*The Annals of Applied Statistics*
2017, Vol. 11, No. 1, 161–184
DOI: 10.1214/16-AOAS998


# A STATISTICAL FRAMEWORK FOR DATA INTEGRATION THROUGH GRAPHICAL MODELS WITH APPLICATION TO CANCER GENOMICS

BY YUPING ZHANG[1], ZHENGQING OUYANG[2] AND HONGYU ZHAO[3]

*University of Connecticut, The Jackson Laboratory for Genomic Medicine and Yale University*

Recent advances in high-throughput biotechnologies have generated various types of genetic, genomic, epigenetic, transcriptomic and proteomic data across different biological conditions. It is likely that integrating data from diverse experiments may lead to a more unified and global view of biological systems and complex diseases. We present a coherent statistical framework for integrating various types of data from distinct but related biological conditions through graphical models. Specifically, our statistical framework is designed for modeling multiple networks with shared regulatory mechanisms from heterogeneous high-dimensional datasets. The performance of our approach is illustrated through simulations and its applications to cancer genomics.

**1. Introduction.** Recent advances in high-throughput technologies have generated unprecedented types and amounts of data for biomedical research. Examples include genome-wide characterizations of DNA variations (e.g., genotyping arrays, whole exome or genome sequencing), gene expression variations (e.g., gene expression microarrays, RNA sequencing), epigenetic variations and protein expression variations. Each data type, for example, genomic, transcriptomic or proteomic data, provides a comprehensive, but one-layer, view of the biological system being studied. Integrating data of diverse types is likely to lead to a more unified and global view. Thus, increasing research attention is being paid to the integrative analysis and modeling of various types of biomedical data. For instance, Varambally et al. (2005) reported the signatures of metastatic progression through integrative genomic and proteomic analysis of prostate cancer. Ouyang, Zhou and Wong (2009) proposed a predictive model to integrate ChIP-Seq and RNA-Seq to capture cooperation among regulators. Chen, Slack and Zhao (2013) developed a statistical framework for joint analysis of expression profiles of microRNA and

Received February 2016; revised September 2016.

[1]Supported in part by the InCHIP Faculty Affiliate Seed Grant at UConn, Faculty Research Excellence Program Award at UConn and the CICATS PreK Career Development Award at UConn.

[2]Supported in part by the Research Starter Grant in Informatics from PhRMA Foundation.

[3]Supported in part by National Science Foundation Grant DMS-11-06738, and National Institutes of Health (NIH) Grants R01 GM59507 and P01 CA154295.

*Key words and phrases.* Cancer genomics, data integration, graphical models.

messenger RNAs from multiple cancers. Troyanskaya et al. (2003) proposed a Bayesian framework for combining heterogeneous data sources for gene function prediction in Saccharomyces cerevisiae. Myers and Troyanskaya (2007) developed a network prediction approach to leveraging biological context information, and applied it to Saccharomyces cerevisiae. Myers et al. (2005) proposed a Bayesian approach to identifying biological networks from diverse functional genomic data. Myers et al. (2006) evaluated several evaluation methods, and suggested a new approach to evaluation in functional genomics. Shen and Tseng (2010) developed a new meta-analysis approach to pathway enrichment analysis when combining multiple genomic studies. For more literature review, see Ge, Walhout and Vidal (2003), Hawkins, Hon and Ren (2010), Hecker et al. (2009), Joyce and Palsson (2006), Ritchie et al. (2015).

In this paper, we focus on the problem of discovering regulatory relationships among heterogeneous genomic variables from biological conditions with potentially shared regulation mechanisms. In this scenario, genomic variables can be genomic variants (for instance, mutations and copy number alterations), epigenetic states (for instance, methylation status) and gene expression profiles. Biological conditions can be different tissue types or different cancer types, etc. The heterogeneous genomic variables can be binary, categorical or continuous. Different biological systems have both shared regulations and tissue or disease specific regulations. Thus, we need a statistical method to jointly learn conditional independence among a set of discrete or continuous variables across a set of distinct but related conditions. Conditional independence among variables can be represented by a graphical model in which nodes represent variables and the absence of an edge between two variables implies conditional independence.

In recent years, many efforts have been devoted to estimating undirected graphical models, especially in the high-dimensional setting under the assumption that the underlying graph is sparse. In most of the published work, the nodes in the graphical models represent either continuous or discrete variables, but not both. In the case of continuous variables, much interest has been focused on estimating Gaussian graphical models of the relationships among a set of random variables with a joint multivariate normal distribution, where zero entries in the precision (or concentration) matrix correspond to conditional independence. Meinshausen and Bühlmann (2006) proposed to estimate the precision matrix via a marginal penalized regression approach. Peng, Zhou and Zhu (2009) extended this approach to estimate partial correlations of Gaussian random variables by joint sparse regression models. Instead of performing regressions, Yuan and Lin (2006), Friedman, Hastie and Tibshirani (2008) and others took a penalized log-likelihood approach. This approach has been extended by Guo et al. (2011) and Danaher, Wang and Witten (2013) to infer multiple Gaussian graphical models based on data collected from distinct but related conditions such as different cancer types. Yin and Li (2011) and Li, Chun and Zhao (2012) considered external effects on the inferred edges through modeling conditional Gaussian graphical models. Chun et al. (2013)

proposed joint conditional Gaussian graphical models with multiple sources of genomic data. In the case of discrete variables, the Ising model can be used to model conditional independence. Höfling and Tibshirani (2009) proposed a pseudolikelihood approach to estimating the sparse binary pairwise Markov networks. Ravikumar, Wainwright and Lafferty (2010) and Guo et al. (2010) formulated the model selection methods for high-dimensional Ising models under a penalized logistic regression framework. In the case that both discrete and continuous variables are considered, Lauritzen (1996) proposed a mixed graphical model in the low-dimensional setting. Recently, several methods have been proposed to estimate a mixed graphical model in the high-dimensional setting. Lee and Hastie (2012) proposed a pairwise graphical model over continuous and discrete variables using a group lasso penalty. Cheng, Levina and Zhu (2013) provided an approach that substitutes the $l_1$ penalty for the group lasso penalty to reduce computation. Fellinghauer et al. (2013) took a random forests approach to mixed variables. Chen, Witten and Shojaie (2015) and Yang et al. (2013) investigated the pairwise graphical model in which the conditional distribution of the nodes belong to an exponential family.

However, in the scenario of multiple networks with mixed types of measurements, for instance, multiple cancer types with copy number variations and mutation measurements, the methods mentioned above are not suitable to be applied directly to gain biologically interpretable results. For instance, if we simply treat biological conditions (cancer types in the example) as categorical variables with equal roles as mutations, we may end up with a network modeling interactions among cancer types. These interactions are not as biologically meaningful as the interactions among mutations and/or copy number variations. Moreover, if we ignore the similarities among biological conditions and estimate the networks separately, we may get less accurate networks. Thus, there is a need to treat biological conditions and genomic measurements differently, and to discover multiple related mixed graphical models in the high-dimensional setting to represent distinct but related relationships under different conditions. We will use cancer genomic data as a motivating example to illustrate our method.

Cancers are complex diseases involving many different mechanisms. High-throughput technologies applied to human cancers have generated large genomic datasets, such as The Cancer Genome Atlas (TCGA) [Tomczak, Czerwińska and Wiznerowicz (2015)]. TCGA provides molecular landscapes of thousands of human cancers at multiple layers, including mutations and copy number alterations. It facilitates the study of regulatory mechanisms underlying various cancers. For example, Ciriello et al. (2013) identified distinct oncogenic processes as well as unexpected similarities among tumors originating from different tissues. However, the molecular regulatory networks underlying cancers are still largely unknown, impeding our understanding of cancer classifications and patient stratification, an important issue in precision medicine.

In this paper, we consider statistical learning of multiple graphical models that consist of both continuous and discrete variables, and develop a method named as Data Integration through Graphical models (DIG). We formally introduce our model in Section 2, and propose appropriate penalty schemes in Section 3 to handle high-dimensional data. Then the problem of estimating multiple mixed graphical model is formulated into an optimization problem. We propose an algorithm for parameter estimation in Section 4 and tuning parameter selection in Section 5. We then illustrate our method through simulations in Section 6 and real application to cancer genomic data in Section 7. We conclude our paper with discussion in Section 8.

**2. Model.** We assume that there are a total of $K$ groups where we have observations consisting of both continuous and discrete variables from each group. Let $(\mathbf{x}_{p\times 1}, \mathbf{y}_{q\times 1})_k$ denote a mixed (i.e., having both continuous and discrete variables) random vector, where $k$ is the group label (such as tissue or disease), $\mathbf{x}_{p\times 1}$ is a $p$-dimensional vector of discrete variables, and $\mathbf{y}_{q\times 1}$ is a $q$-dimensional vector of continuous variables. We assume that the density function $f(\mathbf{x}, \mathbf{y})$ has the following form proposed by Lauritzen (1996), with $k$ omitted for simplicity:

$$f(\mathbf{x}, \mathbf{y}) = \exp\left(g_{\mathbf{x}} + \mathbf{h}_{\mathbf{x}}^{\mathsf{T}}\mathbf{y} - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\boldsymbol{\Omega}_{\mathbf{x}}\mathbf{y}\right), \tag{2.1}$$

where $g_{\mathbf{x}}$ is a real-valued function of $\mathbf{x}$, $\mathbf{h}_{\mathbf{x}}$ is a $q$-vector-valued function of $\mathbf{x}$ taking discrete values, and $\boldsymbol{\Omega}_{\mathbf{x}}$ is a $q \times q$ positive definite symmetric matrix of $\mathbf{x}$.

One can note that equation (2.1) can be rewritten as

$$f(\mathbf{x}, \mathbf{y}) = \exp\left(g_{\mathbf{x}} + \frac{1}{2}\mathbf{h}_{\mathbf{x}}^{\mathsf{T}}\boldsymbol{\Omega}_{\mathbf{x}}^{-1}\mathbf{h}_{\mathbf{x}} - \frac{1}{2}(\mathbf{y} - \boldsymbol{\Omega}_{\mathbf{x}}^{-1}\mathbf{h}_{\mathbf{x}})^{\mathsf{T}}\boldsymbol{\Omega}_{\mathbf{x}}(\mathbf{y} - \boldsymbol{\Omega}_{\mathbf{x}}^{-1}\mathbf{h}_{\mathbf{x}})\right).$$

Thus, the density defined in equation (2.1) implies a conditional Gaussian distribution of $\mathbf{y}|\mathbf{x}$ with mean $\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\Omega}_{\boldsymbol{x}}^{-1}\mathbf{h}_{\mathbf{x}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Omega}_{\mathbf{x}}^{-1}$. The marginal distribution of the discrete variables $\mathbf{x}$ has the following form:

$$P(\mathbf{x}) = (2\pi)^{q/2}\big(\det(\boldsymbol{\Omega}_{\boldsymbol{x}})\big)^{-1/2}\exp\left(g_{\mathbf{x}} + \frac{1}{2}\mathbf{h}_{\mathbf{x}}^{\mathsf{T}}\boldsymbol{\Omega}_{\mathbf{x}}^{-1}\mathbf{h}_{\mathbf{x}}\right).$$

We further simplify equation (2.1) by ignoring all interaction terms between discrete variables of order higher than two, and assuming that discrete variables affect continuous variables in a linear form, that is, the conditional covariance matrix and the canonical mean vector of the continuous (Gaussian) variables is modeled as a linear function of the discrete variables. Therefore, we have the following specifi-

cations for the functional forms of $g_{\mathbf{x}}$, $\mathbf{h}_{\mathbf{x}}$ and $\mathbf{\Omega}_{\mathbf{x}}$:

$$g_{\mathbf{x}} = \lambda_0 + \sum_{j=1}^{p} \lambda_j(x_j) + \sum_{j \neq m} \lambda_{jm}(x_j, x_m),$$

$$\mathbf{h}_{\mathbf{x}} = \boldsymbol{\eta}_0 + \sum_{j=1}^{p} \boldsymbol{\eta}_j(x_j),$$

$$\mathbf{\Omega}_{\mathbf{x}} = \mathbf{\Phi}_0 + \sum_{j=1}^{p} \mathbf{\Phi}_j(x_j),$$

where $\lambda_0$ is the normalizing constant,

$$\lambda_0 = -\frac{q}{2}\ln(2\pi) - \ln\left\{\sum_{\mathbf{x}} \det(\mathbf{\Omega}_{\mathbf{x}})^{-1/2} \exp\left(\sum_{j} \lambda_j(x_j) + \sum_{j \neq m} \lambda_{jm}(x_j, x_m) + \frac{1}{2}\mathbf{h}_{\mathbf{x}}^{\mathsf{T}} \mathbf{\Omega}_{\mathbf{x}}^{-1} \mathbf{h}_{\mathbf{x}}\right)\right\};$$

each $x_j$ takes integer values 1 to $L_j$; $\lambda_j(\cdot)$ is a discrete function taking on $L_j$ possible values; $\lambda_{jm}(\cdot,\cdot)$ is a bi-variate function with $L_j \times L_m$ possible values, and $\lambda_{jm}(x_j, x_m) = \lambda_{mj}(x_m, x_j)$; $\boldsymbol{\eta}_0$ is a $q$-dimensional vector; $\boldsymbol{\eta}_j(\cdot)$ is a $q$-dimensional function with $L_j$ possible values for each dimension; $\mathbf{\Phi}_0$ is a $q \times q$ matrix; and $\mathbf{\Phi}_j(\cdot)$ is a $q \times q$ matrix with each element having $L_j$ possible values, which $\{\mathrm{diag}(\Phi_j)\}_{j=1}^{p} = \{\Phi_{jrr} : j = 1, \ldots, p, r = 1, \ldots, q\}$ are all 0. For identifiability, we set $\lambda_j(1) = 0$, $\eta_{jr}(1) = 0$, $\Phi_{jrs}(1) = 0$ and $\lambda_{jm}(1, \cdot) = \lambda_{jm}(\cdot, 1) = 0$, for $j \in \{1, \ldots, p\}$, $m \in \{1, \ldots, p\}$, $r \in \{1, \ldots, q\}$ and $s \in \{1, \ldots, q\}$.

Now, the model is parametrized by $\lambda_j, \lambda_{jm}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_j, \mathbf{\Phi}_0$ and $\mathbf{\Phi}_j$, where $j \in \{1, \ldots, p\}, m \in \{1, \ldots, p\}$. For simplicity, we use $\mathbf{\Theta}$ to denote the collection of the parameters mentioned above. Among these parameters, $\lambda_j$ and $\boldsymbol{\eta}_0$ are nuisance parameters, which refer to discrete and continuous node potentials, respectively. The rest are responsible for edge potentials, which are the parameters of interest. Specifically, $\Phi_{0rs} + \sum_{j=1}^{p} \Phi_{jrs}(x_j)$ is the continuous-continuous edge potential; $\lambda_{jm}$ is the discrete–discrete edge potential, which is a bivariate function taking on $L_j \times L_m$ values; $\eta_{jr}(x_j)$ is the continuous–discrete edge potential, which takes $L_j$ values.

The model covers the two special situations when there are only discrete or continuous variables naturally. In the case of only discrete variables, the mixed model reduces to a discrete Markov random field,

$$p(\mathbf{x}) \propto \exp\left\{\sum_{j=1}^{p} \lambda_j(x_j) + \sum_{j=1}^{p} \sum_{m=1}^{p} \lambda_{jm}(x_j, x_m)\right\};$$

while in the case of only continuous variables, the mixed model reduces to a multivariate Gaussian graphical model,

$$p(\mathbf{y}) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\Phi_0}^{-1}\boldsymbol{\eta_0})^{\mathsf{T}}\boldsymbol{\Phi_0}(\mathbf{y}-\boldsymbol{\Phi_0}^{-1}\boldsymbol{\eta_0})\right\}.$$

For a graphical model, the conditional distributions are important because they characterize the conditional dependence among the variables. The conditional distributions of the proposed model are as follows:

1. The conditional distribution of $x_j$ given the rest is multinomial,

$$p(x_j|\mathbf{x}_{\setminus j},\mathbf{y};\boldsymbol{\Theta}) = \frac{\exp\{\lambda_j(x_j)+\sum_{m=1}^{p}\lambda_{jm}(x_j,x_m)+\sum_{r=1}^{q}\eta_{jr}(x_j)y_r-\frac{1}{2}\sum_{r=1}^{q}\sum_{s=1}^{q}\Phi_{jrs}(x_j)y_ry_s\}}{\sum_{l=1}^{L_j}\exp\{\lambda_j(l)+\sum_{m=1}^{p}\lambda_{jm}(l,x_m)+\sum_{r=1}^{q}\eta_{jr}(l)y_r-\frac{1}{2}\sum_{r=1}^{q}\sum_{s=1}^{q}\Phi_{jrs}(l)y_ry_s\}}.$$

2. The conditional distribution of $y_r$ given the rest is Gaussian,

$$p(y_r|\mathbf{y}_{\setminus r},\mathbf{x};\boldsymbol{\Theta}) = \sqrt{\frac{\Omega_{xrr}}{2\pi}}\exp\left\{-\frac{\Omega_{xrr}}{2}\left[\frac{\eta_{0r}+\sum_{j=1}^{p}\eta_{jr}(x_j)-\sum_{s\neq r}\Omega_{xrs}y_s}{\Omega_{xrr}}-y_r\right]^2\right\},$$

where $\Omega_{xrs}=\Phi_{0rs}+\sum_{j=1}^{p}\Phi_{jrs}(x_j)$. This implies a regression model as

$$y_r = \frac{1}{\Omega_{xrr}}\left(\eta_{0r}+\sum_{j=1}^{p}\eta_{jr}(x_j) - \sum_{s\neq r}\left(\Phi_{0rs}+\sum_{j=1}^{p}\Phi_{jrs}(x_j)\right)y_s\right)+e_r, \qquad \text{where } e_r\sim N\big(0,\Omega_{xrr}^{-1}\big).$$

We assume that the observations from different classes, labeled by $k$ where $k$ varies from 1 to $K$, are independent. Given the observed data $\{\mathbf{x}^{i(k)},\mathbf{y}^{i(k)}\}_{i=1}^{n_k}$ for class $k$ with $n_k$ samples, the negative log-likelihood for class $k$ is

$$\tilde{\ell}\big(\boldsymbol{\Theta}^{(k)}\big) = -\sum_{i=1}^{n_k}\log f\big(\mathbf{x}^{i(k)},\mathbf{y}^{i(k)};\boldsymbol{\Theta}^{(k)}\big). \tag{2.2}$$

The minimization of the negative log-likelihood incorporates the calculation of the normalization scalar. Because the mixed model includes the discrete model as a special case which is computationally intractable, directly minimizing (2.2) is challenging and impractical. Instead, we propose to use a computationally efficient and consistent estimation approach, the pseudolikelihood method, which is formed

by the product of all conditional distributions as below:

$$\ell\big(\boldsymbol{\Theta}^{(k)}|\mathbf{x}^{(k)},\mathbf{y}^{(k)}\big) = -\sum_{i=1}^{n_k}\Bigg(\sum_{j=1}^{p}\log p\big(x_j^{i(k)}|\mathbf{x}_{\setminus j}^{i(k)},\mathbf{y}^{i(k)};\boldsymbol{\Theta}^{(k)}\big) + \sum_{r=1}^{q}\log p\big(y_r^{i(k)}|\mathbf{x}^{i(k)},\mathbf{y}_{\setminus r}^{i(k)};\boldsymbol{\Theta}^{(k)}\big)\Bigg). \tag{2.3}$$

The model above treats different biological conditions differently from categorical biological measurements such as SNPs, and estimates multiple biological networks from different biological conditions jointly. Mathematically, one may think of a possibility as treating the biological conditions and the categorical biological measurements such as SNPs equally, and make estimation through one mixed graphical model. This approach may result in edges among biological conditions. Such a network is biologically hard to interpret. Thus, we propose to use a framework of joint mixed graphical models instead of treating biological conditions equal to categorical biological measurements and estimating one mixed graphical model.

Another possibility one may think of is to estimate the networks from different biological conditions separately instead of jointly. As illustrated through a toy example that consists of observations from two classes following two normal distributions with distinct covariance matrices in Danaher, Wang and Witten (2013), estimating networks separately in each class results in less accurate estimates than estimating networks jointly. Thus, we adapt a joint graphical model approach in our application scenario.

It is easy to show that the additive negative log-pseudolikelihood is jointly convex in all the parameters $\{\boldsymbol{\Theta}^{(k)}\}$ over the region $\{\Omega_{xrr}^{(k)} > 0\}$ [see the Supplementary Material, Zhang, Ouyang and Zhao (2017)].

**3. Penalty terms.** In the graphical representation of probability distributions, the absence of an edge between two variables corresponds to conditional independence between the two variables. In the proposed mixed model for each class (with $k$ omitted for simplicity), there are three types of edges:

1. Discrete–discrete: If $\lambda_{jm}(x_j, x_m) = 0$ for all values of $x_j$ and $x_m$, then there is no edge between nodes $x_j$ and $x_m$. The corresponding edge potential is in either $p(x_j|\mathbf{x}_{\setminus j}, \mathbf{y}, \boldsymbol{\Theta})$ or $p(x_m|\mathbf{x}_{\setminus m}, \mathbf{y}; \boldsymbol{\Theta})$ of equation (2.3).
2. Discrete–continuous: If $\eta_{jr} = 0$ for all values of $x_j$ and $\Phi_{jrs} = 0$ for all values of $y_s$, $s \in \{1, \ldots, q\}$, then there is no edge between nodes $x_j$ and $y_r$. The corresponding edge potential parameter is in either $p(x_j|\mathbf{x}_{\setminus j}, \mathbf{y}; \boldsymbol{\Theta})$ or $p(y_r|\mathbf{x}, \mathbf{y}_{\setminus r}; \boldsymbol{\Theta})$ of equation (2.3).
3. Continuous–continuous: If $\Phi_{0rs} = 0$ for all values of $y_r$ and $y_s$, and $\Phi_{jrs} = 0$ for all values of $y_r$, $y_s$ and $x_j$, then there is no edge between nodes $y_r$ and $y_s$. These parameters are in either $p(y_r|\mathbf{x}, \mathbf{y}_{\setminus r}; \boldsymbol{\Theta})$ or $p(y_s|\mathbf{x}, \mathbf{y}_{\setminus s}; \boldsymbol{\Theta})$ of equation (2.3).

In summary, we have the following equivalences:

1. $x_j \perp x_m | \{\mathbf{x}_{\setminus j}, \mathbf{x}_{\setminus m}, \mathbf{y}\} \Longleftrightarrow \boldsymbol{\lambda}_{jm} = 0$,
2. $x_j \perp y_r | \{\mathbf{x}_{\setminus j}, \mathbf{y}_{\setminus r}\} \Longleftrightarrow \boldsymbol{\eta}_{jr} = 0$ and $\forall s \in \{1, \ldots, q\}$, $\boldsymbol{\Phi}_{jrs} = 0$,
3. $y_r \perp y_s | \{\mathbf{y}_{\setminus r}, \mathbf{y}_{\setminus s}, \mathbf{x}\} \Longleftrightarrow \Phi_{0rs} = 0$ and $\forall j \in \{1, \ldots, p\}$, $\boldsymbol{\Phi}_{jrs} = 0$.

To estimate the parameters for high-dimensional data, we assume that the underlying true graph is sparse, and incorporate penalization on the number of edges in the minimization of the additive negative log-pseudolikelihood to obtain a sparse graphical model. For the $l_0$-norm, which is defined as the number of nonzero elements in a vector, that is, $\|\mathbf{u}\|_0 := \{\#i, \text{s.t. } \mathbf{u}_i \neq 0\} = \lim_{q \to 0^+} (\sum_{i=1}^n |\mathbf{x}_i|^q)$, we may use the $l_0$ penalty to infer the graphical model for each class separately. Also, we notice that, for edges involving discrete variables, the absence of that edge requires the entire matrix $\boldsymbol{\lambda}_{jm}$ (for discrete–discrete), or matrix $\boldsymbol{\Phi}_{jrs}$ and vector $\boldsymbol{\eta}_{jr}$ (for discrete–continuous) to be 0. Specifically, we set up the following optimization problem (with $k$ omitted for simplicity):

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}) + \rho \Bigg( & \sum_{j \neq m} \mathbb{I}(\boldsymbol{\lambda}_{jm} \neq \mathbf{0}) + \sum_{j=1}^{p} \sum_{r=1}^{q} \mathbb{I}(\boldsymbol{\eta}_{jr} \neq \mathbf{0}) \\ & + \sum_{r \neq s} \mathbb{I}(\Phi_{0rs} \neq 0) + \sum_{j=1}^{p} \sum_{r \neq s} \mathbb{I}(\boldsymbol{\Phi}_{jrs} \neq \mathbf{0}) \Bigg), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\lambda}_{jm}$ is an $L_j \times L_m$ matrix, $\boldsymbol{\eta}_{jr}$ is a vector with length $L_j$, $\boldsymbol{\Phi}_{jrs}$ is a vector with length $L_j$, $\Phi_{0rs}$ is a scalar, and $\rho$ is a tuning parameter. The function involved is integer valued and nonconvex, and it is generally hard to solve the optimization. In the machine learning literature, such kinds of difficult optimization problems are usually solved through appropriate relaxation, for example, Cheng, Levina and Zhu (2013). Notice that, for any vector $\mathbf{b}$, $\mathbb{I}(\mathbf{b} \neq \mathbf{0}) = 0 \iff \|\mathbf{b}\|_2 = 0$, and for any matrix $\mathbf{B}$, $\mathbb{I}(\mathbf{B} \neq \mathbf{0}) = 0 \iff \|\mathbf{B}\|_F = 0$. Thus, we can replace the $l_0$ norm in optimizing (3.1) with an appropriate convex relation, for example,

$$\min_{\boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}) + \rho \Bigg( \sum_{j \neq m} \|\boldsymbol{\lambda}_{jm}\|_F + \sum_{j=1}^{p} \sum_{r=1}^{q} \|\boldsymbol{\eta}_{jr}\|_2 + \sum_{r \neq s} |\Phi_{0rs}| + \sum_{j=1}^{p} \sum_{r \neq s} \|\Phi_{jrs}(x_j)\|_2 \Bigg).$$

We also notice that, for any vector $\mathbf{b}$, $\|\mathbf{b}\|_2 \leq \|\mathbf{b}\|_1$, and for any matrix $\mathbf{B} = (b_{ij})$, $\|\mathbf{B}\|_F \leq \sum_{ij} |b_{ij}|$. Thus, we can replace $\|\cdot\|_2$ and $\|\cdot\|_F$ with the corresponding upper bound penalties, leading to the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}) + \rho \Bigg( & \sum_{j \neq m} \sum_{x_j=1}^{L_j} \sum_{x_m=1}^{L_m} |\lambda_{jm}(x_j, x_m)| \\ & + \sum_{j=1}^{p} \sum_{r=1}^{q} \sum_{x_j=1}^{L_j} |\eta_{jr}(x_j)| + \sum_{r \neq s} |\Phi_{0rs}| + \sum_{j=1}^{p} \sum_{r \neq s} \sum_{x_j=1}^{L_j} |\Phi_{jrs}(x_j)| \Bigg). \end{aligned} \tag{3.2}$$

For simplicity, we let $C$ denote the indices such that $(\theta_{uv})_{u\in C, v\in C}$ contain parameters $\{\boldsymbol{\lambda}_{jm}\}$, $\{\boldsymbol{\eta}_j\}$, $\{\boldsymbol{\Phi_0}\}$ and $\{\boldsymbol{\Phi}_j\}$ only. Then problem (3.2) can be written as

$$\min_{\boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}) + \rho \sum_{u\in C}\sum_{v\in C} |\theta_{uv}|.$$

After discussing the optimization function for a single class, we now formulate our problem for joint analysis of multiple classes. The basic assumption for joint graphical model analysis across different biological conditions is that there are commonalities shared among multiple classes. We propose two penalization approaches (fused lasso and group lasso) to encourage borrowing information from multiple biological conditions for the estimation of the joint mixed graphical models:

$$\arg\min_{\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)}} \sum_{k=1}^{K} \ell\big(\boldsymbol{\Theta}^{(k)}\big) + P\big(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)}\big). \tag{3.3}$$

Specifically, we define the penalty terms as follows.

In the case of the fused graphical lasso,

$$\begin{aligned}
P\big(\{\boldsymbol{\Theta}^{(k)}\}\big) = \rho_1 \sum_{k=1}^{K} \Bigg( & \sum_{j\neq m}\sum_{x_j=1}^{L_j}\sum_{x_m=1}^{L_m} |\lambda_{jm}^{(k)}(x_j,x_m)| \\
& + \sum_{j=1}^{p}\sum_{r=1}^{q}\sum_{x_j=1}^{L_j} |\eta_{jr}^{(k)}(x_j)| + \sum_{r\neq s} |\Phi_{0rs}^{(k)}| + \sum_{j=1}^{p}\sum_{r\neq s}\sum_{x_j=1}^{L_j} |\Phi_{jrs}^{(k)}(x_j)| \Bigg) \\
& + \rho_2 \sum_{k<k'} \Bigg( \sum_{j\neq m}\sum_{x_j=1}^{L_j}\sum_{x_m=1}^{L_m} |\lambda_{jm}^{(k)}(x_j,x_m) - \lambda_{jm}^{(k')}(x_j,x_m)| \\
& + \sum_{r,s} |\Phi_{0rs}^{(k)} - \Phi_{0rs}^{(k')}| + \sum_{j=1}^{p}\sum_{r=1}^{q}\sum_{x_j=1}^{L_j} |\eta_{jr}^{(k)}(x_j) - \eta_{jr}^{(k')}(x_j)| \\
& + \sum_{j=1}^{p}\sum_{r\neq s}\sum_{x_j=1}^{L_j} |\Phi_{jrs}^{(k)}(x_j) - \Phi_{jrs}^{(k')}(x_j)| \Bigg) \\
= \rho_1 \sum_{k=1}^{K}\sum_{u\in C}\sum_{v\in C} & |\theta_{uv}^{(k)}| + \rho_2 \sum_{k<k'}\sum_{u\in C}\sum_{v\in C} |\theta_{uv}^{(k)} - \theta_{uv}^{(k')}|,
\end{aligned}$$

where $\rho_1$ and $\rho_2$ are tuning parameters. The fused graphical lasso penalty implies that graphs from multiple biological conditions are the same except for a few edges.

In the case of the group graphical lasso,

$$
\begin{aligned}
P(\{\boldsymbol{\Theta}^{(k)}\}) = \rho_1 \sum_{k=1}^{K} \Bigg( & \sum_{j \neq m} \sum_{x_j=1}^{L_j} \sum_{x_m=1}^{L_m} |\lambda_{jm}^{(k)}(x_j, x_m)| \\
& + \sum_{j=1}^{p} \sum_{r=1}^{q} \sum_{x_j=1}^{L_j} |\eta_{jr}^{(k)}(x_j)| + \sum_{r \neq s} |\Phi_{0rs}^{(k)}| + \sum_{j=1}^{p} \sum_{r \neq s} \sum_{x_j=1}^{L_j} |\Phi_{jrs}^{(k)}(x_j)| \Bigg) \\
& + \rho_2 \Bigg( \sum_{j \neq m} \sum_{x_j=1}^{L_j} \sum_{x_m=1}^{L_m} \sqrt{\sum_{k=1}^{K} \big(\lambda_{jm}^{(k)}(x_j, x_m)\big)^2} + \sum_{r,s} \sqrt{\sum_{k=1}^{K} \big(\Phi_{0rs}^{(k)}\big)^2} \\
& + \sum_{j=1}^{p} \sum_{r=1}^{q} \sum_{x_j=1}^{L_j} \sqrt{\sum_{k=1}^{K} \big(\eta_{jr}^{(k)}(x_j)\big)^2} + \sum_{j=1}^{p} \sum_{r \neq s} \sum_{x_j=1}^{L_j} \sqrt{\sum_{k=1}^{K} \big(\Phi_{jrs}^{(k)}(x_j)\big)^2} \Bigg) \\
= \rho_1 & \sum_{k=1}^{K} \sum_{u \in C} \sum_{v \in C} |\theta_{uv}^{(k)}| + \rho_2 \sum_{u \in C} \sum_{v \in C} \sqrt{\sum_{k=1}^{K} {\theta_{uv}^{(k)}}^2},
\end{aligned}
$$

where $\rho_1$ and $\rho_2$ are tuning parameters. The group graphical lasso penalty treats the related biological conditions as one group, and implies that the underlying multiple graphs are the same.

**4. Algorithm.** In this section, we focus on the numerical algorithms to solve the optimization problem proposed above. This constrained optimization problem can be simplified and solved by replacing it with a series of distributed problems through an augmented Lagrangian scheme. We first make the objective function separable by rewriting (3.3) as

$$
\arg\min_{\{\boldsymbol{\Theta}^{(k)}\}, \{\mathbf{Z}^{(k)}\}} \sum_{k=1}^{K} \ell\big(\boldsymbol{\Theta}^{(k)}\big) + P(\mathbf{Z}), \tag{4.1}
$$

subject to the constraint that $\mathbf{Z}^{(k)} = \boldsymbol{\Theta}^{(k)}$ for $k = 1, \ldots, K$, where $\{\mathbf{Z}\} = \{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(K)}\}$. Then we carry out the function optimization and regularization locally and coordinate them globally via constraints by further rewriting problem (4.1) using the scaled augmented Lagrangian [Boyd et al. (2011), Hestenes (1969)] as

$$
L_\rho\big(\{\boldsymbol{\Theta}\}, \{\mathbf{Z}\}, \{\mathbf{U}\}\big) = \sum_{k=1}^{K} \ell\big(\boldsymbol{\Theta}^{(k)}\big) + P(\mathbf{Z}) + \frac{d}{2} \sum_{k=1}^{K} \big\| \boldsymbol{\Theta}^{(k)} - \mathbf{Z}^{(k)} + \mathbf{U}^{(k)} \big\|_F^2, \tag{4.2}
$$

where $\mathbf{U} = \{\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}\}$ are the dual feasibility-tolerance variables, and $d$ is a scalar. The augmented Lagrangian optimization problem (4.2) can be solved by

the alternating direction method of multipliers (ADMM), which guarantees to converge to the global optimum [Boyd et al. (2011)]. The skeleton of the algorithm at the $i$th iteration includes the following three steps:

(a) $\{\mathbf{\Theta}_{(i)}\} \leftarrow \arg\min_{\{\mathbf{\Theta}\}} L_d(\{\mathbf{\Theta}\}, \{\mathbf{Z}_{(i-1)}\}, \{\mathbf{U}_{(i-1)}\})$,
(b) $\{\mathbf{Z}_{(i)}\} \leftarrow \arg\min_{\{\mathbf{Z}\}} L_d(\{\mathbf{\Theta}_{(i)}\}, \{\mathbf{Z}\}, \{\mathbf{U}_{(i-1)}\})$,
(c) $\{\mathbf{U}_{(i)}\} \leftarrow \{\mathbf{U}_{(i-1)}\} + (\{\mathbf{\Theta}_{(i)}\} - \{\mathbf{Z}_{(i)}\})$.

Please refer to Supplementary Material for details [Zhang, Ouyang and Zhao (2017)]. Briefly, to estimate $\mathbf{\Theta}$, we use a coordinate-wise descent approach to obtain each parameter in $\mathbf{\Theta}$, and directly apply a well-suited proximal gradient algorithm, which can achieve $\varepsilon$-optimality within $O(1/\sqrt{\varepsilon})$ iterations. The convergence rates and properties of proximal gradient algorithms and their accelerated variants have been well studied [Auslender and Teboulle (2006), Beck and Teboulle (2009)]. To update $\mathbf{Z}$, the optimization problem is separable with respect to each pair of elements in the matrix, and thus can be solved using the fused lasso signal approximator in Hoefling (2010) or the group lasso operator in Friedman, Hastie and Tibshirani (2010) depending on the choice of penalty $P$.

We note that separate regressions were used in estimating a single graphical model, including the Gaussian graphical model [Meinshausen and Bühlmann (2006)], the Ising model [Ravikumar, Wainwright and Lafferty (2010)] and the mixed graphical model [Chen, Witten and Shojaie (2015), Cheng, Levina and Zhu (2013), Yang et al. (2013)]. The regression-type approach is computationally convenient by virtue of effective regression tools, such as glmnet [Friedman, Hastie and Tibshirani (2009)]. However, node-wise regression yields asymmetric estimates of edge potentials for an undirected graph, which results in an arbitrary or ad hoc selection of estimates of parameters. The computational diagram employed in our approach can yield symmetric estimates of edge potentials for the undirected graphs. In this respect, the proposed optimization method outperforms the simple approach to parameter estimation via separate node-wise regression. Moreover, as discussed in Section 2, it is not appropriate to treat biological conditions as discrete measurements in the modeling and estimate one mixed graph. Furthermore, estimating networks separately in each class can result in less accurate estimates than estimating networks jointly [Danaher, Wang and Witten (2013)]. Thus, we use the proposed symmetric pseudolikelihood method to jointly estimate mixed graphical models.

It is also notable that our model covers a special situation when there are only continuous variables across multiple biological conditions, which was studied by joint graphical lasso (JGL) [Danaher, Wang and Witten (2013)]. In this simple case, one only needs to estimate precision matrices of multivariate Gaussian random variables, which results in estimated concentration graphs. It has been shown that the thresholded sample covariance graph induces the same connected components as those induced by the estimated concentration graph under the same regularization parameter [Mazumder and Hastie (2012), Witten, Friedman and Simon

(2011)]. With this nice property of a path of graphical lasso solutions, it can result in a faster computation by employing screenings of empirical covariance matrices to determine whether the solution to concentration graphs is block diagonal upon feature permutation, and by performing the JGL algorithm on the features within each block separately [Danaher, Wang and Witten (2013)].

**5. Tuning parameter selection.** For the selection of tuning parameters, we propose the following Bayesian information criterion (BIC) type of approach:

$$\mathrm{BIC}(\rho_1, \rho_2) = -2 \sum_{k=1}^{K} \ln l\big(\boldsymbol{\Theta}^{(k)}_{\rho_1,\rho_2}\big) + E_k \ln(n_k),$$

where $l(\boldsymbol{\Theta}^{(k)}_{\rho_1,\rho_2})$ is the pseudolikelihood for the observations from the $k$th class with the tuning parameters $\rho_1$ and $\rho_2$, and $E_k$ is the number of edges in the $k$th mixed graphical model. It is notable that the proposed BIC-type approach departures from classical BIC approaches by using the pseduolikelihood rather than likelihoods.

We notice that the Akaike information criterion (AIC) approach has been used for the selection of Gaussian graphical models [Danaher, Wang and Witten (2013)]. Analogously, one may use the following AIC-type approach for model selection in our research scenario using the pseudolikelihood:

$$\mathrm{AIC}(\rho_1, \rho_2) = -2 \sum_{k=1}^{K} \ln l\big(\boldsymbol{\Theta}^{(k)}_{\rho_1,\rho_2}\big) + 2E_k.$$

We compared the two model selection criteria through simulation studies as shown in Section 6. Our analysis suggests that the AIC-type approach tends to choose too large but less accurate models compared to the proposed BIC-type criterion. Thus, we use the proposed BIC-type approach in the real application as illustrated in Section 7.

**6. Simulations.** We demonstrate the performance of our approach through simulations. Without loss of generality, in the following and for the simplicity of simulation, we focus on $\boldsymbol{\Omega}_{\mathbf{x}} = \boldsymbol{\Omega}$, that is, $\Phi_{jrs}(x_j) = 0$ for any $j \in \{1, \ldots, p\}$, $r \in \{1, \ldots, q\}$, $s \in \{1, \ldots, q\}$ and $x_j \in \{1, \ldots, L_j\}$. With this assumption, the covariance matrix for the continuous variables is independent of the values of the discrete variables for the same class.

6.1. *Random networks.* We first considered two mixed graphical models (representing two classes), each consisting of 10 categorical (with two levels, 1 and 2) and 10 Gaussian variables. The topologies for the two simulated networks are shown in Figure 1. The left panel of Figure 1 is the adjacent matrix of the mixed graphical model of the first class, while the right panel represents the second class.

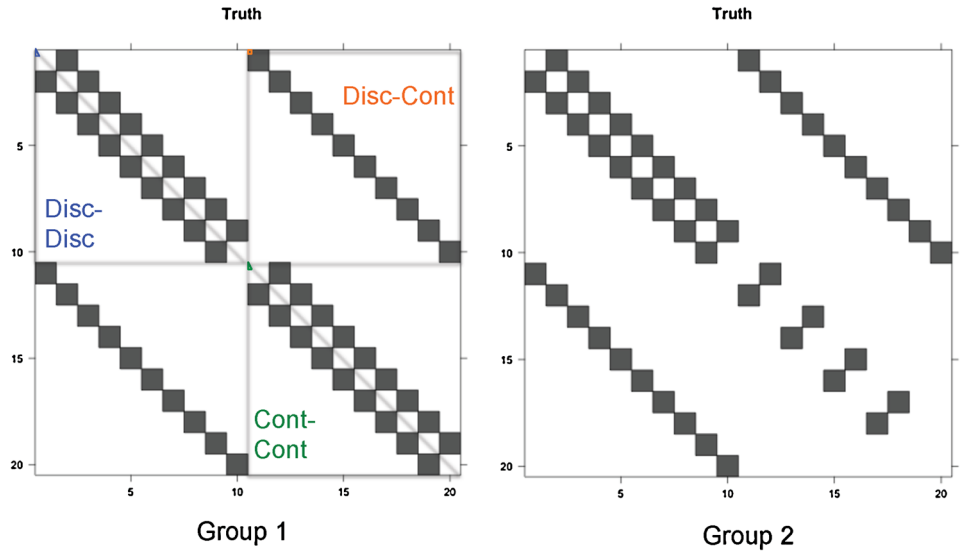


FIG. 1. *The true adjacency matrices of simulated random networks in Section 6.1.*

The first 10 rows/columns correspond to discrete variables, while the second 10 rows/columns correspond to continuous variables. The degree distributions are almost uniform, which are similar to those in the synthetic experiments in Lee and Hastie (2012). Based on the edge sets defined in Figure 1, we assigned nonzero potentials on the edges for each mixed graphical model as follows.

First, we generated a $p \times p$ edge potential matrix connecting discrete variables as below:

$$\lambda_{jm}(x_j, x_m) = \begin{cases} 0.5 & \text{if } (j,m) \in E \text{ and } x_j = 1 \text{ and } x_m = 1, \\ -0.5 & \text{if } (j,m) \in E \text{ and } x_j = 2 \text{ and } x_m = 1, \\ -0.5 & \text{if } (j,m) \in E \text{ and } x_j = 1 \text{ and } x_m = 2, \\ 0.5 & \text{if } (j,m) \in E \text{ and } x_j = 2 \text{ and } x_m = 2, \\ 0 & \text{if } (j,m) \notin E. \end{cases} \tag{6.1}$$

Second, we assigned $2p \times q$ elements for an edge potential matrix connecting discrete and continuous random variables as follows:

$$\eta_{jr}(x_j) = \begin{cases} 1 & \text{if } (j,m) \in E \text{ and } x_j = 1, \\ -1 & \text{if } (j,m) \in E \text{ and } x_j = 2, \\ 0 & \text{if } (j,m) \notin E. \end{cases} \tag{6.2}$$

Third, we generated a precision matrix $\mathbf{\Omega} = (\omega_{rs})$ of continuous variables as below:

$$\omega_{rs} = \begin{cases} 0.25 & \text{if } (r,s) \in E \text{ and } r \neq s, \\ 1 & \text{if } r = s, \\ 0 & \text{if } (r,s) \notin E \text{ and } r \neq s. \end{cases}$$

To draw samples $(\mathbf{x}, \mathbf{y})$ from the joint density $f(\mathbf{x}, \mathbf{y})$, we first drew samples $\mathbf{x} \sim f(\mathbf{x})$ of the following form:

$$f(\mathbf{x}) \propto \exp\left( \sum_{j=1}^{p} \sum_{m=1}^{p} \lambda_{jm}(x_j, x_m) + \frac{1}{2} \eta^{\mathsf{T}}(\mathbf{x}) \mathbf{\Omega}^{-1} \eta(\mathbf{x}) \right)$$

with

$$\big(\eta(\mathbf{x})\big)_r = \sum_{j=1}^{p} \eta_{jr}(x_j).$$

To overcome the difficulty with direct sampling from $f(x)$, we adapted the Gibbs sampling approach in Lee and Hastie (2012). We drew 202,000 samples in total for discrete random variables of each mixed graphical model, and discarded the first 2000 samples which were generated in a burn-in period. Then we took one sample every 100 draws to preserve independence. After sampling x, we sampled $y$ from the conditional distribution $f(y|x)$, which is $N(\mathbf{\Omega}^{-1}(\boldsymbol{\eta_0} + \boldsymbol{\eta}(\boldsymbol{x})), \mathbf{\Omega}^{-1})$ with $\boldsymbol{\eta_0} = \mathbf{0}$.

Using the proposed method DIG, we discovered the network structures for two classes over a range of tuning parameters. We recorded the total number of identified edges for each pair of tuning parameters and calculated the number of true positive edges and the number of false positive edges. It took 52 seconds to obtain the results using the proposed algorithm using a 3-GHz Intel Core i7 processor. For comparison, we also applied the JGL proposed by Danaher, Wang and Witten (2013) by treating the two values (1 and 2) of discrete random variables as continuous variables. Similarly, we calculated the number of true positive edges and the number of false positive edges for a range of tuning parameters of JGL. The results are shown in Figure 2. One can see that our method has better performance with both group lasso and fused lasso penalty schemes. This shows the benefit of explicitly modeling discrete and continuous variables in discovering the underlying networks.

Then we investigate the performance of the tuning parameter selection procedure by checking the sensitivity and specificity of the selected model. The sensitivity and specificity are defined as below, respectively,

$$\text{sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}},$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}},$$

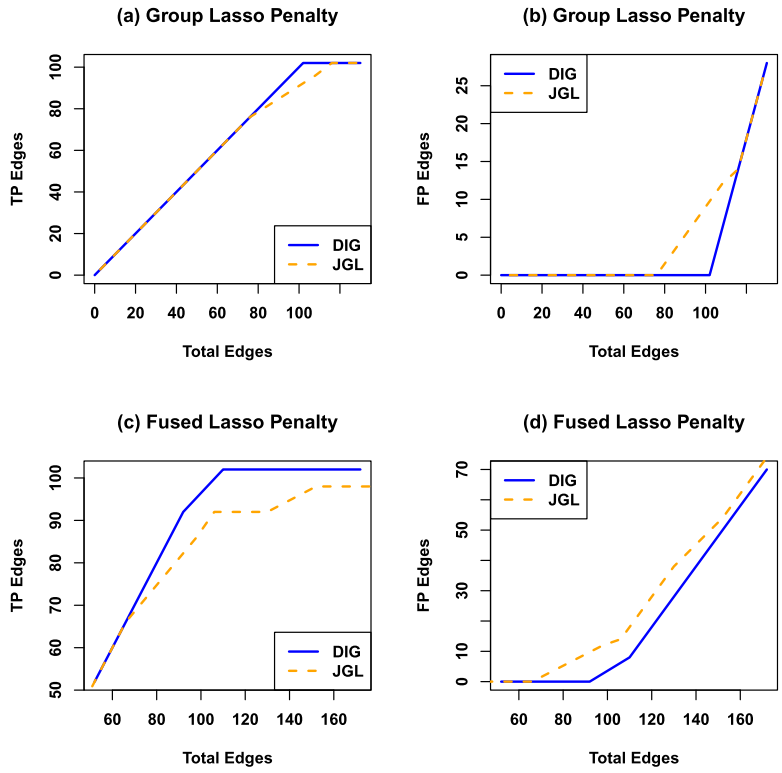


FIG. 2. *Comparison of DIG and JGL on random networks using the simulated data in Section 6.1.*

where TP refers to true positives, FP refers to false positives, TN refers to true negatives and FN refers to false negatives. For each pair of tuning parameters, we calculated the corresponding sensitivity, specificity and score for the proposed BIC-type model selection criterion as shown in Figure 3. Figure 3(a) shows the sensitivities of DIG with the group lasso penalty over a range of tuning parameters, while Figure 3(b) shows the sensitivities of DIG with the fused lasso penalty. Figure 3(c) shows the specificities of DIG with the group lasso penalty over a range of tuning parameters, while Figure 3(d) shows the specificities of DIG with the fused lasso penalty. Figures 3(e) and (f) show the corresponding BIC-type scores for the group lasso and fused lasso, respectively, with the selected model for each type of penalty indicated by a purple diamond. The results show that DIG achieves high sensitivities (1 for both group lasso and fused lasso penalties) and specificities (0.93 for group lasso penalty, 0.96 for fused lasso penalty) with the proposed BIC-type model selection approach. To compare, we also investigated model selection performance through the AIC-type approach. The orange circles in Figure 3 indicate the corresponding selected models for the group lasso penalty and the fused lasso penalty. In this simulation study, although the graphical models selected by

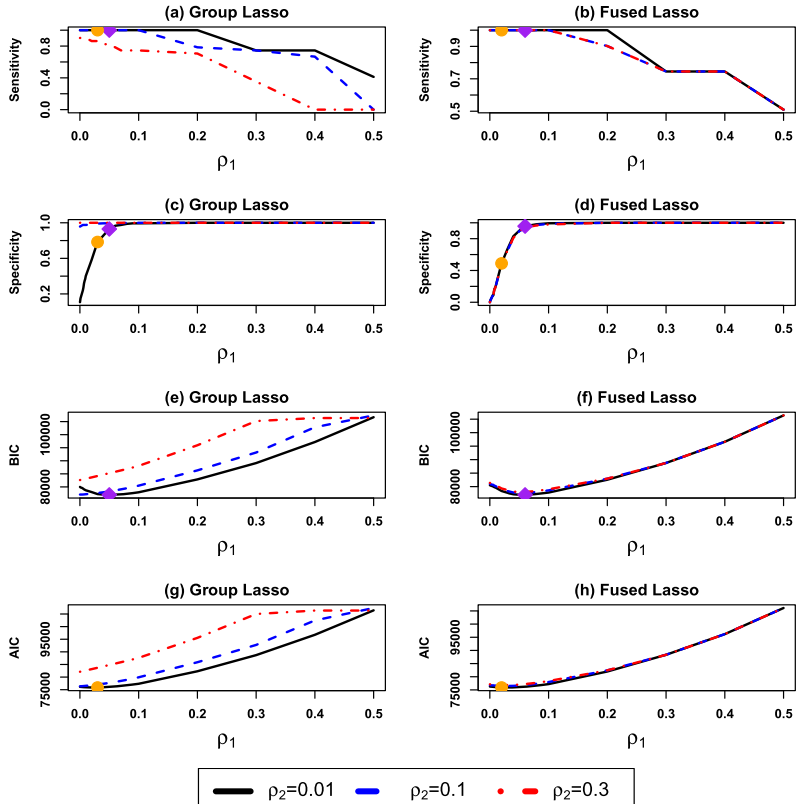


FIG. 3. *Performance of DIG with a range of tuning parameter pairs on random networks using the simulated data in Section 6.1. Purple diamonds indicate the model selected by the proposed BIC procedure. Orange circles indicate the model selected by the AIC-type criterion.*

the AIC-type approach can achieve as high sensitivities as the BIC-type approach, the corresponding specificities are lower (0.78 for group lasso penalty, 0.49 for fused lasso penalty) than those of the BIC-type approach. The results suggest that the AIC-type approach tends to choose a larger but less accurate model than the BIC-type approach.

6.2. *Scale-free networks*. It has been shown that many real networks are scale-free, of which degree distributions follow power law. In this simulation, we investigate the performance of our approach for scale-free networks where the probability that a node has a connectivity of $d$ is proportional to $d^{-\gamma}$. It has been found that the $\gamma$ values for real-world networks usually vary from 2 to 3 [Albert, Jeong and Barabási (2000), Barabási and Albert (1999), Govindan and Tangmunarunkit (2000), Jeong et al. (2001), Yook, Oltvai and Barabási (2004)]. Thus, we randomly generated two networks for two classes where $\gamma$ is 2.433 and 2.317, respectively.

The nodes in each network consist of 10 discrete variables and 10 continuous variables. The simulated structures of the two scale-free networks guided us to assign edge potentials. We used the formula of equation 6.1 for the potentials of the edges connecting discrete variables. We used the formula of equation 6.2 for the potentials of the edges between discrete and continuous variables. For the potentials of the edges connecting continuous variables, we drew random values using the following approach such that the precision matrix $\mathbf{\Omega}$ is a positive definite matrix. For each class, we first created a $q \times q$ matrix with ones on the diagonal and zeros on elements not corresponding to network edges. Then we drew nonzero random values on elements corresponding to edges. To obtain each of these nonzero random values, we first drew a random number $a$ from a uniform distribution $U(0, 1)$, then we randomly picked a number from $\{0.1, -0.4\}$ with equal probability. The nonzero random values assigned to the elements corresponding to edges are $0.3a + b$. Then we divided each off-diagonal element by 1.5 times the sum of the absolute values of the off-diagonal elements in its row. Finally, we added the transpose of the matrix to the matrix itself to achieve a symmetric and positive definite matrix $\mathbf{\Omega}$. To draw samples for each graphical model, we adapted the sample generation procedure in Section 6.1. We set the burn-in threshold as 2000 in the Gibbs sampler. For each graphical model, we took one observation every 100 draws to preserve independence, and obtained 2000 samples in total for the following investigation.

Using the proposed DIG approach, we discovered network topologies using the simulated samples over a range of tuning parameters. Figures 4(a) and (b) show the sensitivities of DIG with group lasso penalty and fused lasso penalty, respectively, while Figures 4(c) and (d) show the corresponding specificities. Figures 4(e) and (f) show the corresponding scores of the proposed BIC-type approach, with the selected model for each type of penalty indicated by a purple diamond. The results show that DIG can achieve high sensitivity (1 for both group lasso and fused lasso penalties) and specificity (0.87 for group lasso penalty, 0.86 for fused lasso penalty) with the proposed BIC-type of approach. It suggests that the proposed BIC-type model selection approach can select suitable tuning parameters. We also investigated the effects of sample sizes on the performance of DIG. With a smaller sample size of 800, we obtained sensitivity as high as 1 for both group lasso and fused lasso, with a considerable specificity of 0.61 for group lasso and 0.62 for fused lasso.

We also compared our approach with JGL in Danaher, Wang and Witten (2013). The results are shown in Figure 5. In the scenario of scale-free mixed networks, our method outperforms JGL with higher true positive discovery rates and lower false positive discovery rates in both group lasso and fused lasso penalty schemes.

**7. Application to cancer genomic data.** We applied our DIG approach to TCGA datasets of two cancer types: colorectal carcinoma (coadread) and breast invasive carcinoma (brca). We obtained the mutation and copy number variation (CNV) data compiled by Ciriello et al. (2013) for 491 coadread subjects and 488

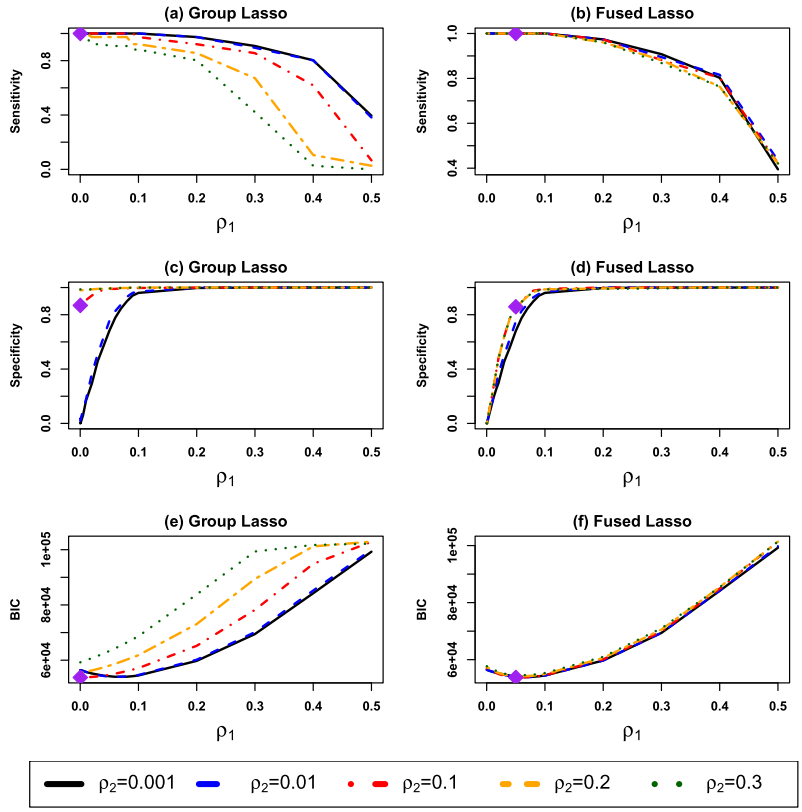


FIG. 4. *Performance of DIG with a range of tuning parameter pairs on scale-free networks using the simulated data in Section 6.2. Purple diamonds indicate the models* (*group lasso penalty*: $\rho_1 = 0.0001$, $\rho_2 = 0.1$; *fused lasso penalty*: $\rho_1 = 0.05$, $\rho_2 = 0.1$) *selected by the proposed BIC procedure.*

brca subjects, respectively. We used the PI3K-mTOR-AKT pathway to illustrate our method. Our data includes 62 genes with mutation information and their corresponding copy number measurements for each subject. We treated mutations as discrete variables with two levels representing the presence and absence of mutations, and copy number variations as continuous variables. We used the fused lasso penalty to the datasets from the two cancer types, and the proposed BIC approach to choose the tuning parameters ($\rho_1 = 0.3$, $\rho_2 = 0.5$). The selected optimum tuning parameter $\rho_2 = 0.5$ for fused lasso indicates similarities exist in our data between coadread and brca. It took 134 seconds to obtain the results using a 3-GHz Intel Core i7 processor with our current DIG implementation. We have identified 1660 edges for the coadread class and 1632 edges for the brca class. Among the interactions of coadread, 16% of them are mutation-mutation interactions, 42.3% of them are mutation-CNV interactions, and the rest are CNV-CNV interactions. For brca, 16.2% of edges are mutation-mutation interactions, 43% of them are

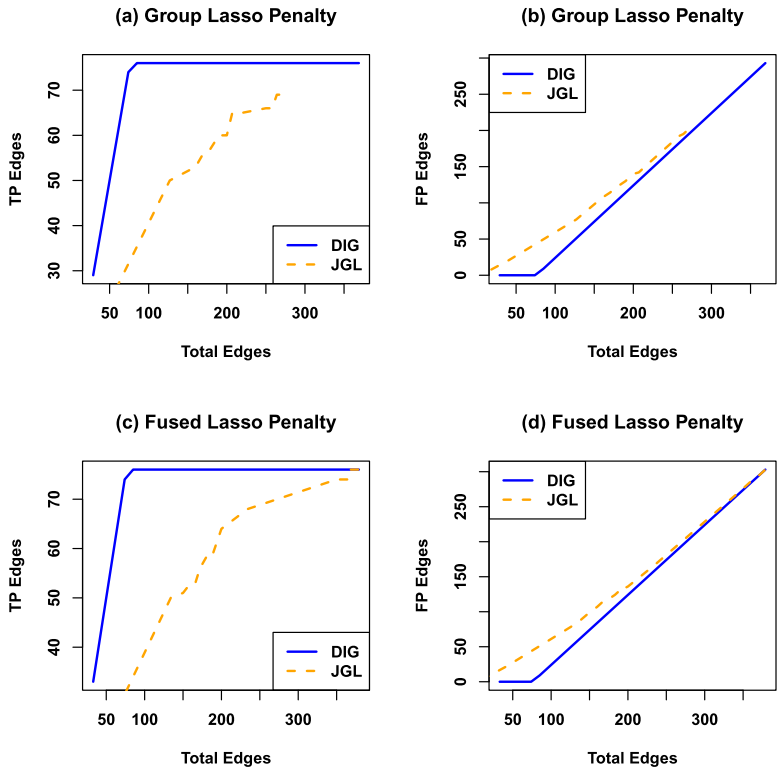


FIG. 5. *Comparison of DIG and JGL on scale-free networks using the simulated data in Section 6.2.*

mutation-CNV interactions, and the rest are CNV–CNV interactions. The two tumor networks share 1584 edges. We studied the community structures or modules in the identified networks through the eigenspectrum decomposition of the modularity matrices described in Newman (2006). We found four modules in coadread with sizes of 16, 66, 22 and 20. We also found two modules in brca with sizes 58 and 66. Interestingly, the second module is shared in both coadread and brca. This module is shown in the left panel of Figure 6 with common interactions in coadread and brca plotted. The nodes with the highest degrees are indicated by their names. The hubs in this common module for both coadread and brca are important oncogenes including TP53 that play important roles in many cancers. We also performed enrichment analysis using INGENUITY (www.ingenuity.com) on the genes evolved in this common module. As shown in the right panel of Figure 6, we found that the identified functions are very relevant to the studied biological context and critical to tumor development, for example, melanoma signaling and cell-cycle events. The rest of the tumor-specific communities for coadread and brca are presented in Supplementary Figure 1, and indicated by different colors. The representative nodes for each module with the highest degrees are indicated

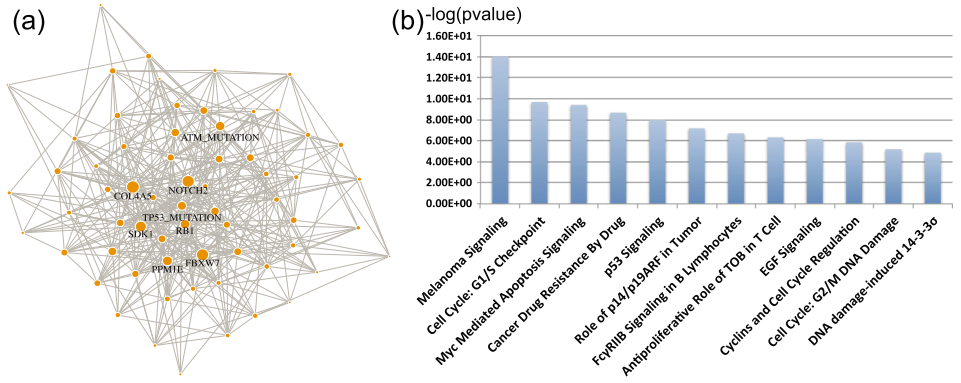


FIG. 6. *Common module and functions for networks identified by DIG.* (a) *The shared module with common interactions presented in both coadread and brca.* (b) *Identified functions by enrichment analysis of genes evolved in the common module. P-values are obtained by Fisher's exact tests.*

by their names as well. We also plotted the detailed tumor-specific interactions in Supplementary Figure 2. There are 76 coadread-specific interactions and 48 brca-specific interactions. Nodes are shown in different colors corresponding to degree differences between the two types of tumors. Specifically, red nodes have the same degrees in the two conditions, while blue nodes have higher degrees in the coadread condition, and orange nodes have higher degrees in the brca condition. We also zoomed in some known oncogenes with their names shown in the pictures. We identified genetic variants that are known to be implicated in individual cancer types. For example, the MTOR copy number has a higher degree in coadread versus brca, which is consistent with the activation of the PI3K-mTOR-AKT pathway in coadread [Ciriello et al. (2013)]. Also, the BRCA1 mutation has a higher degree in brca versus coadread, which corresponds to the inactivation of BRCA1 in breast tumors that leads to defective cell cycle arrest in response to DNA damage [Network et al. (2012)].

To compare, we also performed the analysis without considering the similarities of coadread and brca, for which $\rho_2 = 0$. We selected the optimal value for the tuning parameter controlling sparsity as $\rho_1 = 0.5$. We found 1222 edges for coadread and 981 edges for brca, respectively. Among them, there are 1000 coadread-specific interactions, 759 brca-specific interactions and 222 common interactions. The number of common interactions in this analysis is much less than that in the joint analysis above. We also performed modularity analysis for the resulting networks. We found three modules for caodread with sizes of 67, 1 and 56, as well as four modules for brca with sizes of 36, 45, 22 and 21. Moreover, the results show that the shared edges cannot form a common community. Furthermore, we also applied JGL [Danaher, Wang and Witten (2013)] to this cancer dataset by treating mutations as continuous variables of CNVs. It ended up with 890 interactions for coadread and 1218 interactions for brca. Among these interactions,

for coadread, 1.69% are mutation-mutation interactions and 21.5% are mutation-CNV interactions. For brca, 0.41% are mutation-mutation interaction and 14.4% are mutation-CNV interactions. The results suggest that JGL is less capable of identifying interactions related to mutations. Specifically, JGL identified only one interaction (NRG1) associated with the mTOR mutation in coadread, and no interaction was identified with the mTOR mutation in brca. However, DIG identified 20 and 19 interactions associated with the mTOR mutation in coadread and brca, respectively. In coadread, DIG suggests that the mTOR mutation is associated with the mutations of TP53, RB1, MAP3K1, COL4A5 and PTEN, as well as CNVs of MTOR, ARID1A, DNMT3A, TET2, FBXW7, SDK1, NRG1, CDKN1B, HCN4, CTCF, CDH1, MAP2K4, SMAD4, PHLPP1 and EP300. In brca, DIG indicates that the mTOR mutation is connected with the mutations of TP53, RB1, MAP3K1, COL4A5 and PTEN, as well as MTOR, DNMT3A, TET2, FBXW7, SDK1, NRG1, CDKN1B, HCN4, CTCF, CDH1, MAP2K4, SMAD4, PHLPP1 and EP300. As the PI3K-mTOR-AKT pathway is the biological context considered in our application, JGL is likely to miss important interactions relevant to the mTOR mutation compared to DIG. One of the evidences is that activation of p53 inhibits mTOR activity, and inhibited mTOR also affects p53 activity [Feng et al. (2005)].

**8. Discussion.** In this paper, we have proposed a coherent statistical framework, DIG, for the problem of estimating multiple related mixed graphical models from high-dimensional data with both discrete and continuous variables and with observations belonging to distinct but related biological conditions. The application has been illustrated using cancer studies. DIG is a general statistical framework that can be applied to the genomics of other diseases. For future work, it is natural to extend the proposed framework employing exponential families for a mixed graphical model [Chen, Witten and Shojaie (2015), Yang et al. (2013)]. Furthermore, it would be interesting to develop hypothesis testing methods such that the final mixed graphical models are accompanied by a $p$-value on each edge and an overall estimate of edge false discovery rate. A systematic investigation of model selections and hypothesis testing for the components of the mixed graphical models would be important future work.

**Acknowledgments.** We thank the Associate Editor, two referees and Zhiyi Chi for their careful reading of the manuscript and many helpful suggestions.

## SUPPLEMENTARY MATERIAL

**Supplement to "A statistical framework for data integration through graphical models with application to cancer genomics."** (DOI: 10.1214/16-AOAS998SUPP; .pdf). We present technical and methodological details regarding the model and algorithm in Section 2 and 4. Furthermore, complementary results for the application in Section 7 are provided.

## REFERENCES


ALBERT, R., JEONG, H. and BARABÁSI, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* **406** 378–382.

AUSLENDER, A. and TEBOULLE, M. (2006). Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.* **16** 697–725 (electronic). MR2197553

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

BECK, A. and TEBOULLE, M. (2009). Gradient-based algorithms with applications to signal recovery. *Convex Optim. Signal Process. Commun.* 42–88.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

CHEN, X., SLACK, F. J. and ZHAO, H. (2013). Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions. *Bioinformatics* **29** 2137–2145.

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64.

CHENG, J., LEVINA, E. and ZHU, J. (2013). High-dimensional mixed graphical models. Preprint. Available at arXiv:1304.2810.

CHUN, H., CHEN, M., LI, B. and ZHAO, H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Front. Genet.* **4** Article ID 294. DOI:10.3389/fgene.2013.00294.

CIRIELLO, G., MILLER, M. L., AKSOY, B. A., SENBABAOGLU, Y., SCHULTZ, N. and SANDER, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45** 1127–1133.

DANAHER, P., WANG, P. and WITTEN, D. M. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871

FELLINGHAUER, B., BÜHLMANN, P., RYFFEL, M., VON RHEIN, M. and REINHARDT, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Statist. Data Anal.* **64** 132–152. MR3061894

FENG, Z., ZHANG, H., LEVINE, A. J. and JIN, S. (2005). The coordinate regulation of the p53 and mTOR pathways in cells. *Proc. Natl. Acad. Sci. USA* **102** 8204–8209.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2009). Glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version 1.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group lasso. Technical report, Dept. Statistics, Stanford Univ., Stanford.

GE, H., WALHOUT, A. J. and VIDAL, M. (2003). Integrating 'omic' information: A bridge between genomics and systems biology. *Trends Genet.* **19** 551–560.

GOVINDAN, R. and TANGMUNARUNKIT, H. (2000). Heuristics for Internet map discovery. In *Proceedings IEEE INFOCOM* 2000. *Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies* **3** 1371–1380. IEEE, New York.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2010). Joint structure estimation for categorical Markov networks. Technical report, Dept. Statistics, Univ. of Michigan, Ann Arbor.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206

HAWKINS, R. D., HON, G. C. and REN, B. (2010). Next-generation genomics: An integrative approach. *Nat. Rev. Genet.* **11** 476–486.

HECKER, M., LAMBECK, S., TOEPFER, S., VAN SOMEREN, E. and GUTHKE, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **96** 86–103.

HESTENES, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* **4** 303–320. MR0271809

HOEFLING, H. (2010). A path algorithm for the fused lasso signal approximator. *J. Comput. Graph. Statist.* **19** 984–1006. Supplementary materials available online. MR2791265

HÖFLING, H. and TIBSHIRANI, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

JEONG, H., MASON, S. P., BARABÁSI, A-L. and OLTVAI, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411** 41–42.

JOYCE, A. R. and PALSSON, B. Ø. (2006). The model organism as a system: Integrating "omics" data sets. *Nat. Rev., Mol. Cell Biol.* **7** 198–210.

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LEE, J. D. and HASTIE, T. J. (2012). Learning mixed graphical models. Preprint. Available at arXiv:1205.5012.

LI, B., CHUN, H. and ZHAO, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Amer. Statist. Assoc.* **107** 152–167. MR2949348

MAZUMDER, R. and HASTIE, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.* **13** 781–794. MR2913718

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MYERS, C. L. and TROYANSKAYA, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **23** 2322–2330.

MYERS, C. L., ROBSON, D., WIBLE, A., HIBBS, M. A., CHIRIAC, C., THEESFELD, C. L., DOLINSKI, K. and TROYANSKAYA, O. G. (2005). Discovery of biological networks from diverse functional genomic data. *Genome Biol.* **6** Article ID R114. DOI:10.1186/gb-2005-6-13-r114.

MYERS, C. L., BARRETT, D. R., HIBBS, M. A., HUTTENHOWER, C. and TROYANSKAYA, O. G. (2006). Finding function: Evaluation methods for functional genomic data. *BMC Genomics* **7** 187.

NETWORK, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** Article ID 036104. MR2282139

OUYANG, Z., ZHOU, Q. and WONG, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **106** 21521–21526.

PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

RITCHIE, M. D., HOLZINGER, E. R., LI, R., PENDERGRASS, S. A. and KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16** 85–97.

SHEN, K. and TSENG, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* **26** 1316–1323.

TOMCZAK, K., CZERWIŃSKA, P. and WIZNEROWICZ, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19** A68–A77.

TROYANSKAYA, O. G., DOLINSKI, K., OWEN, A. B., ALTMAN, R. B. and BOTSTEIN, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. USA* **100** 8348–8353.

VARAMBALLY, S., YU, J., LAXMAN, B., RHODES, D. R., MEHRA, R., TOMLINS, S. A., SHAH, R. B., CHANDRAN, U., MONZON, F. A., BECICH, M. J. et al. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* **8** 393–406.

WITTEN, D. M., FRIEDMAN, J. H. and SIMON, N. (2011). New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.* **20** 892–900. MR2878953

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2013). On graphical models via univariate exponential family distributions. Preprint. Available at arXiv:1301.4183.

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129

YOOK, S.-H., OLTVAI, Z. N. and BARABÁSI, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* **4** 928–942.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, Y., OUYANG, Z. and ZHAO, H. (2017). Supplement to "A statistical framework for data integration through graphical models with application to cancer genomics." DOI:10.1214/16-AOAS998SUPP.

Y. ZHANG
DEPARTMENT OF STATISTICS
INSTITUTE FOR SYSTEMS GENOMICS
CENTER FOR QUANTITATIVE MEDICINE
INSTITUTE FOR COLLABORATION
ON HEALTH, INTERVENTION, AND POLICY
THE CONNECTICUT INSTITUTE
FOR THE BRAIN AND COGNITIVE SCIENCES
UNIVERSITY OF CONNECTICUT
STORRS, CONNECTICUT 06269
USA
E-MAIL: yuping.zhang@uconn.edu

Z. OUYANG
THE JACKSON LABORATORY
FOR GENOMIC MEDICINE
DEPARTMENT OF BIOMEDICAL ENGINEERING
DEPARTMENT OF GENETICS AND GENOME SCIENCES
INSTITUTE FOR SYSTEMS GENOMICS
UNIVERSITY OF CONNECTICUT
FARMINGTON, CONNECTICUT 06030
USA
E-MAIL: zhengqing.ouyang@jax.org

H. ZHAO
DEPARTMENT OF BIOSTATISTICS
YALE SCHOOL OF PUBLIC HEALTH
NEW HAVEN, CONNECTICUT 06510
USA
E-MAIL: hongyu.zhao@yale.edu