the TCGA data

Pathway	NetGSA	GSA-s	GSA-c
Epithelial cell signaling in Helicobacter pylori infection	$5e^{-95}$	0.00	1.00
Cell cycle	$2e^{-47}$	0.00	1.00
Galactose metabolism	$3e^{-31}$	0.00	1.00
Glutathione metabolism	$1e^{-27}$	0.00	1.00
NOD-like receptor signaling pathway	$1e^{-24}$	0.00	1.00
Pyrimidine metabolism	$4e^{-23}$	0.00	1.00
Cysteine and methionine metabolism	$1e^{-22}$	0.00	1.00
Starch and sucrose metabolism	$1e^{-18}$	0.00	1.00
Toll-like receptor signaling pathway	$1e^{-18}$	0.00	1.00
Glycolysis / Gluconeogenesis	$3e^{-17}$	0.00	1.00
Jak-STAT signaling pathway	$9e^{-15}$	0.00	1.00
Chemokine signaling pathway	$3e^{-14}$	0.00	1.00
ErbB signaling pathway	$7e^{-13}$	0.00	1.00
p53 signaling pathway	$7e^{-12}$	0.00	1.00
Hedgehog signaling pathway	$5e^{-10}$	0.00	1.00
beta-Alanine metabolism	$1e^{-7}$	0.00	1.00
Fc epsilon RI signaling pathway	$5e^{-7}$	0.00	1.00
Fructose and mannose metabolism	$2e^{-6}$	0.00	1.00
Pentose phosphate pathway	$2e^{-6}$	0.00	1.00
PPAR signaling pathway	$5e^{-6}$	0.00	1.00
Adipocytokine signaling pathway	$4e^{-5}$	0.00	1.00
Purine metabolism	$6e^{-5}$	0.00	1.00
Valine, leucine and isoleucine degradation	$5e^{-4}$	$1e^{-3}$	1.00
GnRH signaling pathway	$2e^{-3}$	0.00	1.00
TGF-beta signaling pathway	$3e^{-3}$	0.00	1.00
Neurotrophin signaling pathway	0.02	0.00	1.00
Fatty acid metabolism	0.03	0.01	1.00
Oxidative phosphorylation	0.04	0.00	1.00
Lysine degradation	0.04	0.00	1.00
Arginine and proline metabolism	0.06	0.00	1.00
VEGF signaling pathway	0.07	0.00	1.00
mTOR signaling pathway	0.08	0.00	1.00
Glycine serine and threonine metabolism	0.10	0.00	1.00
Phosphatidylinositol signaling system	0.17	0.00	1.00
Notch signaling pathway	0.65	0.00	1.00
MAPK signaling pathway	0.82	0.00	1.00
Citrate cycle (TCA cycle)	1.00	0.00	1.00
Tryptophan metabolism	1.00	0.00	1.00
Amino sugar and nucleotide sugar metabolism	1.00	0.00	1.00
Calcium signaling pathway	1.00	0.00	1.00
Wnt signaling pathway	1.00	0.00	1.00
RIG-I-like receptor signaling pathway	1.00	0.00	1.00
T cell receptor signaling pathway	1.00	0.00	1.00
B cell receptor signaling pathway	1.00	0.00	1.00
Insulin signaling pathway	1.00	0.00	1.00