



# Fast hybrid Bayesian integrative learning of multiple gene regulatory networks for type 1 diabetes

BOCHAO JIA

*Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA*

FAMING LIANG\*, The TEDDY Study Group

*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*

fmlaing@purdue.edu

## SUMMARY

Motivated by the study of the molecular mechanism underlying type 1 diabetes with gene expression data collected from both patients and healthy controls at multiple time points, we propose a hybrid Bayesian method for jointly estimating multiple dependent Gaussian graphical models with data observed under distinct conditions, which avoids inversion of high-dimensional covariance matrices and thus can be executed very fast. We prove the consistency of the proposed method under mild conditions. The numerical results indicate the superiority of the proposed method over existing ones in both estimation accuracy and computational efficiency. Extension of the proposed method to joint estimation of multiple mixed graphical models is straightforward.

*Keywords:* Data integration; Meta-analysis; Multiple Gaussian graphical models;  $\psi$ -Learning.

## 1. INTRODUCTION

Type 1 diabetes (T1D) is one of the most common autoimmune diseases. The Environmental Determinants of Diabetes in the Young (TEDDY) study is designed to identify environmental exposures triggering islet autoimmunity and T1D in genetically high-risk children. A large dataset has been collected through the study, including clinical data, genetic data, and demographical data. While great efforts have been made for identifying the genetic and environmental factors that contribute to the etiology of the disease, the molecular mechanism underlying the disease is still far from understanding. To enhance our understanding to the molecular mechanism, this work aims to learn gene regulatory networks (GRNs) by integrating the gene expression data measured from both the patients and healthy controls at multiple time points. Figure 1 shows the structure of the data, where the gene expression was measured for each of the case and control children at nine time points within 4 years of age. How to integrate the data collected under the 18 distinct conditions have posed a great challenge on the current statistical methods.

During the past decade, a variety of approaches have been proposed for estimating multiple GRNs with data collected under multiple distinct conditions. These approaches can be roughly grouped into two categories, namely, regularization and Bayesian.

\*To whom correspondence should be addressed.

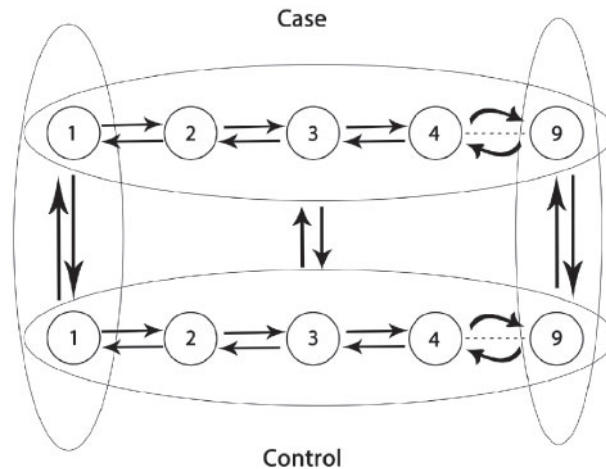


Fig. 1. Structure of the T1D data considered in the article, where the numbers represent nine time points at which gene expression data were collected and the arrows represent joint estimation of Gaussian graphical models by integrating the data across different time points and case–control groups.

The regularization approaches work with some specific penalty functions that enhance the shared structure of the graphical models. For example, [Guo and others \(2011\)](#) employed a hierarchical penalty that targets the removal of common zeros in the precision matrices across conditions. [Danaher and others \(2014\)](#) employed penalized fused lasso or group lasso penalties that encourage shared elements of the precision matrices. [Chun and others \(2015\)](#) employed a class of nonconvex penalty functions that regularize the common and condition-specific structures hierarchically. A shortcoming of these approaches is that they assume the observations under different conditions are independent. This is hard to be satisfied for the temporal data, where the observations were taken from the same cohort at multiple time points. To address this issue, [Zhou and others \(2010\)](#) and [Qiu and others \(2016\)](#) proposed to model the temporal data in a high-dimensional time series and then estimate the time-varying graphical structure using a nonparametric method by assuming that the covariance changes smoothly over time. These approaches usually require the time series to be fairly long, say, 50 or longer.

As an analog to regularization approaches, Bayesian approaches enhance the shared structure of multiple graphical models by employing some specific priors. For example, [Peterson and others \(2015\)](#) links the estimation of graph structures via a Markov random field (MRF) prior which encourages common edges. However, since this method involves repeated calculations of concentration matrices (i.e., inverse of covariance matrices), it is only applicable when the graph is not very large. To accelerate computation, [Lin and others \(2017\)](#) proposed a Bayesian analog of the neighborhood selection method ([Meinshausen and Bühlmann, 2006](#)) to learn the structure of multiple graphical models with the MRF prior.

In this article, we propose a fast hybrid Bayesian integrative analysis (FHBIA) method for jointly estimating multiple Gaussian graphical models. The proposed method consists of both frequentist and Bayesian components. First, it applies a  $\psi$ -learning method, which is a frequentist method, to transform the original data to edge-wise  $\psi$ -scores. The  $\psi$ -score, which forms an equivalent measure of the partial correlation coefficient, provides a good summary for the graph structure information contained in the data under each condition. Then, it applies a Bayesian method to model the  $\psi$ -scores for edge clustering and applies a meta-analysis method for integrating data information across distinct conditions. Finally, it applies a multiple hypothesis test method for edge determination. Due to the use of the  $\psi$ -score transformation, FHBIA avoids inversion of high-dimensional covariance matrices and thus can be executed

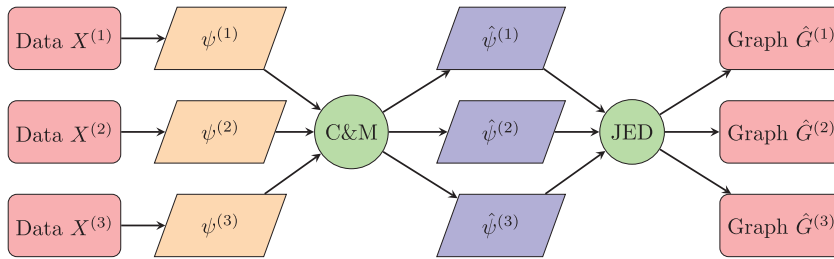


Fig. 2. Diagram of the FHBIA method: (i) datasets  $X^{(k)}$ 's for  $k = 1, 2, \dots, K$  are first transformed to edgewise  $\psi$ -scores  $\psi^{(k)}$ 's through the step of  $\psi$ -score transformation; (ii)  $\psi$ -scores are processed through the step of Bayesian clustering and meta-analysis to get Bayesian integrated  $\psi$ -scores denoted by  $\hat{\psi}^{(k)}$ 's; (iii) Bayesian integrated  $\psi$ -scores are further processed through the step of JED to get graph estimates  $\hat{G}^{(k)}$ 's.

very fast. The multiple hypothesis test produces a  $q$ -value (Storey, 2002), which can be viewed as an uncertainty measure, for each potential edge of the multiple Gaussian graphs. We prove consistency of the proposed method under mild conditions and illustrate its performance using simulated and real data examples. The numerical results indicate the superiority of the proposed method over the existing ones in both estimation accuracy and computational efficiency.

## 2. FAST HYBRID BAYESIAN INTEGRATIVE ANALYSIS

The FHBIA method consists of a few steps, including  $\psi$ -score calculation, Bayesian clustering and meta-analysis, and joint edge detection (JED), with the diagram shown in Figure 2.

### 2.1. $\psi$ -Score transformation

Suppose that we have a dataset of  $p$  variables observed under  $K$  distinct conditions. Let  $\mathbf{X}^{(k)} = (\mathbf{X}_1^{(k)}, \dots, \mathbf{X}_{n_k}^{(k)})^T$  denote the dataset observed under condition  $k$ , where  $n_k$  denotes the sample size under condition  $k$ ; and  $\mathbf{X}_i^{(k)} = (X_{i1}^{(k)}, \dots, X_{ip}^{(k)})^T$  is a  $p$ -dimensional random vector distributed according to the multivariate normal distribution  $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ , and  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}_k$  are the mean and covariance matrix of the distribution, respectively. The sample size  $n_k$  is not necessarily the same for all conditions. Without loss of generality, we assume that  $\boldsymbol{\mu}_k$  is a zero vector for all  $k$ . With slight abuse of notation, we let  $X_1, \dots, X_p$  denote the  $p$  variables that are common for all  $K$  datasets. Let  $V = \{1, 2, \dots, p\}$  denote the index set of the variables, where each variable is called a node in the terminology of graphs.

We adopt the  $\psi$ -learning algorithm to transform each dataset  $\mathbf{X}^{(k)}$  to edge-wise scores independently. The  $\psi$ -learning algorithm first produced a  $\tilde{\psi}$ -partial correlation coefficient for each pair of nodes. Denote the  $\psi$ -partial correlation coefficients by  $(\tilde{\psi}_{ij})$  for all  $i, j \in V$ , which are equivalent to the true partial correlation coefficients  $(\rho_{ij|V \setminus \{i,j\}})$  for determining the structure of the Gaussian Graphical Model (GGM) in the sense that  $\tilde{\psi}_{ij} = 0 \iff \rho_{ij|V \setminus \{i,j\}} = 0, \forall i, j \in V$ . Further, the  $\psi$ -learning algorithm converts the  $\psi$ -partial correlation coefficients to  $\psi$ -scores, denoted by  $(\psi_{ij})$ , via Fisher's transformation and the probit transformation such that  $\psi_{ij} \sim N(0, 1)$  approximately holds under the null hypothesis  $H_0 : \rho_{ij|V \setminus \{i,j\}} = 0$  for any  $i, j \in V$ . Therefore, the  $\psi$ -score can be used as a test statistic for identifying nonzero partial correlation coefficients and thus the structure of the GGM. The use of  $\psi$ -scores enables the proposed method to avoid inversion of high-dimensional covariance matrices that the existing Bayesian methods often need to deal with, and hence the proposed method can be executed very fast. Refer to the supplementary material available at *Biostatistics* online for the detail of the  $\psi$ -learning algorithm.

Since the GGM is undirected, we have a total of  $p(p-1)/2$   $\psi$ -scores to calculate for each dataset  $X^{(k)}$ . For convenience, we re-arrange the  $\psi$ -scores for each dataset  $X^{(k)}$  into an  $N$ -vector  $\psi^{(k)}$  with  $N = p(p-1)/2$  and re-arrange the  $\psi$ -scores for all  $K$  datasets into an  $N \times K$  matrix  $(\psi_l^{(k)})$  with  $l = 1, 2, \dots, N$  and  $k = 1, 2, \dots, K$ .

## 2.2. Bayesian clustering and meta-analysis

Consider the  $\psi$ -scores  $(\psi_l^{(k)})$ , where each pair  $(l, k)$  corresponds to one candidate edge in the graph  $G^{(k)}$ . Let  $e_l^{(k)}$  be the indicator for the status of edge  $l$  in the underlying graph  $G^{(k)}$ ;  $e_l^{(k)} = 1$  if the edge exists and 0 otherwise. The  $e_l^{(k)}$ 's work as latent variables in FHBIA. Conditioned on  $e_l^{(k)}$ , we assume that  $\psi_l^{(k)}$ 's are mutually independent and follow a two-component mixture Gaussian distribution given by

$$p(\psi_l^{(k)} | e_l^{(k)}) = \begin{cases} N(\mu_{l0}, \sigma_{l0}^2), & \text{if } e_l^{(k)} = 0, \\ N(\mu_{l1}, \sigma_{l1}^2), & \text{if } e_l^{(k)} = 1, \end{cases} \quad (2.1)$$

for  $l = 1, 2, \dots, N$  and  $k = 1, 2, \dots, K$ . When  $e_l^{(k)} = 0$ ,  $\psi_l^{(k)}$ 's have a value close to 0, otherwise,  $\psi_l^{(k)}$ 's might have a large negative or positive value depending on the sign of the partial correlation coefficient. Under the assumption that the structure of the GGM changes only slightly under adjacent conditions, it is reasonable to assume that for each  $l$ , the sign of  $\psi_l^{(k)}$ 's are not changed when the edge exists; therefore,  $\psi_l^{(k)}$ 's can be modeled by a two-component mixture Gaussian distribution. In some cases, e.g., when  $K$  grows, a three-component mixture Gaussian distribution might be needed, which allows us to handle the scenario that an edge is included in multiple graphs, but its partial correlation coefficients have different signs in different graphs. The derivation under this scenario is given in the supplementary material available at *Biostatistics* online, which is a simple extension of the deviation presented below. Regarding the two-component mixture distribution (2.1), we further note that  $\mu_{l0}$  can be simply set to 0 considering the physical mean of  $\psi$ -scores. However, as shown below, this general setup does not cause any computational difficulty.

Essentially, we have formulated the problem of inference of  $e_l^{(k)}$ 's as a clustering problem, grouping  $\psi_l^{(k)}$  to up to two different clusters. For the case of three-component mixture distribution, this is similar. Let  $\boldsymbol{\psi}_l = (\psi_l^{(1)}, \dots, \psi_l^{(K)})$  and  $\mathbf{e}_l = (e_l^{(1)}, \dots, e_l^{(K)})$ . Conditioned on  $\mathbf{e}_l$ , the joint likelihood function of  $\boldsymbol{\psi}_l$  is given by

$$p(\boldsymbol{\psi}_l | \mathbf{e}_l, \mu_{l0}, \sigma_{l0}^2, \mu_{l1}, \sigma_{l1}^2) = \prod_{\{k: e_l^{(k)}=0\}} \phi(\psi_l^{(k)} | \mu_{l0}, \sigma_{l0}^2) \prod_{\{k: e_l^{(k)}=1\}} \phi(\psi_l^{(k)} | \mu_{l1}, \sigma_{l1}^2), \quad (2.2)$$

where  $\phi(\cdot | \mu, \sigma^2)$  is the density function of the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Taking a product of (2.2) over  $l = 1, 2, \dots, N$ , we have the joint distribution of all  $\psi$ -scores  $(\boldsymbol{\psi}_l^{(k)})$  conditioned on  $\mathbf{e}_l$ 's and other parameters. Then, using the Bayes theorem,  $e_l^{(k)}$ 's can be inferred with an appropriate priors of  $\mathbf{e}_l$ 's and other parameters. For example, the MRF prior used in [Peterson and others \(2015\)](#) and [Lin and others \(2017\)](#) can again be used here as the prior of  $\mathbf{e}_l$ 's. In this case, the posterior distribution can be sampled from using a Monte Carlo Markov chain (MCMC) algorithm.

Instead of specifying a joint prior distribution for all  $\mathbf{e}_l$ 's, we assume in this article that  $\mathbf{e}_l$ 's are *a priori* independent for different  $l$ 's, as we believe that the neighboring dependence of the Gaussian graphical network has been accounted for in calculation of the  $\psi$ -scores. To enhance shared edges among distinct conditions, we consider two types of priors for  $\mathbf{e}_l$ 's, namely, temporal prior and spatial prior, with borrowed terms from geostatistics. Figure 3 illustrates the application scenarios of the two types of priors. The temporal prior can be used in the scenario that the networks  $G^{(k)}$ 's evolve sequentially along with the index

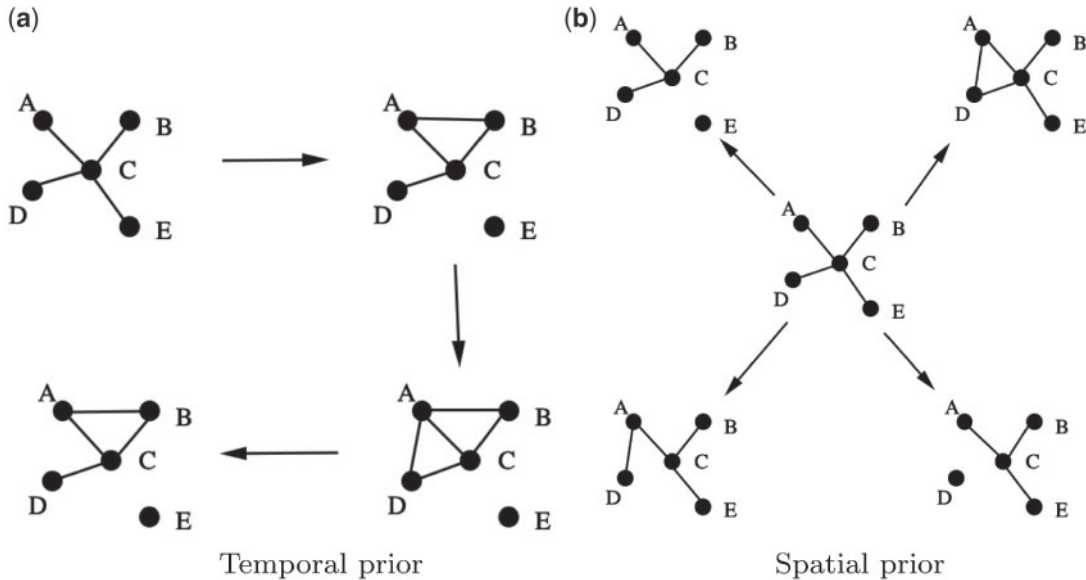


Fig. 3. Illustration of application scenarios of the temporal and spatial priors: (a) networks evolve along with time (temporal prior); (b) networks evolve independently from a common structure (spatial prior).

$k = 1, 2, \dots, K$ . In this scenario, it is quite common to consider the index  $k$  as the time of experiments. The spatial prior can be used in the scenario that the networks or precision matrices evolve independently from a common structure. For example, the temporal prior can be applied when we construct genetic networks using a set of gene expression data measured for the same tissue at multiple time points, and the spatial prior can be applied if the gene expression data are measured for different tissues at the same time point.

2.2.1. *Temporal prior* To enhance the similarity of the networks between adjacent conditions, we let  $e_l$  be subject to the following prior distribution

$$p(e_l|q) = q^{\sum_{i=1}^{K-1} c_l^{(i)}} (1 - q)^{\sum_{i=1}^{K-1} (1 - c_l^{(i)})}, \tag{2.3}$$

where  $c_l^{(i)} = |e_l^{(i+1)} - e_l^{(i)}|$  indicates the change of the status of the edge  $l$  from condition  $i$  to condition  $i + 1$ , and  $q$  is a prior hyperparameter representing the prior probability of edge status changes. In this article, we assume that  $q$  follows a beta distribution  $Beta(a_1, b_1)$ , where  $a_1$  and  $b_1$  are pre-specified parameters. Further, we let  $\mu_{j0}$  and  $\mu_{j1}$  be subject to an improper uniform distribution, i.e.,  $\pi(\mu_{j0}) \propto 1$  and  $\pi(\mu_{j1}) \propto 1$ , and let  $\sigma_{j0}^2$  and  $\sigma_{j1}^2$  be subject to an inverted-gamma distribution, i.e.,  $\sigma_{j0}^2, \sigma_{j1}^2 \sim IG(a_2, b_2)$ , where  $a_2$  and  $b_2$  are pre-specified constants. Then the joint posterior distribution of  $(e_l, \mu_{j0}, \sigma_{j0}^2, \mu_{j1}, \sigma_{j1}^2, q)$  is given by

$$\pi(e_l, \mu_{j0}, \sigma_{j0}^2, \mu_{j1}, \sigma_{j1}^2, q | \psi_l) \propto p(\psi_l | e_l, \mu_{j0}, \mu_{j1}, \sigma_{j0}^2, \sigma_{j1}^2) \pi(\mu_{j0}, \sigma_{j0}^2, \mu_{j1}, \sigma_{j1}^2) \pi(e_l | q) \pi(q),$$

where  $\pi(\cdot)$ 's denote the respective prior distributions. After integrating out the parameters  $\mu_{j0}, \sigma_{j0}^2, \mu_{j1}, \sigma_{j1}^2$ , and  $q$ , we have the marginal posterior distribution of  $e_l$  given by

$$\begin{aligned}
 \pi(\mathbf{e}_l | \boldsymbol{\psi}_l) &\propto \underbrace{\frac{\Gamma(a_1 + k_1)\Gamma(b_1 + k_2)}{\Gamma(a_1 + k_1 + b_1 + k_2)}}_{(H)} \\
 &\times \underbrace{\frac{1}{\sqrt{n_{l0}}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n_{l0}} \Gamma\left(\frac{n_{l0} - 1}{2} + a_2\right) \left[ \frac{1}{2} \sum_{\{k: e_l^{(k)}=0\}} (\psi_l^{(k)})^2 - \frac{(\sum_{\{k: e_l^{(k)}=0\}} \psi_l^{(k)})^2}{2n_{l0}} + b_2 \right]}_{(I)}^{-\left(\frac{n_{l0}-1}{2} + a_2\right)} \\
 &\times \underbrace{\frac{1}{\sqrt{n_{l1}}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n_{l1}} \Gamma\left(\frac{n_{l1} - 1}{2} + a_2\right) \left[ \frac{1}{2} \sum_{\{k: e_l^{(k)}=1\}} (\psi_l^{(k)})^2 - \frac{(\sum_{\{k: e_l^{(k)}=1\}} \psi_l^{(k)})^2}{2n_{l1}} + b_2 \right]}_{(J)}^{-\left(\frac{n_{l1}-1}{2} + a_2\right)} \quad (2.4)
 \end{aligned}$$

when  $n_{l0} > 0$  and  $n_{l1} > 0$  hold, where  $n_{l0} = \#\{k : e_l^{(k)} = 0\}$ ,  $n_{l1} = \#\{k : e_l^{(k)} = 1\}$ ,  $k_1 = \sum_{i=1}^{K-1} c_l^{(i)}$  and  $k_2 = K - 1 - k_1$ . When  $n_{l0} = 0$  and  $n_{l1} > 0$ , we have  $\pi(\mathbf{e}_l | \boldsymbol{\psi}_l) \propto (H) \times (J)$ . When  $n_{l0} > 0$  and  $n_{l1} = 0$ , we have  $\pi(\mathbf{e}_l | \boldsymbol{\psi}_l) \propto (H) \times (I)$ .

Given  $K$  distinct conditions, the total number of possible configurations of  $\mathbf{e}_l$  is  $2^K$ . When  $K$  is small, we can provide an exhaustive evaluation of the  $2^K$  configurations. That is, for each possible configuration of  $\mathbf{e}_l$ , we can calculate its posterior probability and integrated  $\psi$ -scores exactly. For each possible configuration  $e_{ld}$  for  $d \in \{1, 2, \dots, 2^K\}$ , we denote the posterior probability by  $\pi_{ld}$  and the integrated  $\psi$ -score by  $\bar{\psi}_{ld} = (\bar{\psi}_{ld}^{(1)}, \dots, \bar{\psi}_{ld}^{(K)})$ . According to Stouffer’s meta-analysis method (Stouffer and others, 1949), which is also known as the inverse normal method for combining  $p$ -values (Zaykin, 2011), we define the integrated  $\psi$ -score as

$$\bar{\psi}_{ld}^{(k)} = \begin{cases} \sum_{\{i: e_{ld}^{(i)}=0\}} w_i \psi_l^{(i)} / \sqrt{\sum_{\{i: e_{ld}^{(i)}=0\}} w_i^2}, & \text{if } e_{ld}^{(k)} = 0, \\ \sum_{\{i: e_{ld}^{(i)}=1\}} w_i \psi_l^{(i)} / \sqrt{\sum_{\{i: e_{ld}^{(i)}=1\}} w_i^2}, & \text{if } e_{ld}^{(k)} = 1, \end{cases} \quad (2.5)$$

for  $k = 1, \dots, K$ , where the weight  $w_i$  might account for the size or quality of the samples collected under each condition. In this article, we set  $w_i = 1$  for all  $i = 1, \dots, K$ . Such a weighted average score integrates data information on the edge across all conditions. Then the Bayesian Stouffer integrated  $\psi$ -score, or Bayesian integrated  $\psi$ -score in short, is given by

$$\hat{\psi}_l^{(k)} = \sum_{d=1}^{2^K} \pi_{ld} \bar{\psi}_{ld}^{(k)}, \quad l = 1, 2, \dots, N; \quad k = 1, 2, \dots, K. \quad (2.6)$$

When  $K$  is large, the  $\pi_{ld}$ ’s can be estimated with a short MCMC run, say, by running the Gibbs sampler (Geman and Geman, 1984) for a few hundred iterations. Since the MCMC can be run in parallel for different  $l$ ’s, the computation is not a big burden in this case.

It is interesting to point out that the Bayesian Stouffer integrated  $\psi$ -score  $\hat{\psi}_l^{(k)}$  is different from the conventional Bayesian estimator of  $\mu_l^{(k)}$ . The latter is given by

$$\hat{\psi}_{B,l}^{(k)} = \sum_{d=1}^{2^K} \pi_{ld} \bar{\psi}_{B,ld}^{(k)}, \quad k = 1, 2, \dots, K, \quad (2.7)$$

and

$$\bar{\psi}_{B,ld}^{(k)} = \begin{cases} \sum_{\{i:e_{ld}^{(i)}=0\}} \psi_l^{(i)} / \#\{i : e_{ld}^{(i)} = 0\}, & \text{if } e_{ld}^{(k)} = 0, \\ \sum_{\{i:e_{ld}^{(i)}=1\}} \psi_l^{(i)} / \#\{i : e_{ld}^{(i)} = 1\}, & \text{if } e_{ld}^{(k)} = 1. \end{cases} \quad (2.8)$$

It is easy to see that the Bayesian Stouffer integrated  $\psi$ -score amplifies the Bayesian averaged  $\psi$ -score (2.7) by a factor between 1 and  $\sqrt{K}$ . Such amplification makes the two clusters of edges more separable in the scores and, as pointed out in the Proof of Lemma 3.4 (in the [supplementary material](#) available at *Biostatistics* online), helps to improve the power of the proposed method by reducing the false negative error. Also, we would point out that for each  $l$ , if the edge clustering pattern  $d$  is correct and  $e_{ld}^{(k)} = 0$ , then the Stouffer integrated  $\psi$ -score  $\bar{\psi}_{ld}^{(k)}$  has a constant variance of 1, while the simply averaged  $\psi$ -score  $\bar{\psi}_{B,ld}^{(k)}$  has a varied variance depending on the value of  $\#\{i : e_{ld}^{(i)} = 0\}$ . Therefore, the Bayesian Stouffer integrated  $\psi$ -scores are more comparable than the Bayesian averaged  $\psi$ -scores in edge determination.

**2.2.2. Spatial prior** To enhance our prior knowledge that there exists a common structure for all the networks from which they evolve independently, we let  $e_l$ 's be subject to the prior distribution

$$p(e_l|q) = q^{\sum_{i=1}^K c_l^{*(i)}} (1-q)^{\sum_{i=1}^K (1-c_l^{*(i)})}, \quad (2.9)$$

where  $c_l^{*(i)} = |e_l^{(i)} - e_l^{\text{mod}}|$  indicates the status change of the edge  $l$  at condition  $i$  from  $e_l^{\text{mod}}$ , and  $e_l^{\text{mod}}$  is the mode of  $e_l$  and represents the common status of the edge  $l$  across all networks. With this prior distribution, the posterior distribution  $\pi(e_l|\psi_l)$  can also be expressed in the form of (2.4) but with  $k_1 = \sum_{i=1}^K c_l^{*(i)}$  and  $k_2 = K - k_1$ .

### 2.3. Joint edge detection

To jointly estimate the structure of multiple GGMs based on the Bayesian Stouffer integrated  $\psi$ -scores (2.6), a multiple hypothesis test can be applied. The multiple hypothesis test classifies the integrated  $\psi$ -scores into two classes, presence of edges and absence of edges. In this article, we adopt the empirical Bayesian method developed by [Liang and Zhang \(2008\)](#) for the multiple hypothesis test, which models the integrated  $\psi$ -scores by a mixture distribution  $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ , where  $z$  denotes an integrated  $\psi$ -score;  $\pi_0$  and  $\pi_1$  denote the probabilities of edge absence and edge presence, respectively; and  $f_0(\cdot)$  and  $f_1(\cdot)$  denote the probability density functions of integrated  $\psi$ -scores with edge absence and edge presence, respectively. As in [Liang and Zhang \(2008\)](#), we parameterize  $f_0$  by an exponential power distribution, and parameterize  $f_1$  by a mixture of exponential power distributions. The parameters  $\pi_0$ ,  $\pi_1$  and those contained in  $f_0(\cdot)$  and  $f_1(\cdot)$  are estimated using the stochastic approximation method by minimizing the Kullback–Leibler divergence between the density  $f(z)$  and the empirical one. The threshold values for grouping the integrated  $\psi$ -scores into the classes  $f_0$  and  $f_1$  are determined according to the value of  $\alpha_2$ , a pre-specified false discovery rate level. How to specify the value of  $\alpha_2$  will be discussed in Section 2.4. Note that this multiple hypothesis test method allows for the dependence between test statistics, i.e., integrated  $\psi$ -scores for this problem. Other methods that account for the dependence between test statistics, e.g., the two-stage method by [Benjamini and Yekutieli \(2001\)](#), can also be applied here.

Finally, we would like to point out that the empirical Bayesian method used above produces a  $q$ -value ([Storey, 2002](#)) for each potential edge of the multiple graphs. The  $q$ -value, like the  $p$ -value for the single hypothesis test, provides an uncertainty measure for each potential edge.



#### 2.4. Parameter setting

FHBIA contains two free parameters, i.e.,  $\alpha_1$  and  $\alpha_2$ , which refer to the significance levels of the multiple hypothesis tests conducted in correlation screening and JED, respectively. Following the suggestion of [Liang and others \(2015\)](#), we set  $\alpha_1 = 0.2$  and  $\alpha_2 = 0.05$  as the default values. Otherwise, their values will be stated in the context. In general, a high significance level of correlation screening will lead to slightly large conditioning sets in calculation of  $\psi$ -partial correlation coefficients, which reduces the risk of missing important variables in the conditioning sets. Including a few false variables in the conditioning sets will not hurt much the accuracy of the  $\psi$ -partial correlation coefficients. As shown in [Xu and others \(2019\)](#), the performance of the  $\psi$ -learning algorithm can be quite robust to the choice of  $\alpha_1$ . However, the setting of  $\alpha_2$  is quite free, which determines the sparsity of the resulting graphs. A smaller value of  $\alpha_2$  might be used if sparse graphs are preferred.

In addition to the two free parameters, FHBIA contains four prior-hyperparameters, i.e.,  $a_1, b_1, a_2$ , and  $b_2$ . Since the probability  $q$  usually takes a small value, we set  $(a_1, b_1) = (1, 10)$  for its prior distribution  $\text{Beta}(a_1, b_1)$ . Since the variance of the  $\psi$ -scores is approximately equal to 1 under the null hypothesis that the true partial correlation coefficient is equal to 0, we set  $(a_2, b_2) = (1, 1)$  for its prior distribution  $\text{IG}(a_2, b_2)$ . The same prior hyperparameter settings have been used in all examples of this article.

#### 2.5. Consistency

Under the faithfulness assumption, sparsity assumption, and other regularity conditions for the joint Gaussian distribution, e.g., the dimension  $p = O(\exp(n^\delta))$  is allowed to grow exponentially with the sample size  $n$  for some constant  $0 \leq \delta < 1$  and the largest eigenvalue of the covariance matrix can grow with  $n$  at a restricted rate, [Liang and others \(2015\)](#) showed that the multiple hypothesis test based on the  $\psi$ -scores produces a consistent estimate for the GGM with data observed under single condition. Essentially, [Liang and others \(2015\)](#) showed that the  $\psi$ -scores are separable in probability for the pairs of nodes with edge absence and edge presence.

To accommodate the change from single condition to multiple conditions, we modified the assumptions of [Liang and others \(2015\)](#) and added an assumption about  $K$ . Under the new set of assumptions, we proved that the FHBIA method is consistent.

**THEOREM 2.1** Assume  $(A_1)$ – $(A_6)$  (see supplementary material available at *Biostatistics* online) hold. Then

$$P[\hat{\mathbf{E}}_{\zeta_n}^{(k)} = \tilde{\mathbf{E}}_n^{(k)}, k = 1, 2, \dots, K] \geq 1 - o(1), \quad \text{as } n \rightarrow \infty,$$

where  $\tilde{\mathbf{E}}_n^{(k)} = \{(i, j) : \rho_{ij|V \setminus \{i, j\}}^{(k)} \neq 0, i, j = 1, \dots, p\}$  denotes the true network under condition  $k$ ,  $\hat{\mathbf{E}}_{\zeta_n}^{(k)}$  denotes the FHBIA estimator of  $\tilde{\mathbf{E}}_n^{(k)}$ , and  $\zeta_n$  denotes a threshold value of Bayesian integrated  $\psi$ -scores based on which the edges are determined for all  $K$  graphs.

The proof of the theorem is given in the supplementary material available at *Biostatistics* online. **Theorem 2.1** implies that for all graphs there exists a common threshold with respect to which the Bayesian integrated  $\psi$ -scores are separable in probability for the pairs of nodes with edge presence and edge absence. Here, we would like to highlight three points. First, as indicated by our proof [see inequality (S29) in the Proof of Lemma 3.5 in [supplementary material](#) available at *Biostatistics* online], the data integration step can indeed improve the power of proposed method. Second, following from the inequalities (S29) and (S30) and the condition  $(A_5)$  given in the [supplementary material](#) available at *Biostatistics* online, we can conclude the sign consistency of the estimator  $\hat{\mathbf{E}}_{\zeta_n}^{(k)}$ ; i.e., for any edge of the graph, the sign of



the Bayesian integrated  $\psi$ -score has the same sign as the true partial correlation coefficient when the sample size  $n$  becomes large. Third, the assumption imposed on  $K$ , i.e.,  $K = O(n^{\delta+2d+\epsilon-1})$ , is rather weak, where  $\delta$ ,  $d$ , and  $\epsilon$  are all some positive constants as defined in other assumptions and  $\delta + 2d < 1$  (see the [supplementary material](#) available at *Biostatistics* online). For example, we can choose  $\epsilon = 1 - \delta - 2d$  and thus  $K = O(1)$ . This is consistent with our numerical results; the method can perform very well even with a small value of  $K$ .

### 3. SIMULATION STUDIES

#### 3.1. Scenario with temporal priors

To illustrate the performance of the proposed FHBIA method under the scenario with temporal priors, we consider three types of network structures, namely, autoregressive (AR), scale-free, and hub, which are all allowed to change slightly with the evolution of conditions. For all the types of structures, we fix  $K = 4$  and  $p = 200$ , and varied the sample size  $n = 100$  and  $500$ . We let  $\Omega^{(k)}$  denote the precision matrix at condition  $k$  for  $k = 1, \dots, K$ . At each condition  $k$ , we generated 10 independent datasets of size  $n$  by drawing from the multivariate Gaussian distribution  $N\left(0, (\Omega^{(k)})^{-1}\right)$ .

For the AR network structure, the precision matrix at condition 1 is given by

$$\Omega_{i,j}^{(1)} = \begin{cases} 0.5, & \text{if } |j - i| = 1, i = 2, \dots, (p - 1), \\ 0.25, & \text{if } |j - i| = 2, i = 3, \dots, (p - 2), \\ 1, & \text{if } i = j, i = 1, \dots, p, \\ 0, & \text{otherwise,} \end{cases} \quad (3.10)$$

which represents an AR(2) graphical model. To construct  $\Omega^{(2)}$ , we employed the following random edge deleting–adding procedure: we first randomly removed 5% edges in  $\Omega^{(1)}$  by setting the corresponding nonzero elements to 0, and then added the same number of edges at random by replacing zeros in  $\Omega^{(1)}$  with the values drawn from the uniform distribution defined on  $[-0.1, -0.3] \cup [0.1, 0.3]$ ; to ensure  $\Omega^{(2)}$  to be positive definite, we set the diagonal elements of  $\Omega^{(2)}$  to be the smallest absolute eigenvalue of  $\tilde{\Omega}^{(2)}$  plus a small positive number, where  $\tilde{\Omega}^{(2)}$  is obtained from  $\Omega^{(2)}$  by setting the diagonal elements to zero. In the same procedure, we generated  $\Omega^{(3)}$  conditioned on  $\Omega^{(2)}$  and then generated  $\Omega^{(4)}$  conditioned on  $\Omega^{(3)}$ . We note that similar procedures have been used in [Peterson and others \(2015\)](#) and [Lin and others \(2017\)](#) to generate multiple precision matrices. For the scale-free and hub structures, we first generated the precision matrix  $\Omega^{(1)}$  using the R package “*huge*,” then applied the random edge deleting–adding procedure to generate  $\Omega^{(k)}$ ’s for  $k = 2, 3, 4$  in a sequential manner.

The FHBIA method was first applied to this example. To access the performance of the method, we plot the precision-recall curves in Figure S1 of [supplementary material](#) available at *Biostatistics* online. The same rule applies to other tables and figures included in the [supplementary material](#) available at *Biostatistics* online. The precision and recall are defined by  $\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ ,  $\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ , where TP, FP, and FN denote true positives, false positives, and false negatives, respectively, as defined in Table S1 of [supplementary material](#) available at *Biostatistics* online. To draw the precision-recall curves shown in Figure S1 of [supplementary material](#) available at *Biostatistics* online, we fix the significance level of correlation screening to  $\alpha_1 = 0.2$  and varied the value of  $\alpha_2$ , the significance level of JED. The precision and recall values were calculated by cumulating the TP, FP, FN, and TN values across all  $K$  conditions. In this article, we employ the precision-recall curve instead of the Receiver Operating Characteristic (ROC) curve as the classification problem involved in recovering the network structure is severely imbalanced, which contains a large number of negative cases due to the network sparsity. As pointed out by [Saito and Rehmsmeier \(2015\)](#) and

Davis and Goadrich (2006), the precision-recall curve can be more informative than the ROC curve in the imbalanced classification scenario.

For comparison, the MRF method (Lin and others, 2017), fused graphical Lasso (FGL), and group graphical Lasso (GGL) (Danaher and others, 2014) were applied to this example. The Matlab code of the MRF is available at <https://github.com/linzx06/Spatial-and-Temporal-GGM>. Both FGL and GGL are available in the R package *JGL*. For a thorough comparison, we also applied the original  $\psi$ -learning algorithm to this example, for which the models under each condition were estimated separately. The results are summarized in Figure S1 and Table S2 of [supplementary material](#) available at *Biostatistics* online. The comparison indicates that FHBIA significantly outperforms the existing methods, especially when the sample size is small. When the sample size is large, FHBIA, MRF, FGL, and GGL tend to perform similarly for the scale-free and hub networks. It is not surprising that FHBIA always outperforms the separated  $\psi$ -learning algorithm, which implies the importance of data integration for such high-dimensional problems.

Table S3 of [supplementary material](#) available at *Biostatistics* online reports the CPU time cost by FGL, GGL, MRF, separated  $\psi$ -learning and FHBIA for one dataset of AR(2) structure, where the CPU time was measured on a Linux desktop with Inter Core i7-4790 CPU 3.6Ghz. All computations reported in this article were done on the same computer. The CPU times of these methods for the other two graph structures are about the same. FGL is extremely slow for this example, as it needs to search over a grid of possible values for an optimal setting of  $(\lambda_1, \lambda_2)$ . The grid we used consists of 100 different pairs of  $(\lambda_1, \lambda_2)$ . Moreover, for each pair of  $(\lambda_1, \lambda_2)$ , it needs to solve a generalized fused Lasso problem for which a closed-form solution does not exist when  $K$  is greater than 2. Solving the generalized fused Lasso problem is time consuming and has a computational complexity of  $O(p^2 K \log K)$ . The GGL is better as for which there exists a closed-form solution to the regularized parameter optimization problem under each setting of  $(\lambda_1, \lambda_2)$ , although the optimal setting of  $(\lambda_1, \lambda_2)$  also needs to be searched over a grid of 100 points. The computational complexity of MRF is of  $O(p^3)$  (Lin and others, 2017) while FHBIA is of  $O(p^2 2^K)$ , which can be pretty fast for a small value of  $K$ . The separated  $\psi$ -learning is a little more time consuming than FHBIA because it needs to conduct multiple hypothesis tests under each condition.

### 3.2. Scenario with spatial priors

As in the scenario with temporal priors, we considered three types of network structures: AR(2), scale-free, and hub. For each type of structures, we set  $K = 5$  and  $p = 200$ , and tried two sample sizes  $n = 100$  and  $n = 500$ . For AR(2), we first generated the precision matrix  $\Omega^{(0)}$  according to (2.1). Conditioned on  $\Omega^{(0)}$ , we generated the precision matrices  $\Omega^{(k)}$ ,  $k = 1, 2, \dots, 5$ , independently using the random edge deleting-adding procedure as described in the scenario of temporal priors. For the other two types of structures, we generated the precision matrices  $\Omega^{(0)}$  using the R package *huge*, and then generated  $\Omega^{(k)}$ ,  $k = 1, 2, \dots, 5$  independently using the random edge deleting-adding procedure. Given the precision matrices, we then generated 10 independent datasets of size  $n$  by drawing from the multivariate Gaussian distribution  $N\left(0, (\Omega^{(k)})^{-1}\right)$  for each condition  $k$ .

The FHBIA, MRF, FGL, GGL, separated  $\psi$ -learning and graphical EM (Xie and others, 2016) methods were applied to this example. The graphical EM algorithm was specially designed for jointly estimating multiple dependent Gaussian graphical models under this scenario.

Figure S2 of [supplementary material](#) available at *Biostatistics* online shows the precision-recall curves produced for two datasets by FHBIA, MRF, FGL, GGL, separated  $\psi$ -learning, and graphical EM. Table S4 of [supplementary material](#) available at *Biostatistics* online summarizes the performance of these methods for all simulated datasets of this example. The comparison indicates that FHBIA significantly outperforms all other methods, especially when the sample size is small.

Table S5 of [supplementary material](#) available at *Biostatistics* online reports the CPU time cost by MRF, FGL, GGL, separated  $\psi$ -learning, graphical EM, and FHBIA for one dataset of AR(2) structure. The CPU times for the other two graph structures are about the same. For FGL, this example is even more time consuming than the previous one, although it was run under exactly the same setting for the two examples. One reason is that  $K$  has increased from 4 to 5. For FHBIA, the CPU time is not much increased compared to the previous example.

#### 4. TEDDY DATA ANALYSIS

This section applied the FHBIA method to the mRNA gene expression data collected in the study of TEDDY. In the study, to reduce potential bias and retain study power while reducing the costs by limiting the numbers of samples requiring laboratory analyses, the gene expression data were collected from the nested matched case–control cohort. A subject who developed two primary outcomes, persistent confirmed islet autoimmunity (i.e., the presence of one confirmed autoantibody, GADA65A, IA-2A, or IAA, on two or more consecutive samples) and/or T1D, was defined as a case. The controls are randomly selected among cohort members who have not yet developed the disease at the time a case is diagnosed. For each subject, the gene expression data were collected at multiple time points within 4 years of age. Refer to [Lee and others \(2014\)](#) for the detailed description for the study. Our goal is to integrate all the data to construct one gene network under each distinct condition.

The dataset consists of 21 285 genes and 742 samples collected at multiple time points from a total of 313 subjects. Among the 742 samples, half of them are for the case and half of them are for the control. The dataset also contains some external variables for each patient, which include age (the time of data collected), gender, race, race ethnicity, season of birth, number of older siblings, and country. To simplify the analysis, we first filtered out some non-differentially expressed genes across the case and control conditions. This was done by conducting a paired  $t$ -test for each gene at each time point and then applied the multiple hypothesis test method by [Liang and Zhang \(2008\)](#) to identify the set of genes that are significantly differentially expressed under the two conditions at least at one time point. With this filtering process, 572 genes were selected for further study. Figure S3 of [supplementary material](#) available at *Biostatistics* online shows the histogram of the ages of the samples. Based on this histogram, we selected only the samples fallen into the first nine groups for the further analysis, where each mode of the histogram is treated as a group. The respective group sizes are 29, 40, 49, 43, 32, 27, 27, 23, and 21, which are the same for both the case and control. Since the samples were grouped in ages, the index  $k = 1, 2, \dots, 9$  can be understood as the time of experiments. In grouping the samples, we have ensured that in each group, each sample corresponds to a different patient and thus the samples within the same group can be treated as mutually independent. Since the sample size of each group is small, we set  $\alpha_1 = 0.05$  and  $\alpha_2 = 0.01$ , which are smaller than the default values.

To adjust the effect of external variables, we adopted the method proposed by [Liang and others \(2015\)](#). Let  $W_1^{(k)}, \dots, W_q^{(k)}$  denote the external variables observed at condition  $k$ . To adjust for their effects, we can replace the empirical correlation coefficient used in the  $\psi$ -score calculation step by the  $p$ -value obtained in testing the hypotheses  $H_0 : \beta_{q+1} = 0 \leftrightarrow H_1 : \beta_{q+1} \neq 0$  for the regression

$$X_i^{(k)} = \beta_0 + \beta_1 W_1^{(k)} + \dots + \beta_q W_q^{(k)} + \beta_{q+1} X_j^{(k)} + \epsilon, \quad (4.11)$$

where  $X_i^{(k)}$  denotes the expression value of gene  $i$  measured at condition  $k$ , and  $\epsilon$  denotes a vector of Gaussian random errors. Similarly, we can replace the  $\psi$ -partial correlation coefficient calculated in the  $\psi$ -score calculation step by the  $p$ -value obtained in testing the hypotheses  $H_0 : \beta_{q+1} = 0 \leftrightarrow H_1 : \beta_{q+1} \neq 0$

for the regression

$$X_i^{(k)} = \beta_0 + \beta_1 W_1^{(k)} + \cdots + \beta_q W_q^{(k)} + \beta_{q+1} X_j^{(k)} + \sum_{s \in S_{ij}^{(k)}} \gamma_s X_s^{(k)} + \epsilon, \quad (4.12)$$

where  $S_{ij}^{(k)}$  is the separator of  $X^{(i)}$  and  $X^{(j)}$  under condition  $k$ . With the  $p$ -values, we can define the adjusted  $\psi$ -score as  $\psi_l^{(k)} = \Phi^{-1}(1 - p_l^{(k)})$ , where  $p_l^{(k)}$  is the  $p$ -value obtained from equation (4.12) for edge  $l$  at condition  $k$ .

For this dataset, the effect of all available demographical variables, including age (the time of data collection), gender, race, race ethnicity, season of birth, number of older siblings, and country, have been adjusted. With the adjusted  $\psi$ -scores, the FHBIA method is ready to be applied to construct the gene networks. Given the complexity of the dataset, which contains case and control groups and multiple time points for each group, we calculated the integrated  $\psi$ -scores in two steps. First, we integrated the  $\psi$ -scores across nine time points under the case and control, separately. Then, for each time point, we integrated the  $\psi$ -scores across the case and control conditions. In this way, all information of the data collected under the 18 conditions were integrated together. Figure 1 shows a schematic diagram for this two-step procedure. Finally, we applied the multiple hypothesis test to the Bayesian integrated  $\psi$ -scores to determine the structure of the gene networks under the 18 conditions. The total CPU time cost by FHBIA was 19.2 h, which is pretty long as  $K = 9$  is large. For a larger value of  $K$ , we might resort to MCMC for estimating the posterior probabilities  $\pi(e_l | \psi_l)$ 's.

Figure 4 shows the networks constructed by FHBIA for the case samples at nine time points. The networks have identified quite a few hub genes, which refer to the genes with high connectivity. Table 1 shows the top 5 hub genes identified at each time point for the case samples. The lists of hub genes are pretty stable. For example, RPS26P11 and RPS26 consistently appear as top 2 genes at all time points, the gene ADAM10 appeared at five out of nine time points, and quite a few genes appeared twice or more times, such as PRF1, POGZ, BCL11B, GGNBP2, and TMEM159. Note that RPS26P11 is a pseudo-gene, which represents a segment of the gene RPS26.

Table 1 includes 11 different genes in total. Among the 11 genes, 9 genes have been verified in the literature to be T1D associated genes. For example, [Schadt and others \(2008\)](#) reported that RPS26 is a T1D causal gene, and [Ma and Hart \(2013\)](#) reported that the gene O-GlcNAc transferase (OGT) is directly linked to many metabolic diseases including diabetes. Other than identifying some verified T1D associated genes, we have also some new findings such as gene PRF1. [Orilieri and others \(2008\)](#) claimed that PRF1 variations are susceptibility factors for T1D development. In Table 1, PRF1 appeared as a hub gene twice, which suggests that the connection between PRF1 and T1D might be worth to be further explored. Moreover, we also identifies some connection changes in the networks. As showed in Figure 4, the new appearing and disappearing connections are marked in different colors at each time point, which identify some evolvment patterns of the network.

For comparison, the GGL method was also applied to this example, for which the regularization parameters were chosen according to the minimum AIC criterion. The total CPU time cost by the method was 20.2 h. FGL was not applied to this example, as it would take extremely long CPU time. Figure S4 available at *Biostatistics* online shows the networks constructed by GGL for the case samples at all nine time points. Table S6 available at *Biostatistics* online shows the top 5 hub genes identified by GGL at each time point for the case samples. The lists of hub genes are pretty stable, which consists of seven different genes only. Among the seven genes, only three genes RPS26, OGT, and JMJD1C have been verified in the literature as T1D-associated genes. Moreover, as showed in Figure S4 available at *Biostatistics* online, the hub genes in networks are almost identical at each time point. In summary, FHBIA tends to outperform GGL for this real data example which can identify more hub genes which are associated with T1D.

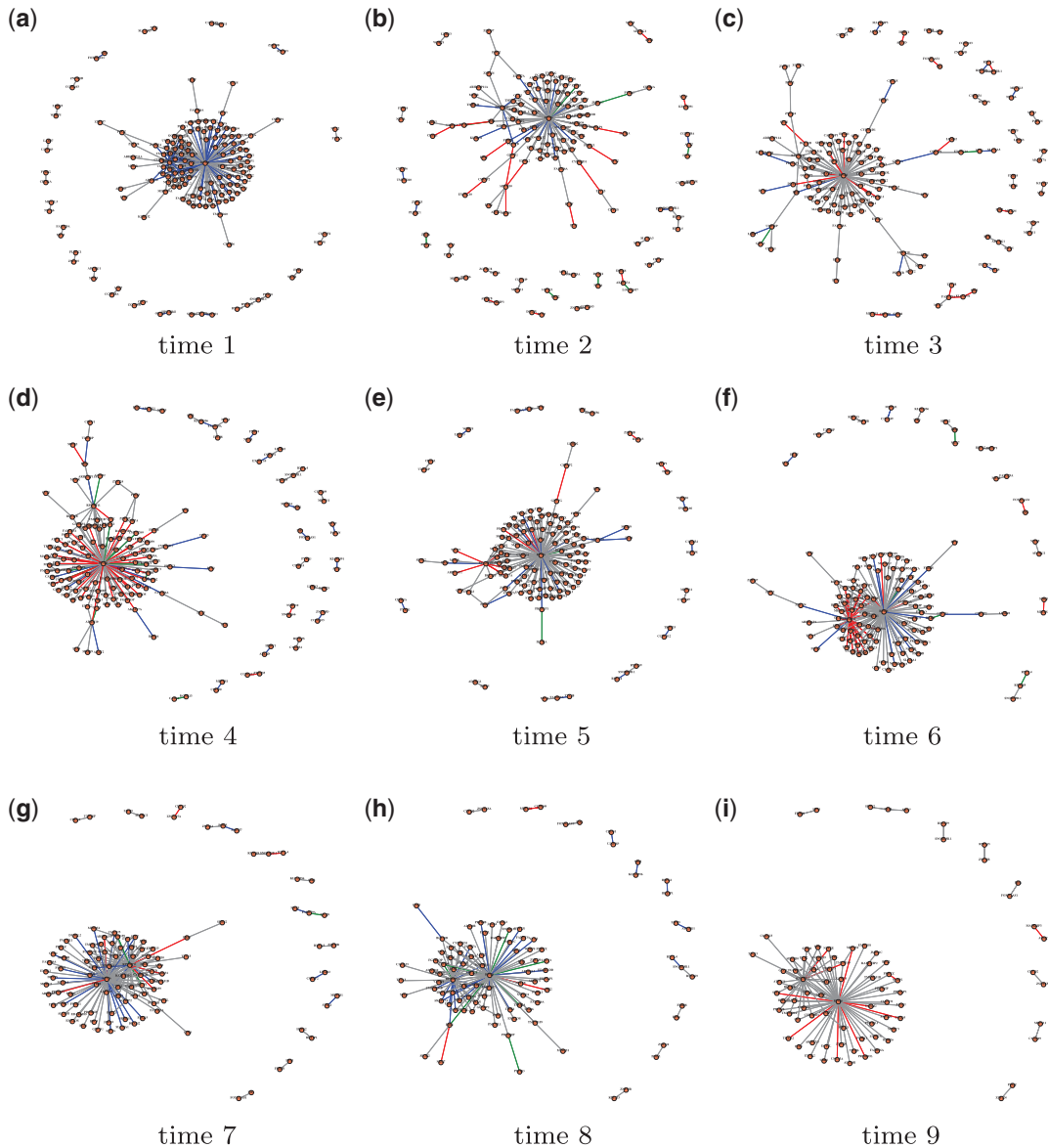


Fig. 4. Gene networks produced by FHBIA for the case TEEDY samples at nine time points. The red edge lines denote new connections appearing in the current network compared with the network at the previous time point; the blue edge lines denote the disappearing connections in the network of the next time point; the green edge lines denote the lines that are both newly appearing connections and disappearing connections; the gray edge lines denote the unchanged connections in the current network and network at the previous time point. (a) Time 1, (b) time 2, (c) time 3, (d) time 4, (e) time 5, (f) time 6, (g) time 7, (h) time 8, and (i) time 9.

From the perspective of data analysis, one might also be interested in estimating the gene networks constructed from the controls, as well as the differences between the networks from the cases and controls. For comparing the networks from the cases and controls, we can adopt the method described in Section 6 of *Liang and others (2015)*. However, since the method by *Liang and others (2015)* requires that the two

Table 1. Top 5 hub genes identified by FHBIA for the case TEDDY samples at nine time points: “Links” denotes the number of links of the gene to other genes,  $k$  is the index of time points, \* indicates that there exist other genes which has the same number of links with this genes,  $\Delta$  indicates that this gene has been verified as a T1D-related gene in the literature

		Case group					
		Gene	Links	Gene	Links	Gene	Links
$k = 1$	$\Delta$ RPS26	104	$\Delta$ RPS26	68	$\Delta$ RPS26	64	
	$\Delta$ RPS26P11	40	$\Delta$ RPS26P11	15	$\Delta$ RPS26P11	12	
	$\Delta$ ADAM10	4	$k = 2$ $\Delta$ ADAM10	5	$k = 3$ $\Delta$ ADAM10	5	
	$\Delta$ POGZ	3	$\Delta$ PRF1	4	U2SURP	4	
	$\Delta$ TMEM159*	3	$\Delta$ POGZ	3	$\Delta$ BCL11B*	3	
$k = 4$	$\Delta$ RPS26	99	$\Delta$ RPS26	91	$\Delta$ RPS26	86	
	$\Delta$ RPS26P11	14	$\Delta$ RPS26P11	18	$\Delta$ RPS26P11	42	
	$\Delta$ ADAM10	6	$k = 5$ $\Delta$ ADAM10	4	$k = 6$ $\Delta$ BCL11B	3	
	$\Delta$ BCL11B	3	$\Delta$ BCL11B	3	GNPTG	3	
	$\Delta$ POGZ*	3	$\Delta$ POGZ*	3	$\Delta$ GGNBP2	3	
$k = 7$	$\Delta$ RPS26	78	$\Delta$ RPS26	70	$\Delta$ RPS26	61	
	$\Delta$ RPS26P11	46	$\Delta$ RPS26P11	39	$\Delta$ RPS26P11	30	
	$\Delta$ BCL11B	3	$k = 8$ $\Delta$ PRF1	4	$k = 9$ $\Delta$ TMEM159	3	
	$\Delta$ TMEM159	3	$\Delta$ BCL11B	3	$\Delta$ GGNBP2	3	
	$\Delta$ GGNBP2	3	$\Delta$ GGNBP2	3	$\Delta$ OGT*	2	

networks under comparison are independent, the sample information from the cases and controls should not be integrated in this case. We left this work to the future.

## 5. DISCUSSION

We have proposed the FHBIA method for jointly estimating multiple GGMs under distinct conditions and applied it to TEDDY data. The FHBIA method consists of a few important steps, which is to first summarize the graph structure information contained in the data using the  $\psi$ -learning algorithm, then integrate information via a meta-analysis procedure under the Bayesian framework, and finally determine the structures of multiple graphs via a multiple hypothesis test. Compared to the existing methods, FHBIA has a few significant advantages. First, FHBIA includes a meta-analysis procedure to explicitly integrate information across distinct conditions. In contrast, the existing methods integrate information through prior distributions or penalty function, which is often less efficient. Second, FHBIA can be run very fast, especially when  $K$  is small. The overall computational complexity of FHBIA is  $O(p^2 2^K)$ , where the factor  $2^K$  is the total number of possible configurations of an edge across all  $K$  conditions. When  $K$  is large, we need to resort to MCMC for an efficient estimation of the posterior probabilities  $\pi(e_l | \psi_l)$ 's for  $l = 1, 2, \dots, p(p-1)/2$ . Since  $\pi(e_l | \psi_l)$ 's can be estimated for each  $l \in \{1, 2, \dots, p(p-1)/2\}$  independently, this step can be done in parallel. In addition, we note that the correlation coefficients and  $\psi$ -scores can also be calculated in parallel. Hence, the whole method can be executed very fast on a parallel architecture. Moreover, instead of working on the original data, the Bayesian integration step chooses to work on the edge-wise  $\psi$ -scores, which avoids to invert high-dimensional covariance matrices and thus can be very fast. Note that, in calculation of  $\psi$ -scores, the  $\psi$ -learning algorithm also successfully avoids to invert high-dimensional covariance matrices through correlation screening. Third, the empirical Bayesian



method that FHBIA employed for multiple hypothesis tests produces a  $q$ -value (Storey, 2002) for each potential edge of the multiple graphs. The  $q$ -value provides an uncertainty measure for each potential edge. This has been beyond the ability of many of the existing methods, especially when  $p$  is large.

The FHBIA method has a very flexible framework, which can be easily extended to joint estimation of multiple mixed graphical models. For example, consider the scenario that the data consist of only Gaussian and multinomial random variables, for which the joint distribution is well defined (Lee and Hastie, 2015). For such mixed data, the  $\psi$ -learning algorithm can be performed under the framework of generalized linear models; that is, we can replace the correlation coefficients and  $\psi$ -partial correlation coefficients used in the algorithm by the corresponding  $p$ -values obtained in the marginal variable screening tests (Fan and Song, 2010) and conditional independence tests. Then, we can replace the  $\psi$ -scores by the  $Z$ -scores corresponding to the  $p$ -values of the conditional independence tests. For other types of continuous random variables, we can apply the nonparanormal transformation (Liu and others, 2009) to Gaussianize them prior to the application of the FHBIA method. For certain types of discrete data, e.g., next generation sequencing data, we can apply the transformations developed in (Jia and others, 2017) to continuize and Gaussianize them prior to the application of the FHBIA method.

Finally, we would like to mention that the sparsity assumption imposed on the networks does not limit the applications of the FHBIA method. Sparsity is a just trick adopted by people for making statistical inference where there is no enough amount of data available, e.g., in dealing with small- $n$ -large- $p$  problems. In this article, the sparsity assumption A4 (in the [supplementary material](#) available at *Biostatistics* online) is given in the form of the sample size  $n$ , which leads to a bounded neighborhood size of  $O(n/\log(n))$  for each node, see Lemma 3.2 and the  $\psi$ -learning algorithm presented in the [supplementary material](#) available at *Biostatistics* online. Therefore, when the sample size  $n$  is large enough, e.g., when using the UK biobank data which consists of over 500 K samples to construct GRNs, FHBIA can be applied to learn dense genetic networks (Boyle and others, 2017). For UK biobank data, we have  $500\,000/\log(500\,000) \approx 38\,103$ , which is much larger than the number of genes ( $\approx 20\,000$ ) we usually considered and thus neighborhood truncation in the correlation screening step of  $\psi$ -learning will not be triggered. When the data size is not large enough, the FHBIA will identify only the strongest connections in terms of partial correlation coefficients. This property is inherited from the  $\psi$ -learning algorithm.

## 6. SOFTWARE

The software accompanied with this article is available as a module called *JGGM* in the R package *equSA* at <https://cran.r-project.org/web/packages/equSA/index.html>.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

## ACKNOWLEDGMENTS

The authors thank the editor, associate editor, two referees, and Dr. George Tseng for their constructive comments which have led to significant improvement of this article. Members of the TEDDY Study Group are listed in the [Supplementary File](#). *Conflict of Interest*: None declared.

## FUNDING

Leona M. and Harry B. Helmsley Charitable Trust (2015PG-T1D050); F.L.'s research was support in part by the grants USF-ITN-15-11-MH, DMS-1612924, DMS/NIH R01-GM117597, and NIH R01-GM126089. The TEDDY Study is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4



DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and Contract No. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082), in part.

## REFERENCES

- BENJAMINI, Y. AND YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.
- BOYLE, E. A., LI, Y. I. AND PRITCHARD, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186.
- CHUN, H., ZHANG, X. AND ZHAO, H. (2015). Gene regulation network inference with joint sparse Gaussian graphical models. *Journal of Computational and Graphical Statistics* **24**, 954–974.
- 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**, 373–397.
- DAVIS, J. AND GOADRICH, M. (2006). The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA. New York, NY, USA: ACM, pp. 233–240.
- FAN, J. AND SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38**, 3567–3604.
- GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- GUO, J., LEVINA, E., MICHAILIDIS, G. AND ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.
- JIA, B., XU, S., XIAO, G., LAMBA, V. AND LIANG, F. (2017). Learning gene regulatory networks from next generation sequencing data. *Biometrics* **73**, 1221–1230.
- LEE, H. S., BURKHARDT, B. R., MCLEOD, W., SMITH, S., EBERHARD, C., LYNCH, K., HADLEY, D., REWERS, M., SIMELL, O., SHE, J. X. and others (2014). Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes/Metabolism Research and Reviews* **30**, 424–434.
- LEE, J. AND HASTIE, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics* **24**, 230–253.
- LIANG, F., SONG, Q. AND QIU, P. (2015). An equivalent measure of partial correlation coefficients for high dimensional Gaussian graphical models. *Journal of the American Statistical Association* **110**, 1248–1265.
- LIANG, F. AND ZHANG, J. (2008). Estimating the false discovery rate using the stochastic approximation algorithm. *Biometrika* **95**, 961–977.
- LIN, Z., WANG, T., YANG, C. AND ZHAO, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* **73**, 769–779.
- LIU, H., LAFFERTY, J. AND WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**, 2295–2328.
- MA, J. AND HART, G. W. (2013). Protein O-GlcNAcylation in diabetes and diabetic complications. *Expert Review of Proteomics* **10**, 365–380.

- MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 1436–1462.
- ORILIERI, E., CAPPELLANO, G., CLEMENTI, R., COMETA, A., FERRETTI, M., CERUTTI, E., CADARIO, F., MARTINETTI, M., LARIZZA, D., CALCATERRA, V. and others (2008). Variations of the perforin gene in patients with type 1 diabetes. *Diabetes* **57**, 1078–1083.
- PETERSON, C., STINGO, F. C. AND VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* **110**, 159–174.
- QIU, H., HAN, F., LIU, H. AND CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 487–504.
- SAITO, T. AND REHMSMEIER, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432.
- SCHADT, E. E., MOLONY, C., CHUDIN, E., HAO, K., YANG, X., LUM, P. Y., KASARSKIS, A., ZHANG, B., WANG, S., SUVER, C. and others (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6**, e107.
- STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479–498.
- STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A., and WILLIAMS, R. M., Jr. and others (1949). *The American Soldier: Adjustment during Army Life*, (Studies in social psychology in World War II) Volume 1. Oxford, England: Princeton University Press.
- XIE, Y., LIU, Y. AND VALDAR, W. (2016). Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika* **103**, 493–511.
- XU, S., JIA, B. AND LIANG, F. (2019). Learning moral graphs in construction of high-dimensional Bayesian networks for mixed data. *Neural Computation* **31**, 1183–1214.
- ZAYKIN, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology* **24**, 1836–1841.
- ZHOU, S., LAFFERTY, J. AND WASSERMAN, L. (2010). Time varying undirected graphs. *Machine Learning* **80**, 295–319.

[Received September 26, 2018; revised June 1, 2019; accepted for publication June 23, 2019]