Inference of Differential Gene Regulatory Networks Based on Gene Expression and Genetic Perturbation Data

Xin Zhou[1], Xiaodong Cai[1,2,*]

[1]Department of Electrical and Computer Engineering, University of Miami, Coral Gables, Florida 33146, United States;[2]Sylvester Comprehensive Cancer Center, University of Miami, Miami, Florida 33136, United States

*x.cai@miami.edu

Abstract

Motivation: Gene regulatory networks (GRNs) of the same organism can be different under different conditions, although the overall network structure may be similar. Understanding the difference in GRNs under different conditions is important to understand condition-specific gene regulation. When gene expression and other relevant data under two different conditions are available, they can be used by an existing network inference algorithm to estimate two GRNs separately, and then to identify the difference between the two GRNs. However, such an approach does not exploit the similarity in two GRNs, and may sacrifice inference accuracy.

Results: In this paper, we model GRNs with the structural equation model (SEM) that can integrate gene expression and genetic perturbation data, and develop an algorithm named fused sparse SEM (FSSEM), to jointly infer GRNs under two conditions, and then to identify difference of the two GRNs. Computer simulations demonstrate that the FSSEM algorithm outperforms the approach that estimates two GRNs separately. Analysis of a gene expression and SNP dataset of lung cancer and normal lung tissues with FSSEM inferred a GRN largely agree with the known lung GRN reported in the literature, and it identified a differential GRN, whose genes with largest degrees were reported to be implicated in lung cancer. The FSSEM algorithm provides a valuable tool for joint inference of two GRNs and identification of the differential GRN under two conditions.

Availability: The software package for the FSSEM algorithm is available at https://github.com/Ivis4ml/FSSEM.git

Contact: x.cai@miami.edu

Keywords

Gene network; Differential network; Structural equation model

Introduction

A gene regulatory networks (GRN) consists of a set of genes that interact with each other to govern their expression and molecular functions. For example, transcription factors (TFs) can bind to promoter regions of their target genes and regulate the expression of target genes (Harbison *et al.*, 2004). Gene-gene interactions can change under different environments, in different tissue types or disease states, and during development and speciation (Ideker and Krogan, 2012). Therefore, GRNs undergo substantial rewiring depending on specific molecular context in which they operate

(Califano, 2011). Identification of condition-specific GRNs is critical to unravel the molecular mechanism of various tissue or disease-specific biological processes (Sonawane *et al.*, 2017).

Although a number of computational methods have been developed to infer GRNs from gene expression and other relevant data, they are mainly concerned with the static structure of gene networks under a certain condition. Several methods aim to infer GRNs using only gene expression data; they include the approaches that construct relevance network based on a similarity measure, such as correlation or mutual information (Butte and Kohane, 1999; Faith *et al.*, 2007; Margolin *et al.*, 2006), Gaussian Graphical Model (GGM) (Friedman *et al.*, 2008), Bayesian networks (Statnikov and Aliferis, 2010), and linear regression model (Haury *et al.*, 2012). Several other methods infer GRNs by integrating genetic perturbations with gene expression data; these methods include approaches using Bayesian networks incorporating expression quantitative trait loci (eQTLs) (Zhu *et al.*, 2007), likelihood-based causal models (Neto *et al.*, 2008), and structural equation models (SEMs) (Cai *et al.*, 2013; Liu *et al.*, 2008; Logsdon and Mezey, 2010).

While it is possible to apply these methods to identify GRNs under different conditions separately, such an approach is apparently not optimal to identify the difference in GRNs, because it does not exploit the similarity in two GRNs. Several methods have been proposed to use the gene expression data of different conditions to jointly estimate GRNs under different conditions. Particularly, GRNs under multiple conditions are modeled with multiple GGMs, and these GGMs are inferred jointly from gene expression data (Danaher *et al.*, 2014). When a gene is mutated, its regulatory effect on all its target gene may changed. Taking into account such effects, a node-based approach to joint inference of multiple GGMs were developed in (Mohan *et al.*, 2014). GGMs exploit the sample covariance of the gene expression levels, but they cannot integrate genetic perturbations with gene expression data. Moreover, it has been demonstrated that genetic perturbation along with gene expression data can determine directed edges in GRNs (Logsdon and Mezey, 2010), but GGMs can only identify undirected edges.

In this paper, we employ SEMs to model GRNs as described in (Cai *et al.*, 2013; Liu *et al.*, 2008; Logsdon and Mezey, 2010). This enables us to integrate genetic perturbation data with gene expression data. Taking into account the sparsity in GRNs, we have developed a sparse-aware maximum likelihood (SML) method (Cai *et al.*, 2013) to infer a single GRN based on SEM. Here, taking into account not only the sparsity in GRNs but also the sparsity in the differences between GRNs under two different conditions, we develop an algorithm, named fused sparse SEM (FSSEM), to infer two GRNs from different conditions jointly, and then to identify difference in two GRNs. Computer simulations demonstrate the superior performance of our novel approach relative to the existing one that infers GRNs under two conditions separately.

Methods

1.1 GRN model

Suppose that expression levels of n genes under two different conditions are measured using e.g. micro-array or RNA-Seq technique. Let $\mathbf{y}_i^{(k)} = [y_{i1}^{(k)}, y_{i2}^{(k)}, ..., y_{in}^{(k)}]^T$ denote expression level of n genes in individual i under condition k, where $k = 1, 2$ and $i = 1, 2, ..., n_k$, with n_k being the number of individuals where gene expression levels are measured under condition k. Supposed that a set of perturbations of these genes have also been measured. These perturbations can be due to e.g., eQTLs or gene copy number variants (CNVs). In this paper, we will consider only eQTLs. As in (Cai *et al.*, 2013; Logsdon and Mezey, 2010), we assume that each gene in the GRN of interest

has at least one *cis*-eQTL, so that the structure of underlying GRN is uniquely identifiable. Let $\mathbf{x}_i^{(k)} = [x_{i1}^{(k)}, x_{i2}^{(k)}, ..., x_{iq}^{(k)}]^T$ denote the genotypes of q *cis*-eQTLs in individual i under condition k, where $k = 1,2$ and $i = 1,2,...,n_k$. Since the expression level of a particular gene may be regulated by other genes and is affected by its eQTLs, we employ the following SEM to model the expression of n genes

$$\mathbf{y}_i^{(k)} = \mathbf{B}^{(k)}\mathbf{y}_i^{(k)} + \mathbf{F}^{(k)}\mathbf{x}_i^{(k)} + \mu_i^{(k)} + \varepsilon_i^{(k)}, \tag{1}$$

where $i = 1,...,n_k$, $k = 1,2$, $n \times n$ matrix $\mathbf{B}^{(k)}$ defines the unknown network structure under condition k, $n \times q$ matrix $\mathbf{F}^{(k)}$ captures the effect of *cis*-eQTLs on gene expression levels under condition k, $n \times 1$ vector $\mu^{(k)}$ accounts for the model bias in SEM, and $n \times 1$ vector $\varepsilon^{(k)}$ denotes the residual error, which is modeled as a Gaussian vector with zero mean and variance σ^2. It is assumed that no self-loops are presented per gene in GRN, which implies that the diagonal entries of $\mathbf{B}^{(k)}$ are zero, and it is also assumed that q *cis*-eQTLs have been identified using an existing eQTL method, but their regulatory effects are unknown, thus, $\mathbf{F}^{(k)}$ has q nonzero entries with known locations.

1.2 Joint Inference of two GRNs

Let $\mathbf{Y}^{(k)} = [\mathbf{y}_1^{(k)}, ..., \mathbf{y}_{n_k}^{(k)}]$, $\mathbf{X}^{(k)} = [\mathbf{x}_1^{(k)}, ..., \mathbf{x}_{n_k}^{(k)}]$ and $\mathbf{E}^{(k)} = [\varepsilon_1^{(k)}, ..., \varepsilon_{n_k}^{(k)}]$, where $k = 1,2$, and assume that $n_1 + n_2$ observations are independent. Then, the negative log-likelihood function of the data can be written as

$$\begin{aligned}\mathbb{L}(\mathbf{B},\mathbf{F},\mu,\sigma^2) &= -\log \prod_{k=1}^{2}\prod_{i=1}^{n_k} \mathbb{P}(\mathbf{y}_i^{(k)} \,|\, \mathbf{x}_i^{(k)}, \mu_i^{(k)}, \mathbf{B}^{(k)}, \mathbf{F}^{(k)}) \\ &= -\sum_{k=1}^{2} \frac{n_k}{2} \log|\mathbf{I} - \mathbf{B}^{(k)}|^2 + \frac{(n_1+n_2)n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{k=1}^{2}\left|\left|(\mathbf{I}-\mathbf{B}^{(k)})\mathbf{Y}^{(k)} - \mathbf{F}^{(k)}\mathbf{X}^{(k)} - \mu^{(k)}\right|\right|_F^2,\end{aligned} \tag{2}$$

where $\mathbf{B} = [\mathbf{B}^{(1)}, \mathbf{B}^{(2)}]$, $\mathbf{F} = [\mathbf{F}^{(1)}, \mathbf{F}^{(2)}]$, $\mu = [\mu^{(1)}, \mu^{(2)}]$, and $\|\,.\,\|_F$ stands for the Frobenius norm. Our goal is to estimate $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ in (2). It is not difficult to show that minimizing (2) with respect to μ yields $\hat{\mu}^{(k)} = (\mathbf{I} - \mathbf{B}^{(k)})\tilde{\mathbf{Y}}^{(k)} - \mathbf{F}^{(k)}\tilde{\mathbf{X}}^{(k)}$, where $\tilde{\mathbf{Y}}^{(k)} = \mathbf{Y}^{(k)} - 1/n_k \sum_{i=1}^{n_k} \mathbf{y}_i^{(k)}\mathbf{1}$, $\tilde{\mathbf{X}}^{(k)} = \mathbf{X}^{(k)} - 1/n_k \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)}\mathbf{1}$, and $\mathbf{1}$ is a vector with all its entries equal to 1.

Since a gene is regulated by a small number of other genes (Gardner *et al.*, 2003; Tegner *et al.*, 2003; Thieffry *et al.*, 1998), GRNs are sparse, meaning that most entries of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ are zeroes. Moreover, it is reasonable to expect that changes in a GRN under two different conditions is relatively small. Therefore, most entries of $\mathbf{B}^{(2)} - \mathbf{B}^{(1)}$ are zeroes. Let $\hat{\sigma}^2$ be an estimate of σ^2 that will be specified later, replacing $\mu^{(k)}$ and σ^2 in (2) with $\hat{\mu}^{(k)}$ and $\hat{\sigma}^2$ respectively, and taking into account the sparsity in $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$, and the sparsity in $\mathbf{B}^{(2)} - \mathbf{B}^{(1)}$, we can estimate $\mathbf{B}$ and $\mathbf{F}$ by minimizing the following penalized negative log-likelihood function

$$\begin{aligned}J(\mathbf{B},\mathbf{F}) = &-\sum_{k=1}^{2} n_k \log|\mathbf{I} - \mathbf{B}^{(k)}| \\ &+ \frac{1}{2\hat{\sigma}^2}\sum_{k=1}^{2} \| (\mathbf{I}-\mathbf{B}^{(k)})\tilde{\mathbf{Y}}^{(k)} - \mathbf{F}^{(k)}\tilde{\mathbf{X}}^{(k)} \|_F^2 \\ &+ \lambda \sum_{k=1}^{2} \| \mathbf{B}^{(k)} \|_{1,w^{(k)}} + \rho \| \mathbf{B}^{(2)} - \mathbf{B}^{(1)} \|_{1,r},\end{aligned} \tag{3}$$

where $\mathbf{B}_{ii}^{(k)} = 0,\ \forall i = 1,...,n,\ k = 1,2,\ \| \mathbf{B}^{(k)} \|_{1,w^{(k)}} = \sum_i \sum_j w_{ij}^{(k)} |\mathbf{B}_{ij}^{(k)}|$ is the weighted ℓ_1-norm, $\| \mathbf{B}^{(2)} - \mathbf{B}^{(1)} \|_{1,r}$ is also a weighted ℓ_1-norm with similar definition, λ and ρ are two nonnegative parameters. Weights $w_{ij}^{(k)}$ and r_{ij} in the penalty terms are introduced to improve estimation accuracy and robustness in line with the adaptive lasso (Zou, 2006) and the adaptive generalized fused lasso (Viallon *et al.*, 2016), and they are selected as $1/|\hat{\mathbf{B}}_{ij}^{(k)}|$ and $1/|\hat{\mathbf{B}}_{ij}^{(2)} - \hat{\mathbf{B}}_{ij}^{(1)}|$, respectively, where $\hat{\mathbf{B}}^{(1)}$ and $\hat{\mathbf{B}}^{(2)}$ are preliminary estimates of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ obtained from the following ridge regression:

$$\{\hat{\mathbf{B}}, \hat{\mathbf{F}}\} = \arg\min_{\{\mathbf{B},\mathbf{F}\}} \left\{ \sum_{k=1}^{2} \frac{1}{2} \left|\left| (\mathbf{I} - \mathbf{B}^{(k)}) \tilde{\mathbf{Y}}^{(k)} - \mathbf{F}^{(k)} \tilde{\mathbf{X}}^{(k)} \right|\right|_F^2 + \tau \left|\left| \mathbf{B}^{(k)} \right|\right|_F^2 \right\} \quad \text{s.t. } \mathbf{B}_{ii}^{(k)} = 0, \forall i = 1,...,n, k = 1,2, \tag{4}$$

where $\hat{\mathbf{B}} = [\hat{\mathbf{B}}^{(1)}, \hat{\mathbf{B}}^{(2)}]$, $\hat{\mathbf{F}} = [\hat{\mathbf{F}}^{(1)}, \hat{\mathbf{F}}^{(2)}]$, and the estimate of σ^2, $\hat{\sigma}^2$ in (3), is given by

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{2} \| (\mathbf{I} - \hat{\mathbf{B}}^{(k)}) \tilde{\mathbf{Y}}^{(k)} - \hat{\mathbf{F}}^{(k)} \tilde{\mathbf{X}}^{(k)} \|_F^2}{(n_1 + n_2) n}. \tag{5}$$

Based on (3), we next develop a proximal alternative linearize minimization algorithm to infer $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$.

1.3 Ridge regression

In the first stage, we solve the ridge regression problem (4) to find initial values of $\mathbf{B}$, $\mathbf{F}$, weights $w^{(k)}$, $k = 1,2$, and r for the FSSEM algorithm to minimize (3). Let $\mathbf{B}_i^{(k)}, \mathbf{F}_i^{(k)}$ and $\mathbf{Y}_i^{(k)}$ be the i-th row of $\mathbf{B}^{(k)}, \mathbf{F}^{(k)}$ and $\mathbf{Y}^{(k)}$, respectively. Define $\mathbf{B}_{i,-i}^{(k)}$ as the $1 \times (n-1)$ vector obtained by removing the i-th entry from $\mathbf{B}_i^{(k)}$. Let $S_q(i)$ be the set of indices of non-zero entries in the $\mathbf{F}_i^{(k)}$, $\mathbf{F}_{i,S_q(i)}^{(k)}$ be the vector that contains the nonzero entries of $\mathbf{F}_i^{(k)}$, $\tilde{\mathbf{X}}_{S_q(i)}$ be the matrix formed by taking rows of $\tilde{\mathbf{X}}$ whose indices are in $S_q(i)$, and $\tilde{\mathbf{Y}}_{-i}$ be the matrix formed by removing ith row of $\tilde{\mathbf{Y}}$.

Then, the ridge regression problem (4) can be decomposed into n separate problems:

$$\arg\min_{\mathbf{B}_{i,-i}, \mathbf{F}_{i,S_q(i)}} \left\{ \sum_{k=1}^{2} \frac{1}{2} \left|\left| \tilde{\mathbf{Y}}_i^{(k)} - \mathbf{B}_{i,-i}^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)} - \mathbf{F}_{i,S_q(i)}^{(k)} \tilde{\mathbf{X}}_{S_q(i)}^{(k)} \right|\right|_F^2 + \tau \left|\left| \mathbf{B}_{i,-i}^{(k)} \right|\right|_F^2 \right\}, i = 1,...,n. \tag{6}$$

Minimizing the objective function in (6) with respect to (w.r.t.) $\mathbf{F}_{i,S_q(i)}^{(k)}$ yields the following closed-form solution

$$\hat{\mathbf{F}}_{i,S_q(i)}^{(k)} = (\tilde{\mathbf{Y}}_i^{(k)} - \hat{\mathbf{B}}_{i,-i}^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)}) \tilde{\mathbf{X}}_{S_q(i)}^{(k)T} (\tilde{\mathbf{X}}_{S_q(i)}^{(k)} \tilde{\mathbf{X}}_{S_q(i)}^{(k)T})^{-1}. \tag{7}$$

Substituting $\hat{\mathbf{F}}_{i,S_q(i)}^{(k)}$ into (6) and minimizing w.r.t. $\hat{\mathbf{B}}_{i,-i}^{(k)}$ gives $\hat{\mathbf{B}}_{i,-i}^{(k)} = \tilde{\mathbf{Y}}_i^{(k)} \mathbf{P}_i^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)T} (\tilde{\mathbf{Y}}_{-i}^{(k)} \mathbf{P}_i^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)T} + \tau \mathbf{I})^{-1}$, which in turn results in $\hat{\mathbf{F}}_{i,S_q(i)}^{(k)} = \tilde{\mathbf{Y}}_i^{(k)} \Gamma_i^{(k)} \tilde{\mathbf{X}}_{S_q(i)}^{(k)T} (\tilde{\mathbf{X}}_{S_q(i)}^{(k)} \tilde{\mathbf{X}}_{S_q(i)}^{(k)T})^{-1}$, where $\Gamma_i^{(k)} = \mathbf{I} - \mathbf{P}_i^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)T} (\tilde{\mathbf{Y}}_{-i}^{(k)} \mathbf{P}_i^{(k)} \tilde{\mathbf{Y}}_{-i}^{(k)T} + \tau \mathbf{I})^{-1} \tilde{\mathbf{Y}}_{-i}^{(k)}$ and $\mathbf{P}_i^{(k)} = \mathbf{I} - \tilde{\mathbf{X}}_{S_q(i)}^{(k)T} (\tilde{\mathbf{X}}_{S_q(i)}^{(k)} \tilde{\mathbf{X}}_{S_q(i)}^{(k)T})^{-1} \tilde{\mathbf{X}}_{S_q(i)}^{(k)}$. After $\hat{\mathbf{B}}^{(k)}$ and $\hat{\mathbf{F}}^{(k)}$ are estimated, the estimate of $\hat{\sigma}^2$ is given by (5). The hyper-parameter τ in ridge regression (4) or (6) is selected by 5-fold cross-validation.

1.4 FSSEM algorithm

In this section, we will develop the FSSEM algorithm to minimize the objective function $J(\mathbf{B},\mathbf{F})$ in (3) with the initial values of $\mathbf{B}^{(k)}$ and $\mathbf{F}^{(k)}$ given in (7). The objective function is non-convex due to the log-determinant term, and non-smooth due to the ℓ_1 norm terms. Recently, the proximal alternating linearized minimization (PALM) method (Bolte *et al.*, 2014) was developed to solve a broad classes of non-convex and non-smooth minimization problems. We next apply the PALM approach to develop the FSSEM algorithm.

Without loss of generality, we define the proximal operator associated with a proper and lower semi-continuous function $h(\mathbf{x}):\mathbb{R}^d\to(-\infty,+\infty]$ as $\text{prox}_\alpha^h(v)=\arg\min_{u\in\mathbb{R}^d}\left\{h(u)+\alpha/2\|u-v\|^2\right\}$, where $\alpha>0$ and $v\in\mathbb{R}^d$ are given. We also define the fused lasso signal approximator (Friedman *et al.*, 2007; Hoefling, 2010) on $x=[x_1,x_2]$ as the following proximal operator:

$$\begin{aligned}\text{prox}_\alpha^{p(x)}(z_k)=\arg\min_{x\in\mathbb{R}^d}\Big\{&\frac{\alpha}{2}\sum_{k=1}^{2}\parallel x_k-z_k\parallel^2\\&+\lambda\sum_{k=1}^{2}\parallel x_k\parallel_1+\rho\parallel x_2-x_1\parallel_1\Big\}.\end{aligned}\tag{8}$$

The solution $(x_1(\lambda),x_2(\lambda))$ of (8) at $\lambda=0$ can be found as

$$(x_1(0),x_2(0))=\begin{cases}(z_1-\rho/\alpha,z_2+\rho/\alpha) & \text{if}\quad z_1-z_2>2\rho/\alpha\\(z_1+\rho/\alpha,z_2-\rho/\alpha) & \text{if}\quad z_1-z_2<-2\rho/\alpha\\(\frac{z_1+z_2}{2},\frac{z_1+z_2}{2}) & \text{if}\quad |z_1-z_2|\le 2\rho/\alpha\end{cases}.\tag{9}$$

Defining soft-thresholding function $S(\beta,\lambda)$ as

$$S(\beta,\lambda)=\begin{cases}\beta-\lambda & \text{if}\quad \beta>\lambda\\\beta+\lambda & \text{if}\quad \beta<-\lambda\\0 & \text{if}\quad |\beta|\le\lambda\end{cases},\tag{10}$$

the solution of (8) at $\lambda>0$ is given in terms of the soft-thresholding operator as follows (Friedman *et al.*, 2007):

$$\text{prox}_\alpha^{f(x)}(z_k)=\big(S(x_1(0),\lambda/\alpha),S(x_2(0),\lambda/\alpha)\big).\tag{11}$$

Minimizing (3) w.r.t. $\mathbf{F}^{(k)}$ yields $\hat{\mathbf{F}}^{(k)}_{i,S_q(i)}$ in (7). Substituting $\hat{\mathbf{F}}^{(k)}_{i,S_q(i)}$ in (7) into (3) gives

$$J(\mathbf{B})=H(\mathbf{B})+\sum_{i=1}^{N_g}f_i(\mathbf{B}_{i,-i}),\tag{12}$$

where

$$\begin{aligned}H(\mathbf{B})=&-\sum_{k=1}^{2}\frac{n_k}{2}\log|\mathbf{I}-\mathbf{B}^{(k)}|^2\\&+\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N_g}\sum_{k=1}^{2}\parallel\tilde{\mathbf{Y}}_i^{(k)}\mathbf{P}_i^{(k)}-\mathbf{B}_{i,-i}^{(k)}\tilde{\mathbf{Y}}_{-i}^{(k)}\mathbf{P}_i^{(k)}\parallel_2^2,\end{aligned}\tag{13}$$

and

$$f_i(\mathbf{B}_{i,-i}) = \lambda(\| \mathbf{B}^{(1)}_{i,-i} \|_{1,w^{(1)}} + \| \mathbf{B}^{(2)}_{i,-i} \|_{1,w^{(2)}}) + \rho \| \mathbf{B}^{(1)}_{i,-i} - \mathbf{B}^{(2)}_{i,-i} \|_{1,r} . \tag{14}$$

Using the inertial version of the PALM approach (Pock and Sabach, 2016), the FSSEM algorithm efficiently minimizes the non-convex non-smooth function $J(\mathbf{B})$ with the block coordinate descent (BCD) method in an iterative fashion. More specifically, in each cycle of the iteration, $J(\mathbf{B})$ is minimized successively w.r.t. $[\mathbf{B}^{(1)}_{i,-i}, \mathbf{B}^{(2)}_{i,-i}]$, while $[\mathbf{B}^{(1)}_{j,-j}, \mathbf{B}^{(2)}_{j,-j}], j = 1,...,n, j \neq i$ are fixed.

Let us consider updating the ith block of variables $\mathbf{B}_{i,-i} = [\mathbf{B}^{(1)}_{i,-i}, \mathbf{B}^{(2)}_{i,-i}]$ in the $(t+1)$th cycle. Let $\mathbf{B}[t] = [\mathbf{B}^{(1)}[t], \mathbf{B}^{(2)}[t]]$ be the estimate of $\mathbf{B}$ in the tth cycle. Define $\tilde{\mathbf{B}}_{i,-i} = \mathbf{B}_{i,-i}[t_1] + \alpha_t(\mathbf{B}_{i,-i}[t_1] - \mathbf{B}_{i,-i}[t_1 - 1])$, where $t_1 = t+1, \forall i < j$, $t_1 = t, \forall i > j$, and α_t is a constant in the interval $[0,1]$. We obtain $\mathbf{B}_{i,-i}$ from the FLSA proximal operator (11) as follows:

$$\mathbf{B}_{i,-i} = \text{prox}^{f_i(.)}_{\gamma_i}\left(\tilde{\mathbf{B}}_{i,-i} - \frac{1}{\gamma_i}\nabla_{\mathbf{B}_{i,-i}} H(\tilde{\mathbf{B}})\right), \tag{15}$$

where $1/\gamma_i$ is the step-size for the i-th block that will be given later, and $\nabla_{\mathbf{B}_{i,-i}} H(\tilde{\mathbf{B}})$ is the partial derivative of $H(\mathbf{B})$ w.r.t. $\mathbf{B}_{i,-i}$ at $\tilde{\mathbf{B}}$.

Since $\mathbf{B}_{i,-i} = [\mathbf{B}^{(1)}_{i,-i}, \mathbf{B}^{(2)}_{i,-i}]$, we have $\nabla_{\mathbf{B}_{i,-i}} H(\mathbf{B}) = [\nabla_{\mathbf{B}^{(1)}_{i,-i}} H(\mathbf{B}), \nabla_{\mathbf{B}^{(2)}_{i,-i}} H(\mathbf{B})]$. The determinant of $\mathbf{I} - \mathbf{B}^{(k)}$ can be expressed as $\mathbf{c}^{(k)}_{ii} - \mathbf{B}^{(k)}_{i,-i}\mathbf{c}^{(k)}_i$, where $\mathbf{c}^{(k)}_{ii}$ is the (i,i) co-factor of $\mathbf{I} - \mathbf{B}^{(k)}$, and the jth entry of the $(n-1)\times 1$ column vector $\mathbf{c}^{(k)}_i$ is the co-factor of $\mathbf{I} - \mathbf{B}^{(k)}$ corresponding to the jth entry of $\mathbf{B}^{(k)}_{i,-i}$. Defining $\mathbf{B}^{(k)}_{-i} = \{\mathbf{B}^{(k)}_{j,-j}, j = 1,...,n, j \neq i\}$, we can write $\nabla_{\mathbf{B}^{(k)}_{i,-i}} H(\mathbf{B})$, $k = 1,2$, with $\mathbf{B}^{(k)}_{-i}$ fixed, as follows

$$\nabla_{\mathbf{B}^{(k)}_{i,-i}} H(\mathbf{B}) = \frac{n_k \mathbf{c}^{(k)T}_i}{\mathbf{c}^{(k)}_{ii} - \mathbf{B}^{(k)}_{i,-i}\mathbf{c}^{(k)}_i} + \frac{1}{\sigma^2}\left(\mathbf{B}^{(k)}_{i,-i}\tilde{\mathbf{Y}}^{(k)}_{-i}\mathbf{P}^{(k)}_i\tilde{\mathbf{Y}}^{(k)T}_{-i} - \tilde{\mathbf{Y}}^{(k)}_i\mathbf{P}^{(k)}_i\tilde{\mathbf{Y}}^{(k)T}_{-i}\right). \tag{16}$$

In Supplementary Text S, we prove that given $\mathbf{B}^{(k)}_{-i}$, $\nabla_{\mathbf{B}^{(k)}_{i,-i}} H(\mathbf{B}), k = 1,2$ are Lipschitz continuous. Specifically, we can write $\nabla_{\mathbf{B}^{(k)}_{i,-i}} H(\mathbf{B})$ as $\nabla_{\mathbf{B}^{(k)}_{i,-i}} H(\mathbf{B}^{(k)}_{i,-i}, \mathbf{B}^{(k)}_{-i})$, which satisfies:

$$\| \nabla_{\mathbf{B}^{(k)}_{i,-i}} H(x, \mathbf{B}^{(k)}_{-i}) - \nabla_{\mathbf{B}^{(k)}_{i,-i}} H(y, \mathbf{B}^{(k)}_{-i}) \leq L_i(\mathbf{B}^{(k)}_{-i}) \| x - y \|, \tag{17}$$

where the Lipschitz constant $L_i(\mathbf{B}^{(k)}_{-i})$ is derived in the Supplementary Text S, and is given by

$$L_i(\mathbf{B}^{(k)}_{-i}) = n_k \| \mathbf{c}^{(k)}_i \|^2_2 \Big/ \min_{\mathbf{B}^{(k)}_{i,-i}} (\det(\mathbf{I} - \mathbf{B}^{(k)}))^2 + \lambda_{\max}(\tilde{\mathbf{Y}}^{(k)}_{-i}\mathbf{P}^{(k)}_i\tilde{\mathbf{Y}}^{(k)T}_{-i})/\sigma^2. \tag{18}$$

Here $\lambda_{\max}(\tilde{\mathbf{Y}}^{(k)}_{-i}\mathbf{P}^{(k)}_i\tilde{\mathbf{Y}}^{(k)T}_{-i})$ is the maximum eigenvalue of $\tilde{\mathbf{Y}}^{(k)}_{-i}\mathbf{P}^{(k)}_i\ \tilde{\mathbf{Y}}^{(k)T}_{-i}$, and the value of $\min_{\mathbf{B}^{(k)}_{i,-i}} \det(\mathbf{I} - \mathbf{B}^{(k)})^2$ can be computed by solving the optimization problem as shown in (S12) in Supplementary

Algorithm 1 Fused Sparse SEM (FSSEM)

Select τ^* in (4) via cross-validation
Solve (4) with τ^* to obtain $(\hat{\mathbf{B}}, \hat{\mathbf{F}})$, and compute $\hat{\sigma}^2$ from (5).
Set $w_{ij}^{(k)} = 1/|\hat{\mathbf{B}}_{ij}^{(k)}|$, $r_{ij} = 1/|\hat{\mathbf{B}}_{ij}^{(2)} - \hat{\mathbf{B}}_{ij}^{(1)}|$.
Initialize $\mathbf{B}[0] = \tilde{\mathbf{B}} = \hat{\mathbf{B}}$.
for t in $1, 2, \ldots$ **do**
  Select $\alpha_t \in [0, 1]$
  for i in $1, \ldots, n$ **do**
    Compute $L_i(\tilde{\mathbf{B}}_{-i})$ from (18), set $\gamma_i = L_i(\tilde{\mathbf{B}}_{-i})$
    Update $\mathbf{B}_{i,-i}^{(k)}$, $k = 1, 2$, with (15)
    Set $\tilde{\mathbf{B}}_{i,-i} = \mathbf{B}_{i,-i}[t] + \alpha_t(\mathbf{B}_{i,-i}[t] - \mathbf{B}_{i,-i}[t-1])$
  end for
  Update $\mathbf{F}_i^{(k)}$ with (7) and $\hat{\sigma}^2$ with (5)
  if convergence **then**
    Break
  end if
end for
Return $\{\hat{\mathbf{B}}^{(k)}, \hat{\mathbf{F}}^{(k)}, k = 1, 2\}$

Text S. Let $L_i(\mathbf{B}_{-i}) = \max\{L_i(\mathbf{B}_{-i}^{(k)}), k = 1, 2\}$. Then, the step size is chosen to be $1/\gamma_i = 1/L_i(\tilde{\mathbf{B}}_{-i})$.

The FSSEM algorithm is summarized in Algorithm 1. The convergence criterion is defined as

$$\begin{aligned} &\Big\{\sum_{k=1}^{2} \Big|\Big|\mathbf{B}^{(k)}[t+1] - \mathbf{B}^{(k)}[t]\Big|\Big|_F^2 \Big/ \sum_{k=1}^{2} \Big|\Big|\mathbf{B}^{(k)}[t]\Big|\Big|_F^2 \\ &+ \sum_{k=1}^{2} \Big|\Big|\mathbf{F}^{(k)}[t+1] - \mathbf{F}^{(k)}[t]\Big|\Big|_F^2 \Big/ \sum_{k=1}^{2} \Big|\Big|\mathbf{F}^{(k)}[t]\Big|\Big|_F^2 \Big\} < \varepsilon_v \\ &|J(\mathbf{B}[t+1]) - J(\mathbf{B}[t])| \big/ |J(\mathbf{B}[t])| < \varepsilon_o, \end{aligned} \tag{19}$$

where $\varepsilon_v > 0$ and $\varepsilon_o > 0$ are pre-specified small constants. Since the objective function is not convex, it is not guaranteed that the FSSEM algorithm converges to the global minimization. However, we prove in Supplementary Text S that the FSSEM algorithm always converges to a stationary point of the objective function. Note that if we drop the fused lasso term $\rho \parallel \mathbf{B}^{(1)} - \mathbf{B}^{(2)} \parallel_{1,r}$ in (3), then minimizing $J(\mathbf{B}, \mathbf{F})$ is equivalent to estimating two network matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ separately. The BCD approach used in FSSEM can also be employed to solve this problem, because the proximal operator in (15) can be easily solved in terms of the soft-thresholding function $S(\beta, \lambda)$ defined in (10). This BCD approach is much more efficient than the SML algorithm in (Cai *et al.*, 2013), which employs the element-wise coordinate ascent approach. Parameters λ and ρ in (3) can be determined with cross-validation (CV). In Supplementary Text S, we derive the expression for the maximum values of λ and ρ and describe the CV process.

Results

2.1 Computer simulations

In this section, we conduct simulation studies to compare the performance of the FSSEM algorithm with that of the SML algorithm (Cai *et al.*, 2013). FSSEM estimates network matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ jointly, while SML estimates $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ separately. Other algorithm such as AL-based (Logsdon and Mezey, 2010) and QDG (Neto *et al.*, 2008) algorithms are available to estimate $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ separately. However, as shown in (Cai *et al.*, 2013), SML algorithm outperforms AL-based and QDG algorithms. Therefore, only SML is considered in performance comparison.

Following the setup of (Cai *et al.*, 2013), both directed acyclic networks (DAG) and directed cyclic networks (DCG) are simulated in our experiments. Specifically, the adjacency matrix $\mathbf{A}^{(1)}$ of a DAG or DCG of 10 or 30 gene nodes with expected number of edges per gene $d = 3$ is generated for the GRN under condition 1. Another adjacency matrix $\mathbf{A}^{(2)}$ was generated by randomly change 10% entries of $\mathbf{A}^{(1)}$, and the probabilities of changes of entries from 0 to 1 and from 1 to 0 are equal. A network matrix $\mathbf{B}^{(1)}$ was generated from $\mathbf{A}^{(1)}$ as follow. For any entry $\mathbf{A}^{(1)}_{ij} = 1$, $\mathbf{B}^{(1)}_{ij}$ is generated from a random variable uniformly distributed over interval $[0.5, 1]$ or $[-1, -0.5]$; for all $\mathbf{A}^{(1)}_{ij} = 0$, we set $\mathbf{B}^{(1)}_{ij} = 0$. The second network matrix $\mathbf{B}^{(2)}$ was generated from $\mathbf{A}^{(2)}$ and $\mathbf{B}^{(1)}$ as follow. For all $\mathbf{A}^{(2)}_{ij} = 0$, we set $\mathbf{B}^{(2)}_{ij} = 0$; for all $\mathbf{A}^{(2)}_{ij} = \mathbf{A}^{(1)}_{ij}$, we set $\mathbf{B}^{(2)}_{ij} = \mathbf{B}^{(1)}_{ij}$; and for all $\mathbf{A}^{(1)}_{ij} = 0$ but $\mathbf{A}^{(2)}_{ij} = 1$, we generate $\mathbf{B}^{(2)}_{ij}$ from a random variable uniformly distributed over interval $[0.5, 1]$ or $[-1, -0.5]$. The genotypes of eQTLs were simulated from an F2 cross. Values 1 and 3 were assigned to two homozygous genotypes, respectively, and value 2 to the heterozygous genotype. Then, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were generated from ternary random variables taking on values $\{1, 2, 3\}$ with corresponding probabilities $\{0.25, 0.5, 0.25\}$. The number of eQTLs per gene n_e was chosen to be either 1 or 3, and effect sizes of all eQTLs were set to 1 in $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$. Error terms $\mathbf{E}^{(1)}$ and $\mathbf{E}^{(2)}$ were independently sampled from Gaussian random variables with zero mean and variable σ^2; $\mu^{(1)}$ and $\mu^{(2)}$ were set to zero vectors; and the sample sizes n_1 and n_2 vary from 100 to 1,000. Finally, $\mathbf{Y}^{(k)}$ was calculated as $\mathbf{Y}^{(k)} = (\mathbf{I} - \mathbf{B}^{(k)})^{-1}(\mathbf{F}^{(k)}\mathbf{X}^{(k)} + \mathbf{E}^{(k)})$, where $k = 1, 2$.

For each configuration of the two GRNs, 30 replicates of the GRN were simulated. For each replicate, SML and FSSEM were used to infer network matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$. Hyper parameter of SML and FSSEM algorithms were determined with 5-fold CV. Power of detection (PD) and false discovery rate (FDR) for detecting network edges were calculated from $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ estimated from the data of each of 30 network replicates. The differential network was defined as $\Delta\mathbf{B} = \mathbf{B}^{(2)} - \mathbf{B}^{(1)}$, and PD and FDR for the differential network were calculated accordingly.

The results for DAGs with $n = 30$, $n_e = 3$ and $\sigma^2 = 0.25$ are depicted in Figure 1, and results of DAGs under other settings are given in Figures S1 and S2 in Supplementary Text S. First, let us look at the PD and FDR of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ in the left panel of Figure 1. FSSEM offers almost the same PD as SML, but slightly lower FDR when $\sigma^2 = 0.25$. As shown in Figures S1 and S2, when $\sigma^2 = 0.01$, FSSEM and SML have almost same PD and FDR, and increasing n_e from 1 to 3 slightly reduced FDR for both FSSEM and SML. Overall, FSSEM and SML have similar performance for the estimates of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$. Next, let us look at the PD and FDR of $\mathbf{B}^{(2)} - \mathbf{B}^{(1)}$ in the right panel of Figure 1. FSSEM exhibits almost the same PD as SML, but it offers much smaller FDR than SML. Specially, the FDR of FSSEM is < 0.2, but the FDR of SML > 0.8. Similar trends are seen in Figure S1 and S2 for all network settings considered. Again, increasing the number of eQTL n_e from 1 to 3 slightly reduces the FDR. For the GRNs of $n = 30$ genes, there are $2(n^2 - n) = 1740$ unknown entries in $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ to be estimated, and there are $2nn_e$ unknown entries in $\mathbf{F}^{(1)}$ and

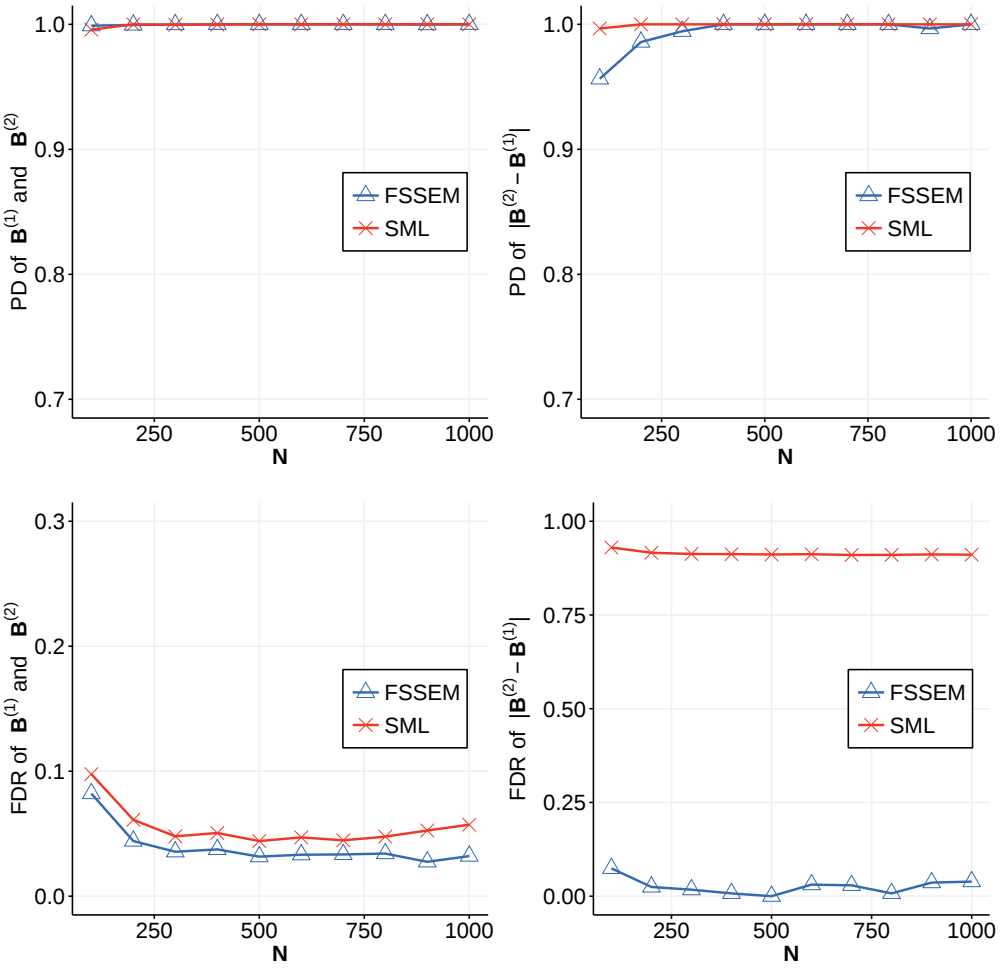

Figure 1. Performance of FSSEM and SML for the DAG with $n = 30$ genes and $n_e = 3$ eQTLs per gene. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.25$. PD and FDR were obtained from 30 network replicates.

$\mathbf{F}^{(2)}$. Interestingly, the performance of both FSSEM and SML did not change much, when the sample size $n_1 + n_2$ varied from 400 to 4,000.

Simulation results for DCGs with $n = 30$, $n_e = 3$ and $\sigma^2 = 0.25$ are depicted in Figure 2, and results of DCGs under other settings are shown in Figures S3 and S4 in Supplementary Text S. As shown in the left panel of Figure 2, FSSEM offers slightly better PD and FDR for the estimates of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ than SML. The results for $\Delta\mathbf{B}$ shown in the right panel indicate that FSSEM offers similar PD comparing with SML, but it exhibits much smaller FDR, as also observed in Figure 1 for DAG networks. Similar trends are also observed for other network settings in Figures S3 and S4. Clearly, FSSEM outperforms SML consistently in terms of both PD and FDR for both DAG and DCG networks. For the convenience of comparison, the simulation results of DAG and DCG with $n_1 = n_2 = 500$, $n_e = 3$ and $\sigma^2 = 0.25$ are summarized in Table 1, which clearly shows that FSSEM outperforms SML. Particularly, FSSEM offers much lower FDR than SML in estimating $\Delta\mathbf{B}$, the matrix of the differential GRN.

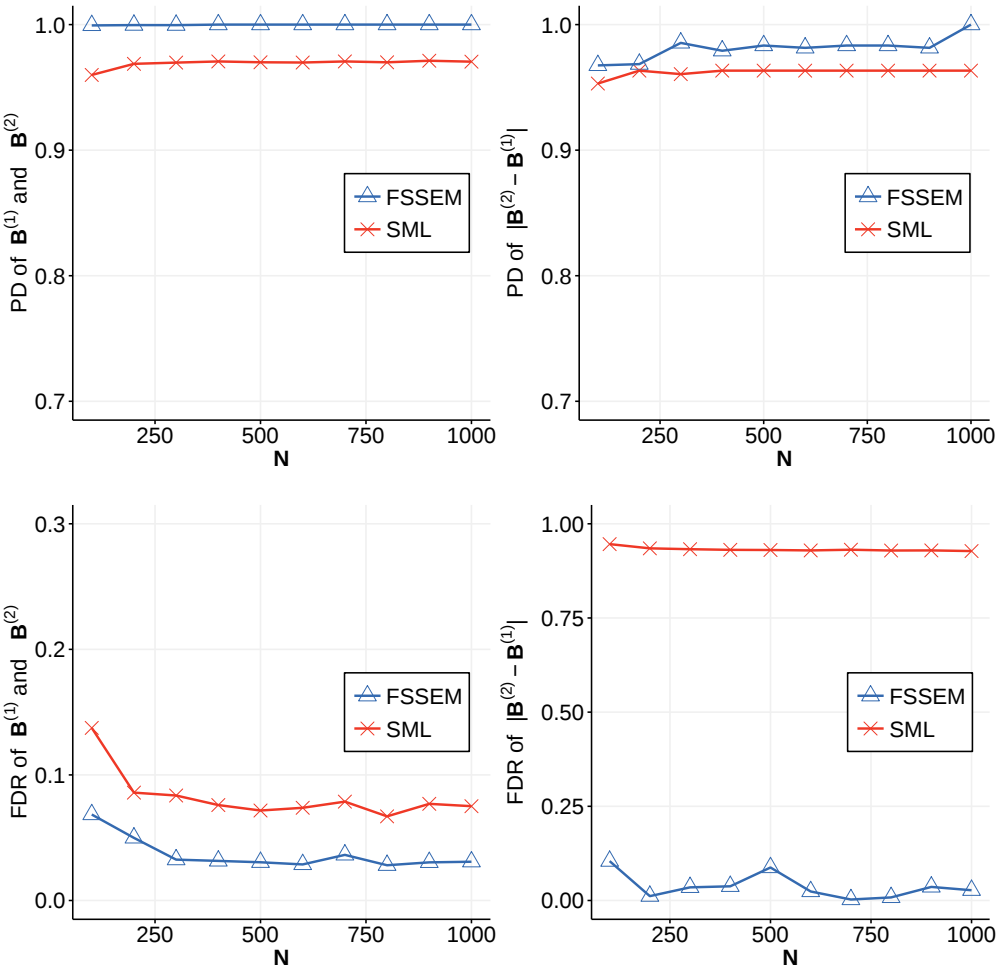

Figure 2. Performance of FSSEM and SML for the DCG with $n = 30$ genes and $n_e = 3$ eQTLs per gene. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.25$. PD and FDR were obtained from 30 network replicates.

Table 1. Performance of FSSEM and SML algorithms. Expected number of eQTLs per gene is $n_e = 3$ and noise variance $\sigma^2 = 0.25$. PD and FDR were obtained from 30 network replicates.

Network	n	FSSEM				SML			
		$PD_{\mathbf{B}}$	$FDR_{\mathbf{B}}$	$PD_{\Delta\mathbf{B}}$	$FDR_{\Delta\mathbf{B}}$	$PD_{\mathbf{B}}$	$FDR_{\mathbf{B}}$	$PD_{\Delta\mathbf{B}}$	$FDR_{\Delta\mathbf{B}}$
DAG	10	1.000	0.037	1.000	0.000	1.000	0.054	1.000	0.889
	30	1.000	0.032	1.000	0.000	1.000	0.044	1.000	0.911
DCG	10	1.000	0.028	1.000	0.016	0.879	0.082	0.856	0.912
	30	1.000	0.004	1.000	0.057	0.962	0.083	0.955	0.920

2.2 Real data analysis

In (Lu *et al.*, 2011), gene expression levels in 42 tumors and their adjacent normal tissues of non-smoking female patients with lung adenocarcinoma were measured with 54,675 probe sets

from Affymetrix Human Genome U133 Plus 2.0 Array. The genotypes of single nucleotide polymorphisms (SNPs) in the same set of tissues were obtained using 906,551 SNP probes from Affymetrix Genome-Wide Human SNP 6.0 array. We applied FSSEM to this data set to infer GRNs in lung cancer and normal tissues.

Both gene expression and SNP data in the gene expression omnibus database (GSE33356) were downloaded. The R package affy (Gautier *et al.*, 2004) was employed to transform raw micro-array data to normalized gene expression levels. Specifically, the raw gene expression data in the custom CDF format (Dai *et al.*, 2005) were normalized using the robust multi-array average (RMA) method (Bolstad *et al.*, 2003; Irizarry *et al.*, 2003a,b). In total, gene expression levels of 18,807 genes with their Entrez IDs were obtained from 54,675 probe sets. the genotypes of the 906,551 SNP probes in the 84 tissue samples were transformed to values $\{0,1,2\}$ using the following mapping: AA $\rightarrow 0$, AB $\rightarrow 1$ and BB $\rightarrow 2$. The missing genotypes of SNP probes were imputed by randomly sampling from $\{0,1,2\}$ using the observed probabilities of genotypes of each SNP. Finally, R package MatrixEQTL (Shabalin, 2012) was adopted to identify *cis*-eQTLs of genes. In total, 1,456 genes were found to have at least one *cis*-eQTLs within 10^6 base pairs from the open reading frame (ORF) of the gene at an FDR $= 0.01$.

Since the number of samples available is 84, which may be too small to be used to reliably infer the network of 1,456 genes with eQTLs, we selected a subset of the 1,456 genes as follow. In (Greene *et al.*, 2015), gene interactions in 144 human tissues and cell types were inferred by integrating a collection of data sets covering thousands of experiments reported in more than 14,000 distinct publications. Each identified interaction between a pair of genes was assigned a confidence score or posterior probability in the Bayesian data integration process. The tissue-specific gene networks are all available in the GIANT (http://hb.flatironinstitute.org) database. For each pair of genes among the 1,456 genes with eQTLs, we searched the GIANT database to see if they interact with each other in the lung network constructed from the gene expression data. We identified the following 19 genes that interact with at least one another gene with high confidence (posterior probability ≥ 0.80): UBA2, CCT7, COX6B1, DBI, DKC1, ETFA, NACA, PSMC4, RPS6, SNRPF, BRIX1, KARS, ECHS1, ATP5G3, UBE2N, CDC123, VBP1, PSMD10, and BTF3. We extracted interactions among these 19 genes with posterior probability ≥ 0.80, and refer to this sub-network as the GIANT reference network. Since the dataset GSE33356 was not used to construct the GIANT network, we would compare the GRN inferred from the GSE33356 by our FSSEM algorithm with the GIANT network.

We applied the FSSEM algorithm to the 84 samples of expression levels of the 19 selected genes and the genotypes of their eQTLs to infer the networks of these genes in lung cancer and normal tissue. An edge from gene j to gene i was detected if $\mathbf{B}^{(k)}_{ij} \neq 0, k = 1,2$, where $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ specify the networks in normal and tumor tissues, respectively. FSSEM yielded a network matrix $\mathbf{B}^{(1)}$ with 93 nonzero entries, or a network with 93 edges, and these edges were regarded as significant gene interactions in normal tissues. This network referred to as the FSSEM network was compared with the GIANT reference network. It was found that 76.9% edges in the FSSEM network $\mathbf{B}^{(1)}$ were also in the GIANT network. This shows that the FSSEM network identified from a small independent samples is in good agreement with the GIANT reference network identified from a large number of data samples.

We also identified the differential network based on $\Delta\mathbf{B} = \mathbf{B}^{(2)} - \mathbf{B}^{(1)}$. Since small changes of coefficients $\mathbf{B}_{ij}$ may not have much biological effect, we regarded the regulatory effect of gene j and

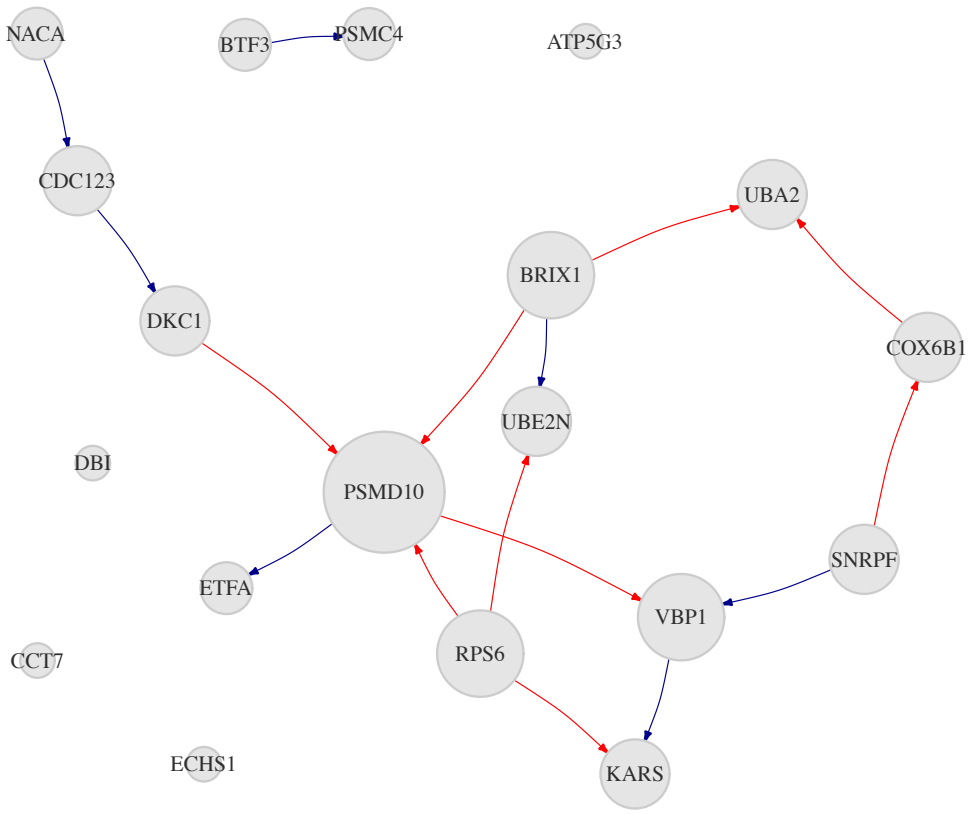

Figure 3. The differential regulatory network of 19 genes inferred from gene expression and eQTL data with the FSSEM algorithm. The size of a node is proportional to its degrees.

i to be different if $|\mathbf{B}_{ij}^{(1)} - \mathbf{B}_{ij}^{(2)}| > \min\{|\mathbf{B}_{ij}^{(1)}|, |\mathbf{B}_{ij}^{(2)}|\}$, which ensured that there is at least one-fold change relative to $\min\{|\mathbf{B}_{ij}^{(1)}|, |\mathbf{B}_{ij}^{(2)}|\}$. However, when one of $\mathbf{B}_{ij}^{(k)}, k = 1,2$ is zero or near zero, this criterion still fails to filter out very small changes $\Delta\mathbf{B}$. To avoid this issue, we added another criterion. Specifically, we obtained all nonzero entries of $\mathbf{B}^{(k)}, k = 1,2$, and compute the 20 percentile value of all nonzero $|\mathbf{B}_{ij}^{(k)}|, k = 1,2$ as η. Then, we defined the second criterion as $\max\{\mathbf{B}_{ij}^{(k)}, k = 1,2\} \geq \eta$. We employed the stability selection technique (Meinshausen and Bühlmann, 2010) to identify the differential network reliably. Specifically, we used 5-fold cross-validation to determine optimal values of λ and ρ, denoted as λ^* and ρ^*, respectively. A set of 21 samples are randomly selected from 42 cancer samples, and another set of 21 corresponding samples were selected from 42 normal samples. This data set of 42 samples was used by FSSEM algorithm to infer $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ with $\lambda = \lambda^*$ and $\rho = \rho^*$. The changed edges were identified based on $\Delta\mathbf{B}$ and the two criteria described earlier. This process of random sampling and network inference was repeated 100 times, and a changed edge was declared to be significant, if it was detected more than 80 times.

Stability selection yielded 16 edges that changed significantly. The differential network that is formed by these 16 edges is shown in Figure 3. Most edges in the differential network are connected with genes PSMD10, RPS6, BRIX1, VBP1 and SNRPF. As will be discussed in the next section, these five genes have been reported in a number of experimental results to be implicated in lung and

other cancers.

Discussion

In this paper, we developed a very efficient algorithm, named FSSEM, for joint inference of two similar GRNs by integrating genetic perturbations with gene expression data under two different conditions with the structural equation model. An R package implementing the FSSEM algorithm is available, which provides a useful tool for inferring GRNs. Computer simulations showed that our FSSEM offered much better accuracy in identifying changed gene interactions than the approach that infers two GRNs separately. Particularly, the FDR of gene interactions in the differential GRN estimated by FSSEM was significantly lower than that resulted from the method estimating two GRNs separately. This result is expected because FSSEM exploits the similarity in the two GRNs and penalizes the changes of gene interactions in the inference process.

Analysis of a data set of lung cancer and normal tissues with FSSEM detected most gene interactions identified in another study that exploited a large number of data sets. Real data analysis also identified several genes that may be involved in cancer development. Specifically, PSMD10 is aberrantly expressed in various cancers, and its expression level is inverse correlated with the expression level of miR-605 (Li *et al.*, 2014), which is repoted to be associated with lung cancer (Yin *et al.*, 2016). RPS6 is a component of the 40S ribosomal subunit; its expression has been shown to increase significantly in non-small cell lung cancer (NSCLC) (Chen *et al.*, 2015). Additionally, RPS6 is regulated in multiple signal pathways, such as the Akt2/mTOR/p70S6K signaling pathway (Yano *et al.*, 2014), that are closely related to the progression of NSCLC (Chen *et al.*, 2015). BRIX1 was identified as a key transcription factor associated with lung squamous cell carcinoma (Zhang *et al.*, 2017) and was also recognized as a key gene in the gastric cancer network (Kutmon *et al.*, 2015). SNRPF is a gene related to mRNA splicing pathway, and it was identified as a biomarker of ovarian cancer (Bengtsson *et al.*, 2007). Recently, it was reported that SNRPF was required for cells to tolerate oncogenic MYC hyperactivation (Hsu *et al.*, 2015). VBP1 is a gene identified to bind to the tumor suppressor gene VHL (Tsuchiya *et al.*, 1996), and it was reported that VBP1 repressed cancer metastasis by enhancing HIF1α degradation induced by pVHL (Kim *et al.*, 2018).

Funding

This work was supported by the National Science Foundation [Grant number CCF-1319981], and National Institute of General Medical Sciences [Grant number 5R01GM104975].
Conflict of interest: none declared.

References

Bengtsson, S., Krogh, M., Szigyarto, C. A.-K., Uhlen, M., Schedvins, K., Silfverswärd, C., Linder, S., Auer, G., Alaiya, A., and James, P. (2007). Large-scale proteomics analysis of human ovarian cancer for biomarkers. *Journal of proteome research*, **6**(4), 1440–1450.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, **146**(1-2), 459–494.

Butte, A. J. and Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific.

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology*, **9**(5), e1003068.

Califano, A. (2011). Rewiring makes the difference. *Molecular Systems Biology*, **7**(1).

Chen, B., Tan, Z., Gao, J., Wu, W., Liu, L., Jin, W., Cao, Y., Zhao, S., Zhang, W., Qiu, Z., *et al.* (2015). Hyperphosphorylation of ribosomal protein s6 predicts unfavorable clinical survival in non-small cell lung cancer. *Journal of Experimental & Clinical Cancer Research*, **34**(1), 126.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., *et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research*, **33**(20), e175–e175.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(2), 373–397.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, **5**(1), e8.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., *et al.* (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**(2), 302–332.

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

Gardner, T. S., Di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**(5629), 102–105.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy-analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**(3), 307–315.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., *et al.* (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, **47**(6), 569.

Harbison, 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**(7004), 99.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, **6**(1), 145.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, **19**(4), 984–1006.

Hsu, T. Y.-T., Simon, L. M., Neill, N. J., Marcotte, R., Sayad, A., Bland, C. S., Echeverria, G. V., Sun, T., Kurley, S. J., Tyagi, S., *et al.* (2015). The spliceosome is a therapeutic vulnerability in myc-driven cancer. *Nature*, **525**(7569), 384.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular systems biology*, **8**(1), 565.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003b). Summaries of affymetrix genechip probe level data. *Nucleic acids research*, **31**(4), e15–e15.

Kim, J., Choi, D. K., Min, J. S., Kang, I., Kim, J. C., Kim, S., and Ahn, J. K. (2018). Vbp1 represses cancer metastasis by enhancing hif-1α degradation induced by pvhl. *The FEBS journal*, **285**(1), 115–126.

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S. R., Miller, R., *et al.* (2015). Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic acids research*, **44**(D1), D488–D494.

Li, J., Tian, F., Li, D., Chen, J., Jiang, P., Zheng, S., Li, X., and Wang, S. (2014). Mir-605 represses psmd10/gankyrin and inhibits intrahepatic cholangiocarcinoma cell progression. *FEBS letters*, **588**(18), 3491–3500.

Liu, B., de La Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**(3), 1763–1776.

Logsdon, B. A. and Mezey, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS computational biology*, **6**(12), e1001014.

Lu, T.-P., Lai, L.-C., Tsai, M.-H., Chen, P.-C., Hsu, C.-P., Lee, J.-M., Hsiao, C. K., and Chuang, E. Y. (2011). Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PloS one*, **6**(9), e24829.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. BioMed Central.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.

Mohan, K., London, P., Fazel, M., Witten, D., and Lee, S.-I. (2014). Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, **15**(1), 445–488.

Neto, E. C., Ferrara, C. T., Attie, A. D., and Yandell, B. S. (2008). Inferring causal phenotype networks from segregating populations. *Genetics*, **179**(2), 1089–1100.

Pock, T. and Sabach, S. (2016). Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, **9**(4), 1756–1787.

Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, **28**(10), 1353–1358.

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K., and Kuijjer, M. L. (2017). Understanding tissue-specific gene regulation. *Cell Reports*, **21**(4), 1077–1088.

Statnikov, A. and Aliferis, C. F. (2010). Analysis and computational dissection of molecular signature multiplicity. *PLoS computational biology*, **6**(5), e1000790.

Tegner, J., Yeung, M. S., Hasty, J., and Collins, J. J. (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, **100**(10), 5944–5949.

Thieffry, D., Huerta, A. M., Pérez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, **20**(5), 433–440.

Tsuchiya, H., Iseda, T., and Hino, O. (1996). Identification of a novel protein (VBP-1) binding to the von Hippel-Lindau (VHL) tumor suppressor gene product. *Cancer Research*, **56**(13), 2881–2885.

Viallon, V., Lambert-Lacroix, S., Hoefling, H., and Picard, F. (2016). On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, **26**, 285–301.

Yano, T., Ferlito, M., Aponte, A., Kuno, A., Miura, T., Murphy, E., and Steenbergen, C. (2014). Pivotal role of mTORC2 and involvement of ribosomal protein S6 in cardioprotective signaling. *Circulation research*, **114**(8), 1268–1280.

Yin, Z., Li, H., Cui, Z., Ren, Y., Li, X., Wu, W., Guan, P., Qian, B., Rothman, N., Lan, Q., *et al.* (2016). Polymorphisms in pre-mirna genes and cooking oil fume exposure as well as their interaction on the risk of lung cancer in a chinese nonsmoking female population. *OncoTargets and therapy*, **9**, 395.

Zhang, F., Chen, X., Wei, K., Liu, D., Xu, X., Zhang, X., and Shi, H. (2017). Identification of key transcription factors associated with lung squamous cell carcinoma. *Medical science monitor: international medical journal of experimental and clinical research*, **23**, 172.

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., Sachs, J. R., and Schadt, E. E. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS computational biology*, **3**(4), e69.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Supplementary Text S

Hyper-parameter selection

We use K-fold cross-validation (CV) to determine the value of τ for ridge regression (4) and values of λ and ρ for FSSEM, where K typically equals to 5 or 10. We search τ over a sequence of 50 values increasing from 10^{-6} to 10^{2} evenly on the logarithm scale, and the optimal value of τ is chosen to minimize the predication error calculated from the test data. We employ a grid search strategy to determine the optimal values of λ and ρ. We first determine the maximum value of λ, namely $\lambda_{\max}$, then choose a set of k_1 values for λ, denoted as sequence $S_\lambda = \{\lambda_{\max}, \alpha_1\lambda_{\max}, \alpha_1^2\lambda_{\max}, ..., \alpha_1^{k_1-1}\lambda_{\max}\}$, where $0 < \alpha_1 < 1$. For each value of $\lambda \in S_\lambda$, we find the maximum value of ρ, namely $\rho_{\max}(\lambda)$, and then choose a set of k_2 values for ρ, denoted as $S_\rho(\lambda) = \{\rho_{\max}(\lambda), \alpha_2\rho_{\max}(\lambda), \alpha_2^2\rho_{\max}(\lambda), ..., \alpha_2^{k_2-1}\rho_{\max}(\lambda)\}$, where $0 < \alpha_2 < 1$. This gives a set of $K = k_1k_2$ pairs of (λ, ρ), and CV is carried out over this parameter space. The optimal values of λ and ρ are chosen to minimize the likelihood calculated from the test data.

Next, we derive the maximum values of λ and ρ needed in CV. The value $\lambda_{\max}$ yields $\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \mathbf{0}$, and can be found from the result in (Cai *et al.*, 2013) as follows:

$$\lambda_{\max} = \max_{i,j=1,...,n,k=1,2} \frac{|-n_k + \frac{1}{\sigma^2}(\tilde{\mathbf{Y}}^{(k)}\tilde{\mathbf{Y}}^{(k)T} - \mathbf{F}^{(k)}(\lambda_{\max})\tilde{\mathbf{X}}^{(k)}\tilde{\mathbf{Y}}^{(k)T})_{ij}|}{w_{ij}^{(k)}}, \tag{S1}$$

where $\mathbf{F}^{(k)}(\lambda_{\max})$ can be determined from (7) by setting $\hat{\mathbf{B}}_{i,-i}^{(k)} = \mathbf{0}, \forall i = 1, ..., n$. When $\rho = \rho_{\max}(\lambda)$, minimizing $J(\mathbf{B}, \mathbf{F})$ in (3) yields $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$. Therefore, we let $\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \mathbf{B}$, which yields

$$\begin{aligned} J(\mathbf{B}, \mathbf{F}) = &-\frac{n_1 + n_2}{2}\log|\mathbf{I} - \mathbf{B}|^2 + \frac{1}{2\hat{\sigma}^2}\sum_{k=1}^{2}||(\mathbf{I} - \mathbf{B})\tilde{\mathbf{Y}}^{(k)} - \mathbf{F}^{(k)}\tilde{\mathbf{X}}^{(k)}||_F^2 \\ &+ \lambda||\mathbf{B}||_{1,w^{(1)}+w^{(2)}} \\ \text{s.t. } &\mathbf{B}_{ii} = 0, \quad \forall i = 1, ..., n. \end{aligned} \tag{S2}$$

Then, we use the SML algorithm (Cai *et al.*, 2013) or the BCD approach of FSSEM mentioned in the main text to get $\hat{\mathbf{B}}$ that minimizes $J(\mathbf{B}, \mathbf{F})$ in (S2). Using the fact that when $\rho = \rho_{\max}(\lambda)$, the sub-gradient of $J(\mathbf{B}, \mathbf{F})$ in (3) w.r.t. $\hat{\mathbf{B}}_{i,j}$ equals to zero at $\hat{\mathbf{B}}^{(1)} = \hat{\mathbf{B}}^{(2)} = \hat{\mathbf{B}}$, we obtain

$$\left|\frac{N_k c_{ij}}{\det(\mathbf{I} - \hat{\mathbf{B}})} + \frac{1}{\sigma^2}((\mathbf{I} - \hat{\mathbf{B}})\tilde{\mathbf{Y}}^{(k)}\tilde{\mathbf{Y}}^{(k)T} - \mathbf{F}^{(k)}(\lambda)\tilde{\mathbf{X}}^{(k)}\tilde{\mathbf{Y}}^{(k)T})_{ij} + \lambda w_{ij}^{(k)}\partial|\hat{\mathbf{B}}_{i,j}^{(k)}|\right| \leq \rho_{\max}(\lambda)r_{ij} \tag{S3}$$

where $\partial(|\beta|)$ is the sub-gradient of $|\beta|$, and $\partial(|\beta|) = 1$, if $\beta > 0$, $\partial(|\beta|) = -1$, if $\beta < 0$, and $\partial(|\beta|) \in [-1, 1]$ if $\beta = 0$, and c_{ij} is the (i, j) co-factor of $\mathbf{I} - \hat{\mathbf{B}}$. From (S3), we obtain

$$\begin{aligned} &\rho_{\max}(\lambda) = \\ &\max_{i,j,k}\left|\frac{N_k c_{ij}}{\det(\mathbf{I} - \hat{\mathbf{B}})} + \frac{1}{\sigma^2}((\mathbf{I} - \hat{\mathbf{B}})\tilde{\mathbf{Y}}^{(k)}\tilde{\mathbf{Y}}^{(k)T} - \mathbf{F}^{(k)}(\lambda)\tilde{\mathbf{X}}^{(k)}\tilde{\mathbf{Y}}^{(k)T})_{ij} + \lambda w_{ij}^{(k)}\partial|\hat{\mathbf{B}}_{i,j}^{(k)}|\right|/r_{ij} \end{aligned} \tag{S4}$$

Convergence analysis

When the objective function in an optimization problem is non-convex and non-smooth, it is possible that the coordinate descent method fails to converge. We next prove that the FSSEM algorithm converges to a stationary point, because the objective function satisfies the conditions for the convergence of the PALM method specified in (Bolte *et al.*, 2014). Specifically, $J(\mathbf{B})$ in (15) has the following properties:

1. $\inf H(\mathbf{B}) > -\infty$ and $\inf f_i(\mathbf{B}_{i,-i}) > -\infty, i = 1,...n.$
2. $\nabla_{\mathbf{B}_{i,-i}} H(\mathbf{B}), i = 1,...,n$, is gradient Lipschitz continuous with constant $L_i(\mathbf{B}_{-i})$ when $\mathbf{B} \in \mathrm{dom}\, H = \{\mathbf{B} : \det(\mathbf{I} - \mathbf{B}^{(k)}) \neq 0, k = 1,2\}$:
$$||\nabla_{\mathbf{B}_{i,-i}} H(x, \mathbf{B}_{-i})) - \nabla_{\mathbf{B}_{i,-i}} H(y, \mathbf{B}_{-i}))|| \leq L_i(\mathbf{B}_{-i})\, ||x - y||,$$
where $\mathbf{B}_{-i} = \{\mathbf{B}^{(k)}_{j,-j}, j = 1,...,n, j \neq i, k = 1,2\}$.
3. $H(\mathbf{B})$ has continuous first and second derivatives when $\mathbf{B} \in \mathrm{dom}\, J = \{\mathbf{B} : \det(\mathbf{I} - \mathbf{B}^{(k)}) \neq 0, k = 1,2\}$.
4. $J(\mathbf{B})$ satisfies the Kurdyka-Łojasiewicz(KL) property.

Note that properties 1-3 are identical to the properties in assumption B of (Bolte *et al.*, 2014), and these 4 properties guarantee that FSSEM algorithm converges to a critical point. First, it is apparent that $H(\mathbf{B}) > -\infty$ and therefore $J(\mathbf{B}) > -\infty, \forall \mathbf{B} \in \mathrm{dom}\, J$. Second, it is not difficult to show that $H(\mathbf{B})$ is differentiable w.r.t. $\mathbf{B}_{i,-i}, i = 1,...,n$ and the first-order and second-order derivatives are continuous in $\mathrm{dom}\, H$. Therefore, property 3 is satisfied.

Third, we prove in the next section that $H(\mathbf{B}_{i,-i}, \mathbf{B}_{-i})$ is gradient Lipschitz continuous with constant $L_i(\mathbf{B}_{-i})$ given in (21). Moreover, based on assumption B(iii) of (Bolte *et al.*, 2014), $L_i(\mathbf{B}_{-i})$ guarantees that proximal steps in the FSSEM algorithm remain well-defined, because we have

$$\begin{aligned} \inf\{L_i(\mathbf{B}_{-i})\} &= \inf \max\{L_i(\mathbf{B}^{(k)}_{-i}), k = 1,2\} \geq \mu_i \\ \mu_i &= \max\{\lambda_{\max}(\tilde{\mathbf{Y}}^{(k)}_{-i} \mathbf{P}^{(k)}_i \tilde{\mathbf{Y}}^{(k)T}_{-i})/\sigma^2, k = 1,2\}. \end{aligned} \tag{S5}$$

Finally, we prove that property 4 holds. The non-smooth functions $f_i(\mathbf{B}_{i,-i})$ in $J(\mathbf{B})$ is the sparse fused lasso penalty term w.r.t. $\mathbf{B}_{i,-i}$, and it is semi-algebraic as shown in (Xu and Yin, 2013). The ℓ_2 norm $\sum_{k=1}^{2} ||\tilde{\mathbf{Y}}^{(k)}_i \mathbf{P}^{(k)}_i - \mathbf{B}^{(k)}_{i,-i} \tilde{\mathbf{Y}}^{(k)}_{-i} \mathbf{P}^{(k)}_i||_2^2$ is apparently semi-algebraic. We next prove that $-\sum_{k=1}^{2} \frac{n_k}{2} \log\det|\mathbf{I} - \mathbf{B}^{(k)}|^2$ is semi-algebraic too. We can regard $\frac{n_k}{2} \log\det|\mathbf{I} - \mathbf{B}^{(k)}|^2, k = 1,2$, as a composite function of $\mathbf{B}^{(k)}$ as follows

$$\begin{aligned} -\frac{n_k}{2} \log|\mathbf{I} - \mathbf{B}^{(k)}|^2 &= (g \circ F)(\mathbf{B}^{(k)}) \\ g(\cdot) &= -\frac{n_k}{2} \log\det(\cdot) \\ F(\cdot) &= (\mathbf{I} - \cdot)^T (\mathbf{I} - \cdot), \end{aligned} \tag{S6}$$

Function $g(\cdot)$ is locally convex function (Boyd and Vandenberghe, 2004). Based on the result in (Xu and Yin, 2013), it is not difficult to show that $g(\cdot)$ satisfies the KL property, and it can be shown that function $F : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ is continuously differentiable in $\mathrm{dom}\, J$. As all terms of $J(\mathbf{B})$ are KL functions, the sum of theses KL functions should satisfy the KL property (Li and Pong, 2017). This completes the proof that $J(\mathbf{B})$ satisfies properties 1-4.

Derivation of the Lipschitz constant of $\nabla_{\mathbf{B}_{i,-i}} H(\mathbf{B})$

In this section, we derive the Lipschitz constant of $\nabla_{\mathbf{B}_{i,-i}} H(\mathbf{B})$ in (16), where we drop index t in $B[t]$ for notational simplicity. From (16), we obtain the following:

$$\begin{aligned} & \left|\left| \nabla_{\mathbf{B}_{i,-i}} H(x, \mathbf{B}_{-i}) - \nabla_{\mathbf{B}_{i,-i}} H(y, \mathbf{B}_{-i}) \right|\right| \\ = & \left|\left| \frac{-n_k (y-x)\mathbf{c}_i \mathbf{c}_i^T}{(\mathbf{c}_{ii} - x\mathbf{c}_i)(\mathbf{c}_{ii} - y\mathbf{c}_i)} + \frac{1}{\sigma^2}(y-x)\tilde{\mathbf{Y}}_{-i}\mathbf{P}_i\tilde{\mathbf{Y}}_{-i}^T \right|\right| \\ \leq & \left(\frac{n_k ||\mathbf{c}_i||_2^2}{\min\limits_{\mathbf{B}_{i,-i}} \det(\mathbf{I}-\mathbf{B})^2} + \lambda_{\max}(\tilde{\mathbf{Y}}_{-i}\mathbf{P}_i\tilde{\mathbf{Y}}_{-i}^T)/\sigma^2 \right) ||y-x||, \end{aligned} \tag{S7}$$

where $\min_{\mathbf{B}_{i,-i}} \det(\mathbf{I}-\mathbf{B})$ is the minimal value of $\det(\mathbf{I}-\mathbf{B})$ for a given $\mathbf{B}_{-i} \in \operatorname{dom} J$, and $\lambda_{\max}(\tilde{\mathbf{Y}}_{-i}\mathbf{P}_i\tilde{\mathbf{Y}}_{-i}^T)$ is the maximum eigenvalue of $\tilde{\mathbf{Y}}_{-i}\mathbf{P}_i\tilde{\mathbf{Y}}_{-i}^T$. The Lipschitz constant of $\nabla_{\mathbf{B}_{i,-i}} H(\mathbf{B})$ is then given by

$$L_i(\mathbf{B}_{-i}) = \frac{n_k ||\mathbf{c}_i||_2^2}{\min\limits_{\mathbf{B}_{i,-i}} \det(\mathbf{I}-\mathbf{B})^2} + \lambda_{\max}(\tilde{\mathbf{Y}}_{-i}\mathbf{P}_i\tilde{\mathbf{Y}}_{-i}^T)/\sigma^2. \tag{S8}$$

The value of $\min_{\mathbf{B}_{i,-i}} \det(\mathbf{I}-\mathbf{B})^2$ can be determined as follows.

Define $\Theta = \mathbf{I} - \mathbf{B}$, and let θ_i^T be the ith row of Θ, and Θ_{-i} be the sub-matrix of Θ that excludes θ_i^T. Then, we have,

$$\det(\Theta)^2 = \det(\Theta\Theta^T) = \det(\Theta_{-i}\Theta_{-i}^T) \times \theta_i^T (\mathbf{I} - \Theta_{-i}^T(\Theta_{-i}\Theta_{-i}^T)^{-1}\Theta_{-i})\theta_i, \tag{S9}$$

where $\Theta_{-i,-i}$ is the submatrix of Θ excluding the ith row and the ith column. Here we assume $\mathbf{B} \in \operatorname{dom} J$, and thus $\det(\Theta\Theta^T) > 0$. Since $\Theta_{-i}\Theta_{-i}^T$ is a submatrix of $\Theta\Theta^T$, the Cauchy's interlacing theorem (Hwang, 2004) implies $\det(\Theta_{-i}\Theta_{-i}^T) > 0$. Therefore, $(\Theta_{-i}\Theta_{-i}^T)^{-1}$ in (S9) exists.

For notational simplicity, we let $b_i = \mathbf{B}_{i,-i}^T$ and write (S9) as

$$\det(\Theta)^2 = \det(\Theta_{-i}\Theta_{-i}^T)(1 + ||b_i||_2^2 - (\Theta_{-i,i}^T - b_i^T\Theta_{-i,-i}^T)(\Theta_{-i}\Theta_{-i}^T)^{-1}(\Theta_{-i,i} - \Theta_{-i,-i}b_i)). \tag{S10}$$

Minimizing $\det(\Theta)^2$ in (S10) w.r.t. b_i yields

$$\hat{b}_i = -(\mathbf{I} - \Theta_{-i,-i}^T(\Theta_{-i}\Theta_{-i}^T)^{-1}\Theta_{-i,-i})^{-1}\Theta_{-i,-i}^T(\Theta_{-i}\Theta_{-i}^T)^{-1}\Theta_{-i,i}. \tag{S11}$$

Substituting $\hat{b}_i$ into (S10) gives the minimum value of $\det(\Theta)^2$. In practice, to ensure numerical stability, we modify the $\hat{b}_i$ in (S11) as follows,

$$\hat{b}_i = -(\mathbf{I} - \Theta_{-i,-i}^T(\Theta_{-i}\Theta_{-i}^T)^{-1}\Theta_{-i,-i} + \zeta\mathbf{I})^{-1}\Theta_{-i,-i}^T(\Theta_{-i}\Theta_{-i}^T)^{-1}\Theta_{-i,i}, \tag{S12}$$

where ζ is a small positive constant. This modification can be regarded as minimizing $(\det(\Theta))^2$ in (S11) subject to the constraint $||b_i||^2 \leq c$, where c is a positive constant. This is a reasonable assumption because in real GRNs, entries of $\mathbf{B}^{(k)}$ are bounded. In our implementation, we chose $\zeta = 10^{-16}$ and we did not observe any numerical instability in all of our numerical experiments.

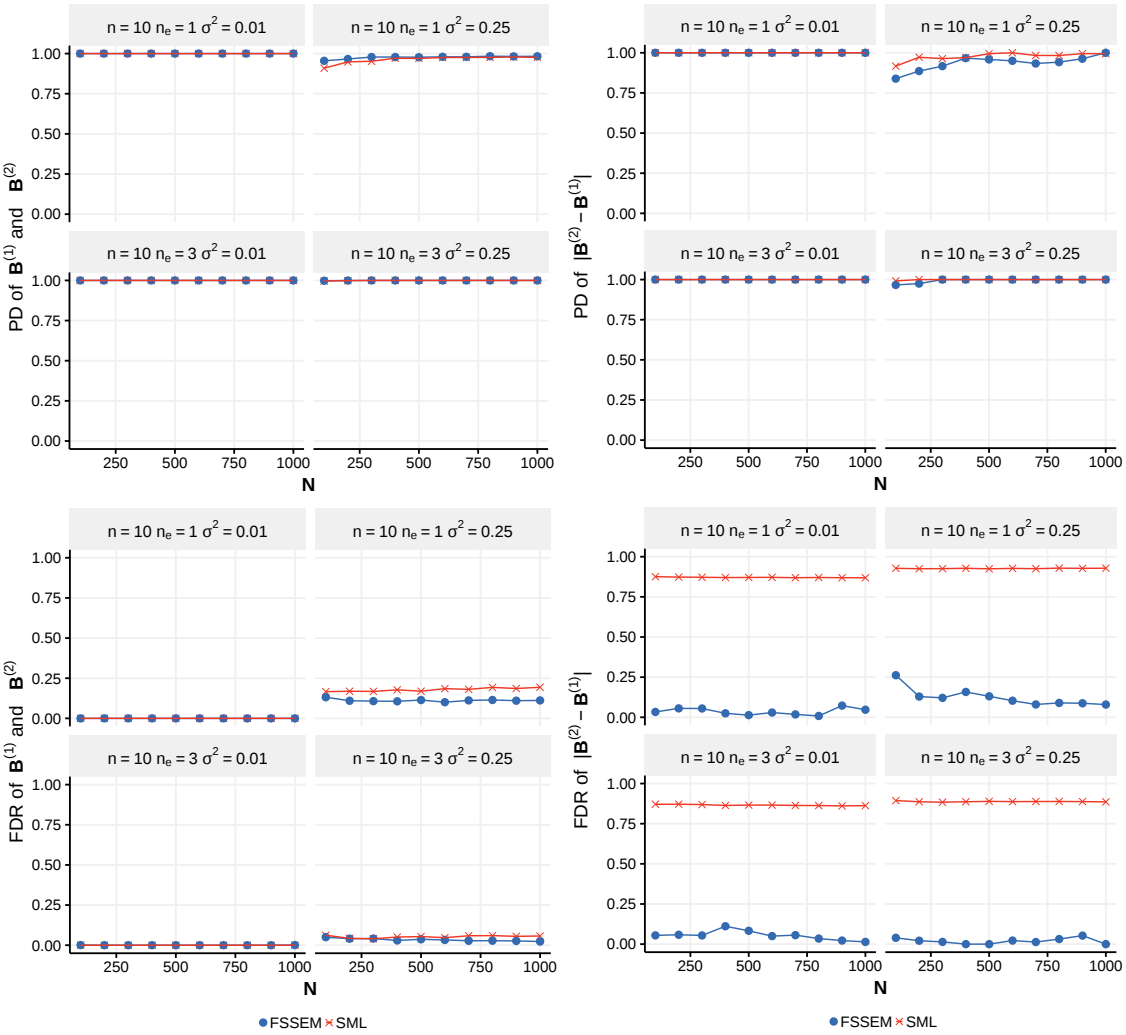

Figure S1. Performance of FSSEM and SML for the DAG with $n = 10$ genes. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.01, 0.25$. PD and FDR were obtained from 30 network replicates.

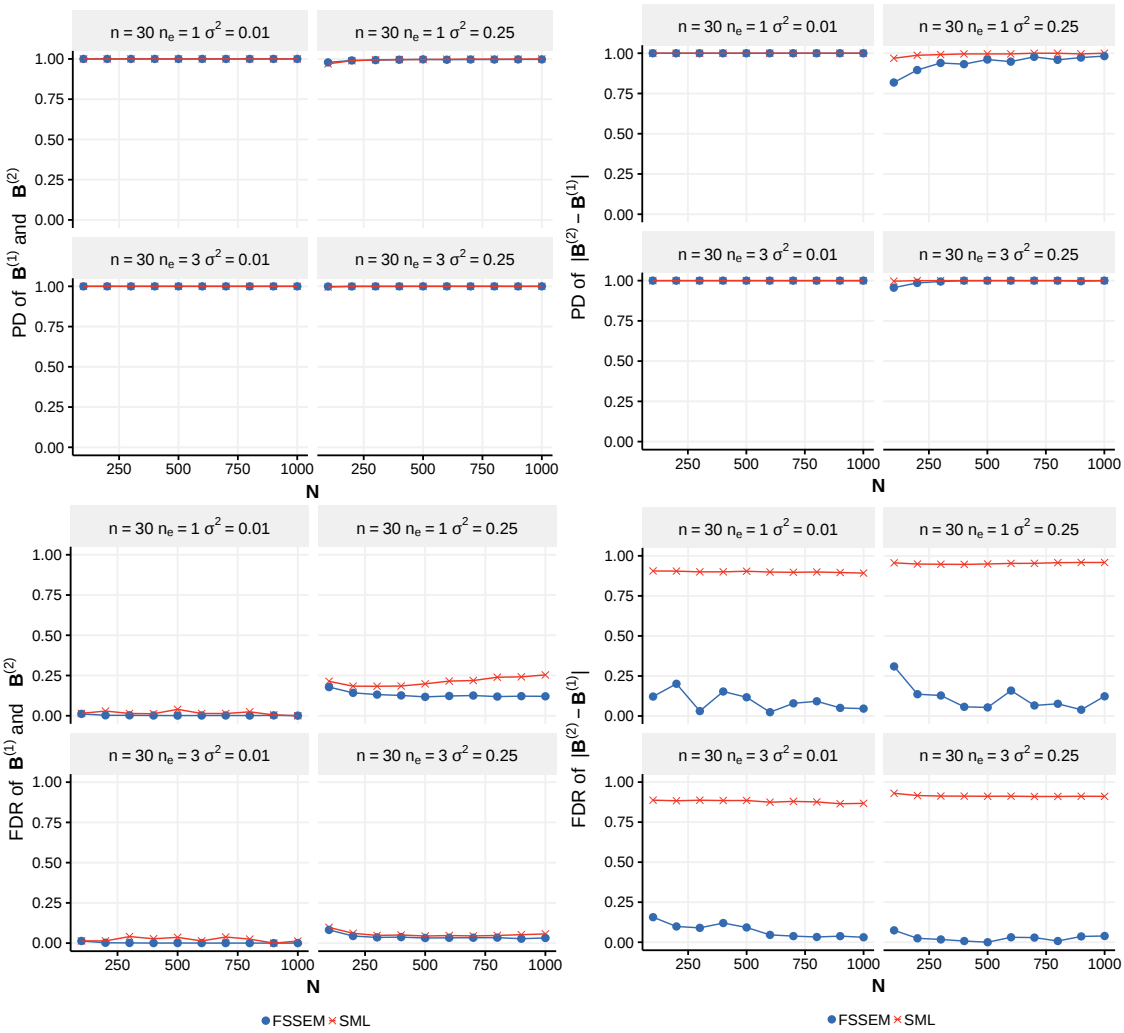

Figure S2. Performance of FSSEM and SML for the DAG with $n = 30$ genes. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.01, 0.25$. PD and FDR were obtained from 30 network replicates.

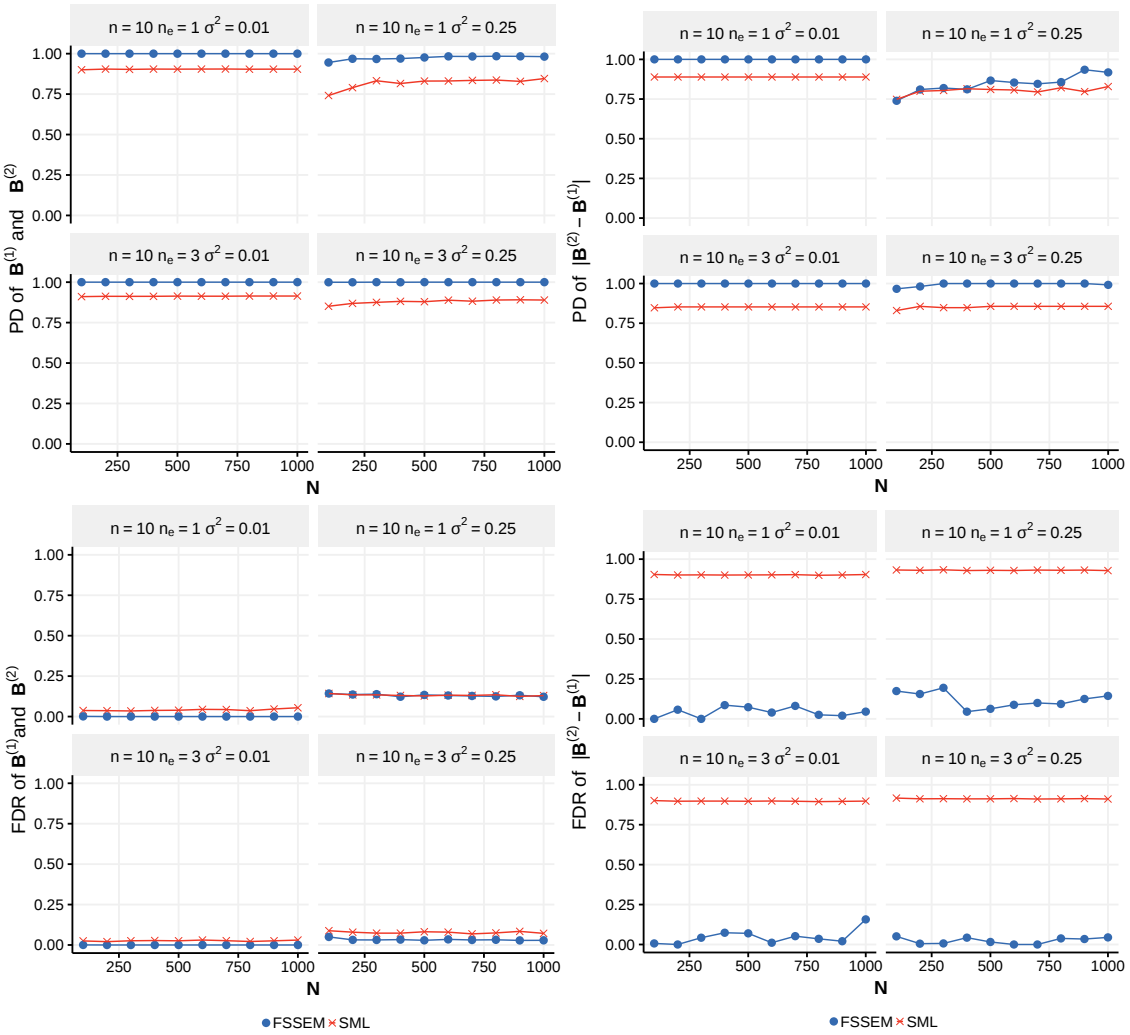

Figure S3. Performance of FSSEM and SML for the DCG with $n = 10$ genes. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.01, 0.25$. PD and FDR were obtained from 30 network replicates.

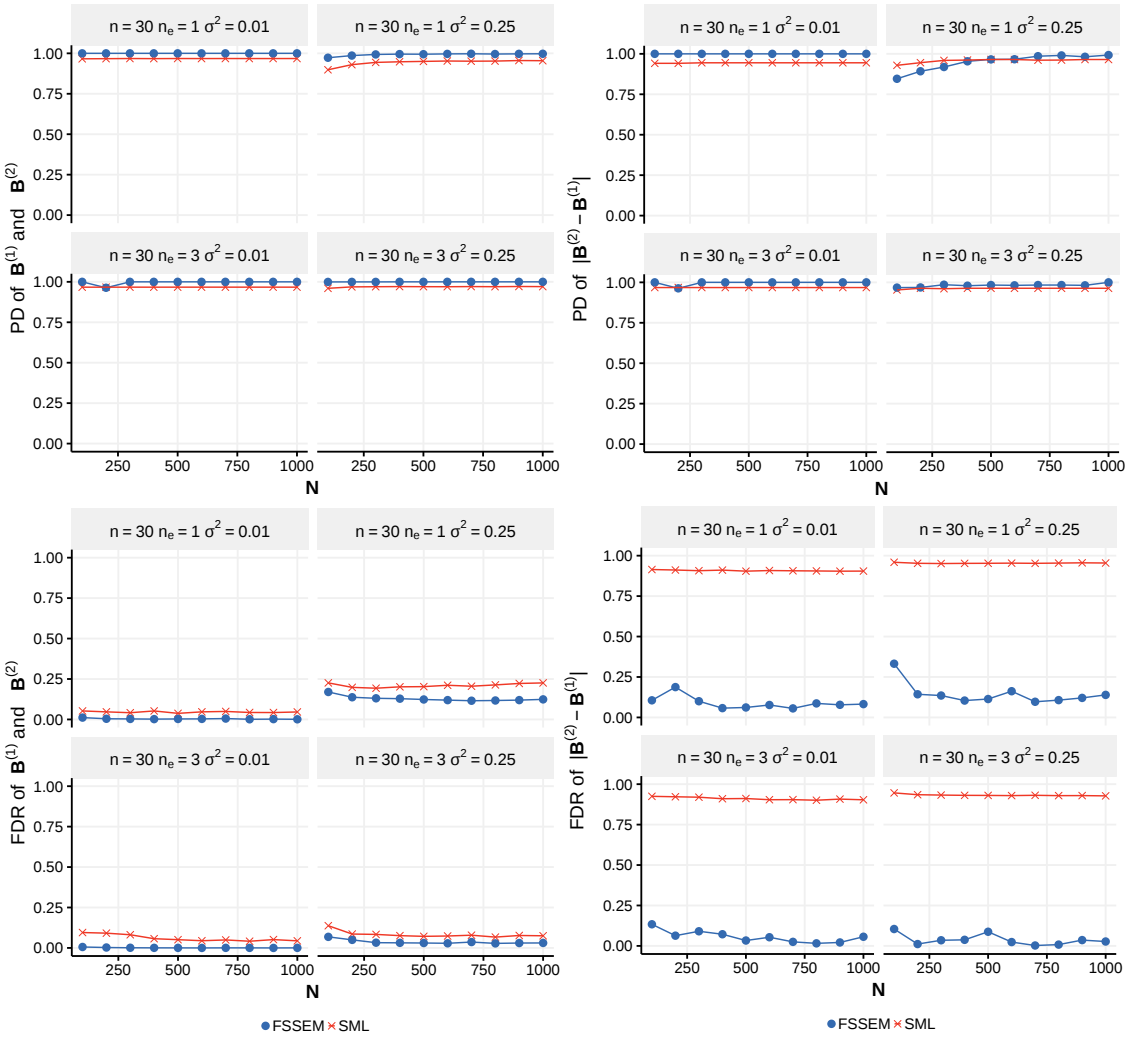

Figure S4. Performance of FSSEM and SML for the DCG with $n = 30$ genes. The number of samples $n_1 = n_2$ varies from 100 to 1,000 and noise variance $\sigma^2 = 0.01, 0.25$. PD and FDR were obtained from 30 network replicates.

References

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, **146**(1-2), 459–494.

Boyd, S., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Hwang, S. G. Cauchy's interlace theorem for eigenvalues of hermitian matrices. *The American mathematical monthly*, 111(2):157–159, 2004.

Li, G., and Pong, T.K. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, pages 1–34, 2017.

Xu, Y., and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.