Statistics in Biosciences
https://doi.org/10.1007/s12561-023-09367-9

Integrative Structural Learning of Mixed Graphical Models via Pseudo-likelihood

Qingyang Liu[1] · **Yuping Zhang**[1]

Received: 10 March 2022 / Revised: 4 March 2023 / Accepted: 10 March 2023

Abstract
Markov random field is a common tool to characterize interactions among a fixed collection of variables. In recent biomedical research, there arise new concerns about the discovery of regulatory and co-expression relationships among different types of features across multiple biological classes. Consequently, we propose a data integration framework to jointly learn multiple mixed graphical models simultaneously. To address the common asymmetry problem in neighborhood selection, we construct a new estimator using regularized pseudo-likelihood, which produces symmetric and consistent estimates of network topologies. We demonstrate the practical merits of our method through learning synthetic networks as well as constructing gene regulatory networks from TCGA data.

Keywords Data integration · Group lasso · Joint modeling · Network · TCGA data

1 Introduction

In recent biomedical research, reconstruction of networks among a group of features is critical for characterizing their functions and mechanisms, so as to unveil etiology of complex diseases and develop targeted therapies. Such correlation networks include typical examples like protein–protein interaction (PPI) network, co-expression network and gene regulatory network. Undirected graphical models, also known as Markov random fields, is a broadly applied tool to identify conditional dependency structures for high-throughput data. The Gaussian graphical model is one of the most representative examples, which is suitable for symmetric and thin-tailed

✉ Yuping Zhang
yuping.zhang@uconn.edu

Qingyang Liu
qingyang.liu@uconn.edu

[1] Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269, USA

Published online: 07 April 2023

continuous data [11, 43]. Broadening the application of graphical models to various types of data, Yang et al. [39] introduced a subclass of Markov Random Fields by assuming all conditional distributions arise from the same univariate exponential family, such as Bernoulli, exponential and Poisson distribution. However, gene regulatory networks usually contain variables of heterogeneous types, for instance, gene expressions, mutations, copy number variations, and epigenetic states, including binary, categorical, count and continuous data. To address this situation, a line of work focused on developing undirected graphical models for heterogeneous data. Lauritzen and Wermuth [16] introduced undirected graphical model for categorical-Gaussian mixtures, and then Cheng et al. [8], Lee and Hastie [17] further simplified it for better scalability. Under the framework of multivariate exponential family, Chen et al. [7], Park et al. [28], Tansey et al.[36], Yang et al. [40] , gradually developed a broader class of mixed Markov Random Fields, which is specified by conditional distributions belonging to potentially different exponential families.

Our research is aimed at constructing regulatory networks among heterogeneous factors, but our data is collected from different biological conditions. These biological conditions can be different types of tissues, subtypes of diseases, phases of progression, or experimental conditions, which potentially share alike regulation mechanisms. Thus, exploiting the prospective similarity among multiple biological classes, integrative modeling is expected to lead to more efficient structural learning. Additionally, it is also critical to capture significant distinctions among networks of different conditions in order to discover class-specific correlations, which requires a flexible and adaptive data integration framework. With regard to joint learning of multiple Gaussian graphical models, Ma and Michailidis [24] assumed prior grouping structure to encourage supervised graph similarity. The works of Guo et al. [13] and Danaher et al. [10] took advantage of different structure-inducing penalties to control network similarity across classes. Under Bayesian framework, Shaddox et al. [32] recently proposed a hierarchical model using Markov random field prior to incorporate data from different biological groups and platforms. As for mixed graphical models, Zhang et al. [44] developed a data-integration framework on mixtures of categorical and Gaussian variables using group lasso and fused lasso regularization.

In this paper, we introduce a joint modeling framework for mixed pairwise exponential graphical models conditionally specified by heterogeneous multi-parameter exponential families. To make structural learning, one common approach is neighborhood selection [25], which can be accomplished by nodewise regularized generalized linear regressions. However, this approach estimates the neighborhood structure of each node separately, which often results in asymmetric edge recovery. To address this problem, our joint learning is established on pseudo-likelihood (PL) regularized by a hierarchical group lasso penalty. It formulates a unified optimization problem to prevent the common asymmetry problem of nodewise regression type approaches, and we also provide statistical guarantee for edge selection consistency. Furthermore, our proposed method is able to perform data-driven joint modeling and take both similarity and divergence among different graphs into account.

To organize the rest of the paper, we first introduce the statistical model as well as the data integration framework in Sect. 2. Theoretical assumptions and results

for edge recovery consistency are discussed in Sect. 3. Then, we present an efficient first-order algorithm to learn multiple networks in Sect. 4 and discuss selection of tuning parameters. We demonstrate the practical merits of the proposed method through simulation studies in Sect. 5. A simple case study is carried out in Sect. 6, where we implement our data integration framework to construct regulatory networks of two different types of human cancer.

2 Integrative Structural Learning of Multiple Heterogeneous Graphical Models

2.1 Pairwise Exponential Markov Random Field

Pairwise exponential Markov Random Field (PEMRF) [22, 28, 36] is a class of Markov random fields to characterize pairwise interactions that can explicitly reveal the underlying conditional dependency structure. In this paper, we assume data come from multiple PEMRF sub-populations with known class labels. The network structures of different classes are highly similar, with only a few numbers of distinctions.

Suppose $\boldsymbol{x} = \{x_1, \ldots, x_p\}$ is a p-variate random vector, of which the p features are potentially heterogeneous, possessing different supports and measures from each other. The joint distribution is a multivariate exponential family specified as

$$\mathcal{P}(\boldsymbol{x};\boldsymbol{\Theta}) = \exp\left\{\sum_{r=1}^{p}\theta_r^{\top}B_r(x_r) + \frac{1}{2}\sum_{r=1}^{p}\sum_{s=1}^{p}\left\langle\theta_{rs}, B_r(x_r)B_s(x_s)^{\top}\right\rangle_F + \sum_{r=1}^{p}C_r(x_r) - A(\boldsymbol{\Theta})\right\}, \tag{1}$$

where $\langle A, B\rangle_F = \mathrm{tr}(A^{\top}B)$ is the Frobenius inner product. This distribution include many popular graphical models as special cases, such as Gaussian graphical model, Ising model and Gaussian–categorical mixture [17]. The collection of all natural parameters is denoted by $\boldsymbol{\Theta} = \left\{\{\theta_r\}_{r=1}^{p}, \{\theta_{rs}\}_{r,s=1}^{p}\right\}$, and $B(x) = \left\{\{B_r(x_r)\}_{r=1}^{p}, \{B_r(x_r)B_s(x_s)^{\top}\}_{r,s=1}^{p}\right\}$ is the corresponding set of sufficient statistics. Since the p features may be of different types, we define the following dimensions of natural parameters and sufficient statistics: $\theta_r \in \mathbb{R}^{m_r}$, $\theta_{rs} \in \mathbb{R}^{m_r \times m_s}$, and $B_r(x_r) \in \mathbb{R}^{m_r}$. For instance, θ_r could be a scalar ($m_r = 1$) for a Gaussian variable or a vector ($m_r > 1$) for a categorical variable with more than two classes. The vertex set of p nodes is denoted by V, and E is the collection of connected edges containing $|E|$ unordered vertex pairs. Without any loss of generality, we assume a symmetric structure in edge potentials, that is, $\theta_{rs} = \theta_{sr}^{\top}$ for any $r < s$. By the Hammersley–Clifford Theorem [15], sparsity in edges potentials reflects conditional dependency, i.e. for any $(r, s) \notin E$, $\theta_{rs} = \theta_{sr}^{\top} = 0$, x_r and x_s are conditionally independent given all other variables. Correspondingly, the set of model parameters can be reduced to

$\boldsymbol{\Theta} = \left\{\{\theta_r\}_{r=1}^p, \{\theta_{rr}\}_{r=1}^p, \{\theta_{rs}\}_{r<s}\right\}$, which still respects the symmetry in undirected graphical model. The normalizing term $A(\Theta) = \log \int_{\mathcal{X}} \exp \exp\{\langle \Theta, B(x)\rangle_F + C(x)\}\nu(\mathrm{d}x) < \infty$ is a finite-valued log-partition function which is intractable in general.

Based on the joint distribution, the conditional distribution of variable x_r given all other variables $\boldsymbol{x}_{\backslash r}$ can be easily derived, that is

$$\begin{aligned}&\mathcal{P}_{r|\backslash r}(x_r|\boldsymbol{x}_{\backslash r};\boldsymbol{\Theta})\\ &\quad= \exp\left\{\phi_r^\top B_r(x_r) + \frac{1}{2}B_r(x_r)^\top \theta_{rr} B_r(x_r) + C_r(x_r) - A_r\left(\phi_r;\theta_{rr}\right)\right\},\end{aligned} \tag{2}$$

where $\phi_r = \theta_r + \sum_{s\in\mathcal{N}(r)} \theta_{rs}B_s(x_s)$ denotes the natural parameter of this node-conditional distribution; $\mathcal{N}(r)$ is the neighborhood of vertex r. Consequently, we are able to specify mixed graphical models through their full conditional distributions [7]. The node conditional distributions can be a mixture of different exponential family distributions, for example, categorical distribution, Poisson distribution and exponential distribution, by assigning different log-partition functions, measures, and supports. These distributions cover a majority of variable types in biomedical research. Please refer to the following Table 1 for details.

It is commonly observed that necessary constraints should be imposed to make the conditionally specified joint distribution normalizable. Yang et al. [40] as well as Chen et al. [7] established the conditions for several simple mixtures to be normalizable in their separate works. However, such constraints can be more complex in general cases. In this paper, we only consider valid conditionally specified mixed models that can be normalized.

2.2 Data Integration Through Pseudo-likelihood (PL)

Although the log-likelihood is convex for exponential family, the intractable log-partition function in form of high-dimensional integral still makes likelihood-based inference computationally infeasible for high-dimensional problems. To address this issue, a natural way to make inference is implied by the node conditional distribution (2),

Table 1 Conditional distributions from exponential family

Distribution	Support	$B_r(x_r)$	θ_{rr}	$A_r(\phi_r;\theta_{rr})$	$C_r(x_r)$
Binary	$\{0,1\}$	$\mathbb{I}(x_r = 1)$	0	$\log(1+e^{\phi_r})$	0
Categorical	$\{0,1,\ldots,m_r\}$	$\begin{bmatrix}\mathbb{I}(x_r=1)\\ \vdots \\ \mathbb{I}(x_r=m_r)\end{bmatrix}$	0	$\log(1+\sum_{j=1}^{m_r} e^{\phi_{r:j}})$	0
Poisson	$\{0,1,\ldots,\infty\}$	x_r	0	e^{ϕ_r}	$-\log(x_r!)$
Trunc. Poisson	$\{0,1,\ldots,R\}$	x_r	0	$\log\left(\sum_{k=0}^{R} e^{k\phi_r}/k!\right)$	$-\log(x_r!)$
Gaussian	$\mathbb{R}$	x_r	$-\frac{1}{\sigma_r^2}$	$-\frac{\phi_r^2}{2\theta_{rr}} - \frac{1}{2}\log(-\theta_{rr})$	$-\frac{1}{2}\log(2\pi)$
Exponential	$\mathbb{R}^+$	x_r	0	$-\log(-\phi_r)$	0

which is based on nodewise regression [25, 36, 40]. This type of approach estimates the neighborhood of each variable separately, often resulting in asymmetric recovery of edges. As a result, some ad-hoc procedures have to be implemented to construct the final topology from all nodewise neighborhoods. Another variational likelihood method was proposed by Park et al. [28], which substitutes the log-partition function by a Gaussian entropy bound. It has good computational efficiency, but sacrifices edge recovery consistency in some situations, for example, graphs with cycles [21, 22]. Instead, we multiply all node conditional distributions to formulate a pseudo-likelihood [PL; 6] problem for our data integration framework. Compared to the Guassian approximation, structural learning based on PL has better theoretical guarantees for cyclic graphs, and meanwhile has similar computational complexity.

Given observations $\{\boldsymbol{x}^i\}_{i=1}^n$ independently sampled from $\mathcal{P}(\boldsymbol{x};\boldsymbol{\Theta})$ specified in (1), the negative log-PL can be expressed by the sum of p negative logarithms of conditional distribution functions

$$\ell(\boldsymbol{\Theta};\{\boldsymbol{x}^i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{p} \left[-\phi_r^{i\top} B_r(x_r^i) - \frac{1}{2}\left\langle \theta_{rr}, B_r(x_r^i)B_r(x_r^i)^\top \right\rangle_F + A_r\left(\phi_r^i;\theta_{rr}\right)\right],$$

where $\phi_r^i = \theta_r + \sum_{s\neq r}\theta_{rs}B_s(x_s^i)$ is a linear function of θ_r and θ_{rs}. Thus, the negative log-PL can be viewed as a sum of p generalized linear regressions using canonical link, but the p regressions are not separable because we enforce symmetry in edge potentials. As an approximation to the intractable true likelihood, the PL can lead to good parameter estimates.

In many biological network studies, data of different types are observed from distinct biological conditions, but these conditions may share similar regulation mechanisms. To simultaneously learn the topology of all graphs, we adopt a joint modeling strategy to enhance the efficiency in utilizing data. For instance, suppose the data come from K different biological conditions, all with the same set of features $V = \{1, \ldots, p\}$. For each individual class k, the p features jointly follow a PEMRF specified by parameter set $\boldsymbol{\Theta}^{(k)}$, from which we have n_k independent samples. Next, on the basis of the PL of PEMRF, combining the data from all K biological conditions, we set up the following regularized optimization problem to perform joint structural learning

$$\min_{\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)}} \frac{1}{N}\sum_{k=1}^{K} n_k \ell(\boldsymbol{\Theta}^{(k)};\{\boldsymbol{x}^{i(k)}\}_{i=1}^{n_k}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)}), \tag{3}$$

where $N = \sum_{k=1}^{K} n_k$ is the total sample size. The penalty function P is identical to the one in Liu and Zhang [22], which is designated to encourage borrowing information across different classes and induce graph sparsity. To be specific, the penalty function is defined as

$$
\begin{aligned}
&P_{\lambda_1,\lambda_2}(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)}) \\
&\quad= \lambda_1 \sum_{r<s}\sum_{k=1}^{K} \eta_{rs}^{(k)} \left\|\theta_{rs}^{(k)}\right\|_F + \lambda_2 \sum_{r<s} w_{rs} \left\|\left(\theta_{rs}^{(1)},\ldots,\theta_{rs}^{(K)}\right)\right\|_F \\
&\quad= \lambda \left[(1-\alpha) \sum_{r<s}\sum_{k=1}^{K} \eta_{rs}^{(k)} \left\|\theta_{rs}^{(k)}\right\|_F + \alpha \sum_{r<s} w_{rs} \left(\sum_{k=1}^{K} \left\|\theta_{rs}^{(k)}\right\|_F^2 \right)^{1/2} \right].
\end{aligned} \tag{4}
$$

With a similar intuition to sparse group lasso [12, 34], it consists of two parts of group lasso penalties [42] in a hierarchical structure. This penalty function encourages similarity among different biological conditions, but also allows divergence in the meantime, which is helpful in capturing class-specific interactions. Graph topologies $\{\widehat{E}^{(1)},\ldots,\widehat{E}^{(K)}\}$ are determined by the sparsity of edge potentials in the solution $\{\widehat{\boldsymbol{\Theta}}^{(1)},\ldots,\widehat{\boldsymbol{\Theta}}^{(K)}\}$. The PL-based M-estimator is able to maintain symmetric edge sparsity, which helps prevent difficulties of asymmetric estimation for nodewise regression type approaches. We choose weights $w_{rs} = K^{-1/2} \cdot (m_r m_s)^{1/2}$ to balance the penalty for matrices $\theta_{rs}^{(k)}$ of different sizes. The other set of weights take the form $\eta_{rs}^{(k)} = (n_k/N) \cdot (m_r m_s)^{1/2}$, which adjusts for different magnitudes of PL functions with imbalanced sample sizes. The tuning parameters can also be parameterized by an overall penalty $\lambda = \lambda_1 + \lambda_2 > 0$ and a proportion $\alpha = \lambda_2/(\lambda_1 + \lambda_2) \in [0, 1]$ controlling the similarity among classes. Given $\alpha = 1$, the penalty enforces all the K graphs to share the same structure. On the other hand, if $\alpha = 0$, the problem becomes separable, which is equivalent to separate modeling for each biological condition.

It is noteworthy that if our interest is to reconstruct networks consisting of heterogeneous types of features, using only one set of tuning parameters may lead to different levels of penalization on different edge types. As a remedy, we standardize the sufficient statistic $B_{r:j}$, namely the jth entry of $B_r(x_r)$, by pooled mean estimate $\hat{\mu}_{r:j} = N^{-1} \sum_k n_k \hat{\mu}_{r:j}^{(k)}$ and pooled sample standard deviation $\hat{\sigma}_{r:j} = \sqrt{N^{-1} \sum_k n_k \hat{\sigma}_{r:j}^{2(k)}}$. As a linear transformation, standardization on sufficient statistics across all classes does not affect the correlation structure, but the log-partition functions A_r should be transformed accordingly. This procedure can also be justified by the calibrated weighting proposed by Lee and Hastie [17], where edge weights should be proportional to standard deviations of sufficient statistics.

3 Consistency of Edge Recovery

3.1 Assumptions

In this section, we establish high-dimensional edge recovery consistency for learning of one biological condition first, and then generalize the result to our

data integration problem. For the single graph estimation, consider the following regularized PL problem,

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \ \ell(\boldsymbol{\Theta}; \{\boldsymbol{x}^i\}_{i=1}^n) + \lambda \sum_{r<s} w_{rs} \|\theta_{rs}\|_F. \tag{5}$$

The key quantity of the following analysis is the Fisher information matrix, $Q^* = \nabla^2 \ell(\boldsymbol{\Theta}^*)$, which is the Hessian of negative log-PL. Denote the true sparsity set by S^c, which is defined as $S^c = \{(r,s) : \theta^*_{rs} = 0\}$, consisting of indices of all unconnected vertex pairs. Its complement S is the support set, including edge indices, as well as all node potentials in model parameters. The matrix Q^* can be thereby partitioned into the following 2×2 blocked matrix through proper permutation of parameters,

$$Q^* = \begin{bmatrix} Q^*_{SS} & Q^*_{SS^c} \\ Q^*_{S^cS} & Q^*_{S^cS^c} \end{bmatrix}.$$

We assume the true graph structure is sparse, so the maximum degree of the graph is controlled at d. In our context, we use $\|v\|_p$ to denote the ℓ_p norm of vector v. When it is applied to a matrix, it becomes the ℓ_p norm of the vectorized matrix, for example, $\|A\|_\infty = \max_{i,j} |A_{ij}|$. The operator matrix norm induced by ℓ_p norm is denoted by $|||A|||_p$. Specifically, $|||A|||_\infty$ is the maximum ℓ_1 norm of a row of the matrix. Based on these notations, the following assumptions are made to establish edge recovery consistency.

Assumption 1 (*Restricted strong convexity*) There exist constants $L_1 > 0$, $L_2 > 0$ and $B > 0$, such that for any Θ satisfying $\|\Theta - \Theta^*\|_2 \le B$ and any model parameter vector v, we have

$$v^\top \nabla^2 \ell(\boldsymbol{\Theta}) v \ge L_1 \|v\|_2^2,$$

and the hessian of loss function obeys

$$\|\nabla^2 \ell(\boldsymbol{\Theta}) - \nabla^2 \ell(\boldsymbol{\Theta}^*)\|_F \le L_2 \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2.$$

Assumption 2 (*Irrepresentability*) There exists a constant $\tau > 0$, such that

$$\left|\left|\left| Q^*_{S^cS} (Q^*_{SS})^{-1} \right|\right|\right|_\infty \le \frac{w_{\min}(1-\tau)}{w_{\max} m_{\max}},$$

where $m_{\max} = \max_r m_r$, $w_{\min} = \min_{r,s} w_{rs}$, and $w_{\max} = \max_{r,s} w_{rs}$.

The first two conditions are common assumptions in the literature on high-dimensional consistency. The loss function $\ell(\boldsymbol{\Theta})$ is assumed to be twice continuously differentiable. The restricted strong convexity condition ensures the loss function is not "too flat" and guarantees identifiability of $\boldsymbol{\Theta}^*$. The Lipschitz continuity is a result of the duality between strong convexity and strong smoothness. The irrepresentability assumption imposes control on the influence that

the non-edge terms can have on the edge-related terms, so that active parameter groups are not be overly dependent on the inactive parameter groups.

Assumption 3 (*Bounded moments*) For any feature $r \in V$, and $j \in \{1, \ldots, m_r\}$, the first moment of sufficient statistic is bounded, i.e. $|\mathbb{E}(B_{r:j})| < \kappa_m$. Also, the log-partition function A of the joint distribution satisfies the following properties:

$$\max_{|v| \le q} \frac{\partial^2}{\partial \theta_{r:j}^2} A(\mathbf{\Theta}^* + v e_{r:j}) \le \kappa_v,$$

and

$$\max_{|v| \le q} \frac{\partial^2}{\partial \theta_{rr:jj}^2} A(\mathbf{\Theta}^* + v e_{rr:jj}) \le \kappa_h$$

for some constants $q, \kappa_v, \kappa_h > 0$, where $e_{r:j}$ ($e_{rr:jj}$) is a unit vector that is equal to one only at the index corresponding to $\theta_{r:j}$ ($\theta_{rr:jj}$) and zero elsewhere.

Assumption 4 (*Smoothness of conditional distributions*) For any variable $r \in V$, the log-partition function of the nodewise conditional distribution A_r satisfies: there exist functions $\kappa_1(n, p)$ and $\kappa_2(n, p)$ such that, for any $X \in \mathcal{X}$,

$$\frac{\partial^2}{\partial \phi_{r:j}^2} A_r(\phi_r^* + a e_{r:j}; \theta_{rr}^*) \le \kappa_1(n, p),$$

where $\phi_r^* = \theta_r^* + \sum_{s \ne r} \theta_{rs}^* B_s$, and $|a| \le 4q^{-1}\kappa_2(n, p) \max\{\log n, \log p\}$.

The assumptions on the joint and conditional distributions are to construct tail probability bound of the score function $\nabla \ell(\mathbf{\Theta}^*)$. The third assumption requires the first moment of sufficient statistics to be bounded. Also, some higher order moments should be bounded under a small perturbation in true model parameters. Assumption 4 is a smoothness condition to generalize the analyses in Meinshausen et al. [25], Ravikumar et al.[29] to exponential family. Yang et al. [41] has explicitly verified this condition for most exponential family distributions.

3.2 Edge Selection Consistency Result

Under these assumptions, we are able to establish the following edge selection consistency results.

Proposition 1 *Under Assumptions 1, 2, 3, 4, for some constant $\kappa_v < \kappa_3 \le \kappa_h q + \kappa_v$, when sample size n is large enough, we are able to set the regularization parameter λ to satisfy*

$$\frac{16(m_{\max}+1-\tau)}{\tau w_{\min}}\sqrt{\kappa_1(n,p)\kappa_3}\sqrt{\frac{2\log p+2\log m_{\max}+\log 2}{n}}$$
$$\le \lambda \le \min\left\{\frac{8(m_{\max}+1-\tau)}{\tau w_{\min}}\kappa_1(n,p)\kappa_2(n,p)\kappa_3, \left(\frac{\tau w_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)}+w_{\max}\sqrt{pd}\right)^{-2}\right.$$
$$\left.\frac{L_1^2\tau w_{\min}}{16L_2(m_{\max}+1-\tau)}\right\},$$

where M is the total number of parameters in the model. Then, with probability of at least $1-c_1p'^{-2}-\exp(-c_2n)-\exp(-c_3n)$, the solution $\widehat{\mathbf{\Theta}}$ satisfies the following properties:

1. *Consistency* $\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_2 \le \frac{2}{L_1}\left(\frac{\tau w_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)}+w_{\max}\sqrt{pd}\right)\lambda$;
2. *Edge selection consistency* $\widehat{\mathbf{\Theta}}_{S^c}=0$,

where $p'=\max\{n,p\}$;c_1,c_2,c_2 are positive constants. Furthermore, if the true model satisfies

$$\min_{(r,s)\in S}\|\theta^*_{rs}\|_F > \frac{2}{L_1}\left(\frac{\tau w_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)}+w_{\max}\sqrt{pd}\right)\lambda,$$

all edges are correctly estimated, i.e. $\widehat{E}=E^*$.

The proof of this proposition follows the standard primal-dual witness approach. In our analysis, we only consider unrestricted solution. If the model space is a subspace of $\mathbb{R}^M$, similar consistency result of constrained solution can be derived through the general framework in Lee et al. [18]. Our consistency statements are general, which are not limited to Ising and Gaussian graphical models. For Gaussian graphical models where we can choose $\kappa_1(n,p)=1$ and $\kappa_2=\infty$, the sample size required on λ_n to achieve consistent sparsity recovery is $\mathcal{O}(\sqrt{\log p/n})$, the same as nodewise regression [41]. Our result indicates that when the sample size is large enough compared to dimension and graph density, our estimator (5) can recover the correct edge structure with high probability, with a proper selection of tuning parameter.

For the joint modeling problem with fixed K specified by (3), we slightly modify Assumption 2 and obtain a similar consistency result.

Assumption 5 (*Irrepresentability*) Denote $\eta_{\min}=\min_{k,i,j}\eta_{ij}^{(k)}$ and $\eta_{\max}=\max_{k,i,j}\eta_{ij}^{(k)}$. There exists some $\tau>0$ such that for all $k=1,\ldots,K$,

$$\left|\left|\left|Q^{*(k)}_{S^{(k)c}S^{(k)}}\left(Q^{*(k)}_{S^{(k)}S^{(k)}}\right)^{-1}\right|\right|\right|_{\infty} \leq \frac{\eta_{\min}}{\eta_{\max} m_{\max}}(1-\tau).$$

Proposition 2 *Suppose Assumptions 1, 3, 4, 5 are satisfied for all classes, given* $n_1 = \cdots = n_K = n$, *for some constant* $\kappa_v < \kappa_3 \leq \kappa_h q + \kappa_v$, *when sample size* n *is large enough, we are able to set the regularization parameter* λ_1 *to satisfy*

$$\frac{16(m_{\max}+1-\tau)}{K\tau\eta_{\min}}\sqrt{\kappa_1(n,p)\kappa_3}\sqrt{\frac{2\log p + 2\log m_{\max} + \log 2}{n}} \leq \lambda_1 \leq \min\left\{\frac{8(m_{\max}+1-\tau)}{K\tau\eta_{\min}}\kappa_1(n,p)\kappa_2(n,p)\kappa_3,\right.$$

$$\left.\left(\frac{\tau K^2\eta_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)} + K^2\eta_{\max}\sqrt{pd} + hw_{\max}pK^{3/2}\right)^{-2}\frac{L_1^2\tau K\eta_{\min}}{16L_2(m_{\max}+1-\tau)}\right\},$$

and $\lambda_2 \leq \frac{\eta_{\max}\tau}{4w_{\max}(1-\tau)}\cdot\lambda_1 = h\lambda_1$. *Then, with probability of at least* $1 - Kc_1p'^{-2} - K\exp(-c_2n) - K\exp(-c_3n)$, *the solution to the joint learning problem* $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Theta}}^{(1)}, \ldots, \widehat{\boldsymbol{\Theta}}^{(K)})$ *satisfies the following properties*:

1. *Consistency* $\max_k \|\widehat{\Theta}^{(k)} - \Theta^{*(k)}\|_2 \leq \frac{2}{L_1}\left(\frac{\tau K^2\eta_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)} + K^2\eta_{\max}\sqrt{pd} + hw_{\max}pK^{3/2}\right)\lambda_1$

   $\max_k \|\widehat{\Theta}^{(k)} - \Theta^{*(k)}\|_2 \leq \frac{2}{L_1}\left(\frac{\tau K^2\eta_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)} + K^2\eta_{\max}\sqrt{pd} + hw_{\max}pK^{3/2}\right)\lambda_1$;
2. *Edge selection consistency* $\widehat{\boldsymbol{\Theta}}^{(k)}_{S^{(k)c}} = 0$ *for all classes.*

Furthermore, if the true model of the kth class satisfies

$$\min_{(r,s)\in S^{(k)}} \|\theta^{*(k)}_{rs}\|_F > \frac{2}{L_1}\left(\frac{\tau K^2\eta_{\min}\sqrt{M}}{4(m_{\max}+1-\tau)} + K^2\eta_{\max}\sqrt{pd} + hw_{\max}pK^{3/2}\right)\lambda_1,$$

all edges of that class are correctly estimated, i.e. $\widehat{E}^{(k)} = E^{*(k)}$.

In order to achieve full consistency for all K classes, the tuning parameter λ_2 for joint modeling must be sufficiently small. However, in our scenario when data come from distinct but similar conditions, more joint modeling leads to superior performance compared with no data integration. Detailed proofs are provided in the Supplementary Information.

4 Algorithm

4.1 Proximal Gradient Algorithm

The joint structural learning framework is summarized by an optimization problem specified in (3). It can be solved in the scheme of proximal gradient and accelerated proximal gradient algorithms [4]. The proximal gradient algorithm is a first-order method frequently applied in statistical learning when the optimization objective can be decomposed into $f(x) + g(x)$, where f is smooth and convex and g is convex but possibly non-smooth. Another requirement is that the following proximal operator of the second function g can be computed in an efficient manner:

$$\text{prox}_t(x) = \arg\min_u \frac{1}{2t}\|x-u\|_2^2 + g(u).$$

When all these requirements are met, the proximal gradient method solves the original minimization problem with the following first-order model iteratively,

$$\begin{aligned}&\arg\min_u f(x_j) + \nabla f(x_j)^\top (u - x_j) + \frac{1}{2t_j}\left\|x_j - u\right\|_2^2 + g(u)\\ &= \arg\min_u \frac{1}{2t_j}\left\|u - (x_j - t_j \nabla f(x_j))\right\|_2^2 + g(u),\end{aligned}$$

where x_j is the current value after the jth iteration. Therefore, the $(j+1)$th iteration updates the argument value by $x_{j+1} = \text{prox}_{t_j}(x_j - t_j \nabla f(x_j))$. The first-order model can also be viewed as a quadratic approximation to f at x_j, with $\nabla^2 f(x_j)$ replaced by I/t_j. For better convergence of the algorithm, step size t_j is usually determined by backtracking line search. Many acceleration tactics have been further established, such as Auslender and Teboulle [3], Nesterov [26, 27], which slightly change the argument passed to the proximal operator for optimal convergence rate. The theoretical properties of the class of proximal gradient algorithms have been well-studied. The accelerated proximal gradient method can achieve sub-linear to linear convergence rate [4], depending on the strong convexity of the problem.

Our problem is perfectly suitable for the proximal gradient method. In our case, f is the negative log-PL and g is the hierarchical group lasso penalty. The negative log-PL is convex and smooth, and the gradient can be computed efficiently through matrix multiplications. Time complexity of computing the gradient is $\mathcal{O}[N(\sum_{r=1}^p m_r)^2]$. The proximal operator of the penalty function (4) has an analytical solution. Evaluate the following problem,

$$\begin{aligned}\min_{U^{(1)},\ldots,U^{(K)}} \; &\lambda_1 \sum_{r<s}\sum_{k=1}^K \eta_{rs}^{(k)} \left\|u_{rs}^{(k)}\right\|_F + \lambda_2 \sum_{r<s} w_{rs} \left\|\left(u_{rs}^{(1)}, \ldots, u_{rs}^{(K)}\right)\right\|_F\\ &+ \frac{1}{2t}\sum_{k=1}^K \left\|\Theta^{(k)} - U^{(k)}\right\|_F^2,\end{aligned}$$

which is separable across vertex pairs. For a vertex pair $r < s$ of a given class k, the proximal operator takes the form:

$$\text{prox}_t(\theta_{rs}^{(k)}) = S\left(\left\|\theta_{rs}^{(k)}\right\|_F, \lambda_1 \eta_{rs}^{(k)} t\right) \times \left(1 - \frac{\lambda_2 w_{rs} t}{\left[\sum_{k=1}^{K} S\left(\left\|\theta_{rs}^{(k)}\right\|_F, \lambda_1 \eta_{rs}^{(k)} t\right)^2\right]^{1/2}}\right)_+ \cdot \frac{\theta_{rs}^{(k)}}{\left\|\theta_{rs}^{(k)}\right\|_F},$$

where $S(x, \lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator and $(x)_+ = \max(x, 0)$ denotes the positive part of a real function. As for non-regularized node potentials $\theta_r^{(k)}$ and $\theta_{rr}^{(k)}$, the solutions are simply

$$\text{prox}_t(\theta_r^{(k)}) = \theta_r^{(k)}, \quad \text{prox}_t(\theta_{rr}^{(k)}) = \theta_{rr}^{(k)}.$$

In our practice, we realize the above algorithm using the Matlab package TFOCS [5], which allows us to experiment with different variants of proximal gradient and accelerated proximal gradient methods. To check convergence of our algorithm, the following stopping criterion is adopted:

$$\frac{\|x_{q+1} - x_q\|_2}{\max\{1, \|x_{q+1}\|_2\}} \leq \text{tol},$$

which is a combination of absolute and relative tolerance. With a moderate tolerance of 10^{-5}, the algorithm usually converges in an acceptable number of iterations and results in satisfactory estimate of sparsity structure.

However, computation of this algorithm for large graphs can be very slow, and it cannot be easily paralleled. Therefore, the application of our method should be limited to problems with tens to hundreds of nodes. Methods which can take advantage of parallel computing such as nodewise regression may be a better option for inference of large complex networks.

4.2 Selection of Tuning Parameters

The proposed estimator reconstructs the structure of K graphs simultaneously, which is associated with two tuning parameters, namely, λ and α to control the overall sparsity and homogeneity of estimated networks. In this section, we introduce a G-fold stratified cross-validation (CV) scheme to perform tuning parameter selection. As a purely data-driven method, it is adaptive to multi-class data with different levels of homogeneity.

We apply a group-wise G-fold partition to create the training and validation sets by excluding and including each fold. Let us denote $\left(\widehat{\boldsymbol{\Theta}}_{[-g]}^{(1)}(\lambda, \alpha), \ldots, \widehat{\boldsymbol{\Theta}}_{[-g]}^{(K)}(\lambda, \alpha)\right)$ the solution to optimization problem (3) computed only by the training set with the

g-th fold excluded for each biological condition. Therefore, the cross-validation error for the g-th fold can be computed by

$$\mathrm{CV}_{[g]}(\lambda,\alpha)=\frac{1}{N}\sum_{k=1}^{K}n_k\ell\Big(\widehat{\boldsymbol{\Theta}}^{(k)}_{[-g]}(\lambda,\alpha);\boldsymbol{X}^{(k)}_{[g]}\Big),$$

which is the sum negative log-PL derived from trained solution and validation data. Taking arithmetic average of G folds, the CV score regarding a combination of tuning parameters is $\mathrm{CV}(\lambda,\alpha)=(1/G)\sum_{g=1}^{G}\mathrm{CV}_{[g]}(\lambda,\alpha)$.

In practice, we specify a few values of α to define several rough levels of homogeneity, but a relatively dense sequence of λ. Using grid search, we are able to locate the optimal combination of tuning parameters:

$$(\hat{\lambda},\hat{\alpha})=\arg\min_{\lambda,\alpha}\ \mathrm{CV}(\lambda,\alpha).$$

Due to the fact that CV is computationally costly, we can also follow an alternative line search strategy implemented in Danaher et al. [10]. To be more specific, we first fix α at the median level of the searching range, and search the best λ minimizing the CV score. Next, we use the tuned optimal $\hat{\lambda}$ for the best level of α, namely $\hat{\alpha}$. In order to further reduce false discovery rate, the "1-SE rule" can be applied to choose λ with extra penalization on sparsity. Starting from $(\hat{\lambda},\hat{\alpha})$, we fix α at $\hat{\alpha}$ and increase the regularization as much as we can, such that the CV score is still within one standard error of $\mathrm{CV}(\hat{\lambda},\hat{\alpha})$. In other words, we aim to find the maximum of λ maintaining

$$\mathrm{CV}(\lambda,\hat{\alpha})\le\mathrm{CV}(\hat{\lambda},\hat{\alpha})+\mathrm{SE}(\hat{\lambda},\hat{\alpha}),$$

where the standard error is computed by $\mathrm{SE}(\hat{\lambda},\hat{\alpha})=G^{-1/2}\cdot\mathrm{SD}\big(\mathrm{CV}_{[1]}(\hat{\lambda},\hat{\alpha}),\ldots,\mathrm{CV}_{[G]}(\hat{\lambda},\hat{\alpha})\big)$. The effectiveness of our method to select tuning parameters is validated in the following artificial data experiments.

5 Numerical Studies

5.1 Illustration of Single Graph Consistency

In this part, we illustrate the difference in edge selection consistency between our pseudo likelihood based method with the approximate inference approach proposed by Park et al. [28]. A simple binary Ising model for a cycle of four nodes is considered in this example, whose distribution is expressed by

$$\mathcal{P}(x)\propto\exp\{\theta_1x_1+\theta_2x_2+\theta_3x_3+\theta_4x_4+\theta_{12}x_1x_2+\theta_{23}x_2x_3+\theta_{34}x_3x_4+\theta_{14}x_1x_4\},x\in\{0,1\}^4.$$

Park et al. [28] introduced an approximate maximum likelihood approach for PEMRF via a Gaussian entropy bound, which has similar form to graphical group lasso. However, the performance highly depends on the real inverse covariance matrix of sufficient statistics, and Loh and Wainwright [23] showed the inverse

Table 2 Monte Carlo probabilities of correct estimation for pseudo-likelihood single class inference

n	200	500	1000	2000	5000	10000
Pr(correct recovery)	0.380	0.634	0.820	0.930	0.996	1.000

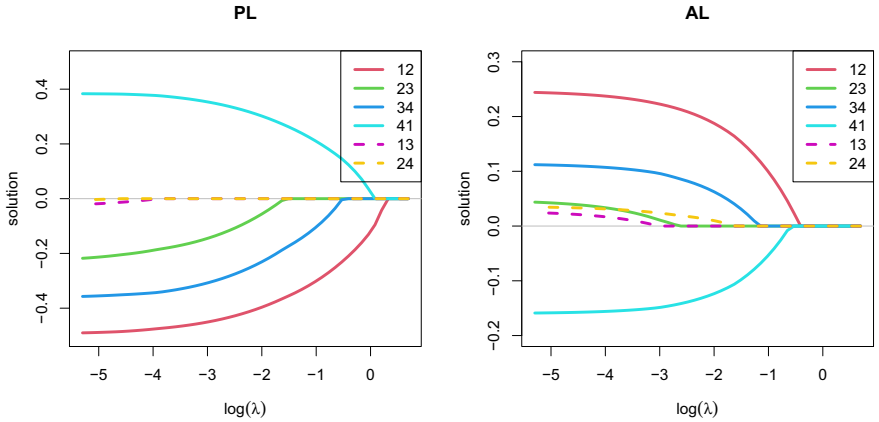

Fig. 1 Solution paths for PL- and AL-based methods with sufficient statistics standardized. The PL method can select the right model with proper λ, but the AL method cannot

covariance matrix of the above Ising model does not reflect the true sparsity of edge structure. Specifically, let $(\theta_1, \theta_2, \theta_3, \theta_4) = (0.5, 4.3, 2.3, -2.3)$, $\theta_{12} = -4.3$, $\theta_{23} = -1.3$, $\theta_{34} = -2.8$, $\theta_{14} = 3.9$, and then the inverse covariance matrix is as follows,

$$[\mathrm{Cov}(X)]^{-1} = \begin{pmatrix} 19.334 & 11.014 & 0.015 & -8.790 \\ 11.014 & 16.491 & 1.043 & -1.124 \\ 0.015 & 1.043 & 5.421 & 3.078 \\ -8.790 & -1.124 & 3.078 & 18.783 \end{pmatrix},$$

where we find that the element corresponding to (x_2, x_4) overtakes the entry for edge (x_2, x_3) in scale. Thus, Liu and Zhang [22] illustrated that consistent edge recovery could not be attained in such situation using the variational likelihood inference. However, our proposed pseudo likelihood based approach shows a different story. In this case, we fix the regularization parameter to be $\lambda = 0.01$, and then inspect the probability of completely correct structure recovery (no false positives or false negatives) given different sample sizes. The probabilities are estimated through Monte Carlo simulation built on 1000 replicates, which results in Table 2 indicating consistent result of structural learning. In addition, Fig. 1 depicts the solution paths for the two methods under $n = 10000$, where we can find their differences in edge selection consistency.

5.2 Simulations

In this part, we assess the performance of our proposed method in jointly estimating multiple pairwise exponential graphical models with high topological similarity. Let us consider $K = 2$ sub-populations composed of $p = 60$ features in our design. Each network consists of two mutually unconnected sub-networks. The first sub-network ($p_1 = 40$) is shared by both classes, but the second network ($p_2 = 20$) is separately generated for each class. All artificial sub-networks are scale-free following power law degree distributions [1]. As a result, the two true graphs are similar but distinct. Specifically, we evaluate two types of mixtures in our study. The first type of mixture (P-E) is composed of 30 truncated Poisson nodes and 30 exponential nodes. The second type of mixture (C–P–G) contains 20 categorical nodes (3-category), 20 truncated Poisson nodes and 20 Gaussian nodes. Within each mixture, all kinds of interactions are present. Model parameters are set up as follows.

- *C–P–G mixture*: $\theta_r = 0$ for categorical and Gaussian nodes; $\theta_r = 0.8$ for truncated Poisson nodes; $\theta_{rr} = -3$ for Gaussian nodes; $\theta_{rs} = \pm 0.03$ for Poisson-Poisson edges; $\theta_{rs} = \pm 0.1$ for Poisson-categorical and Poisson-Gaussian edges; $\theta_{rs} = \pm 0.3$ for other types of edges. All signs of interactions are randomly selected.
- *P–E mixture*: $\theta_r = 1$ for truncated Poisson nodes; $\theta_r = -0.5$ for exponential nodes; $\theta_{rs} = -0.2$ for all types of edges.

We adopt Gibbs sampling to generate synthetic data with per-class sample size $n = 50, 100, 200, 500$. All truncated Poisson variables are truncated at 10 during data generation. However, since truncation points are usually unknown in reality, all truncated Poisson variables are treated as Poisson as we implement our PL-based structural learning.

To elucidate the advantage of joint modeling, we compare the edge selection performance for partial joint modeling ($\alpha = 0.5$) and separate modeling ($\alpha = 0$) using our PL-based data integration (PLDIG). We also assess the difference between PLDIG and the data integration framework based on approximate likelihood (ALDIG) [22]. To compare with these joint learning methods, we also apply node-wise regression (NR) to each individual sub-population so that structures of the two networks are separately learned without data integration. The NR approach is done by the R package mgm [14], in which exponential variables are modeled as Gaussian, and categorical variables are addressed by regularized multinomial regression. The Receiver Operating Characteristic (ROC) curves averaged across 20 independent experiments using a sequence of λ are exhibited in Fig. 2, which reflect the overall prediction power at different levels of data integration. We illustrate the running times for these methods in Fig. 3, on a Mac with 2.3 GHz dual-core Intel i5 processor and 16 GB of RAM.

From Fig. 2, we can easily observe that partial data integration is beneficial to edge recovery accuracy for both mixture types. The lines representing no data integration, either PLDIG (red dashed line) or NR (blue dash-dotted line), are inferior to the others in all scenarios. As for the comparison between PLDIG and ALDIG, we

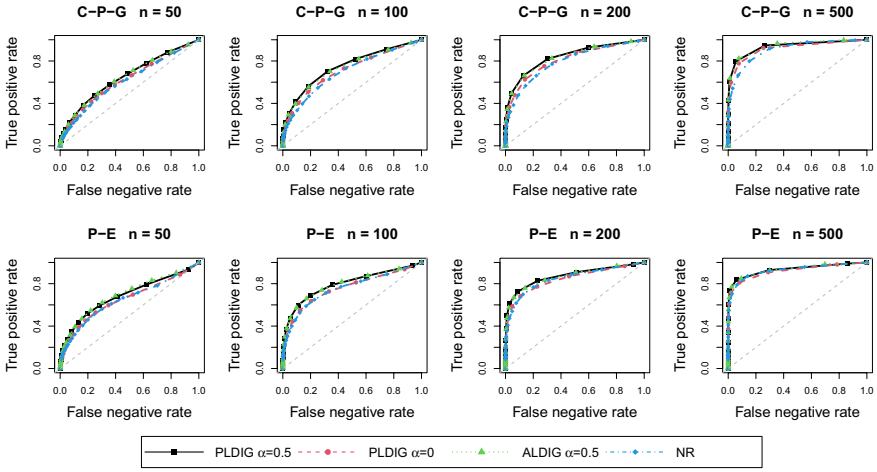

Fig. 2 ROC curves of edge selection for two types of mixtures (averaged across 20 experiments) (Color figure online)

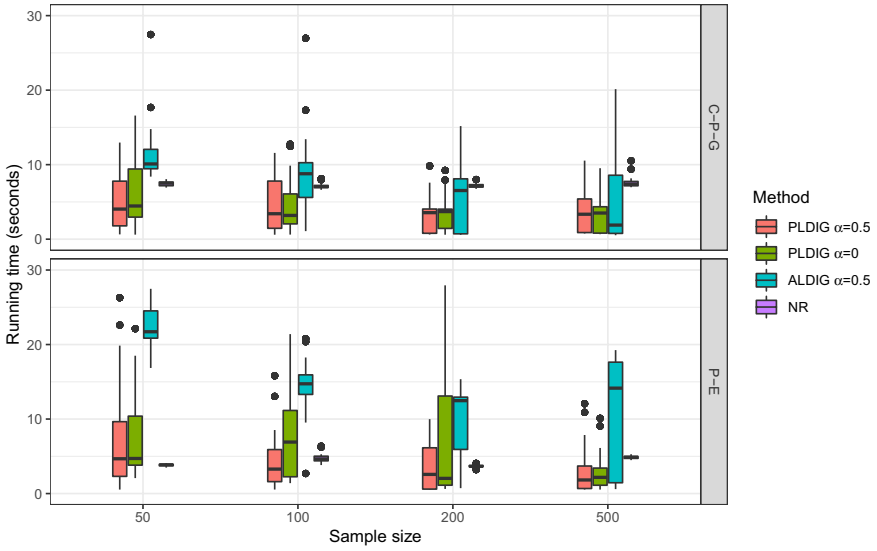

Fig. 3 Boxplots of running time using different methods

find that PLDIG is slightly better than ALDIG in most cases, given that we use the same level of joint modeling. The running time for PLDIG is comparable to ALDIG, and is slightly faster in some cases.

Afterwards, we utilize the C–P–G mixture to demonstrate how CV selects the level of joint modeling and how CV adapts to different similarity scenarios. In addition to the "similar but distinct" scenario described previously, we also take the situation that the two sub-populations are identical into consideration, where we use

the previous class 1 network to simulate data for both classes. In each of the 20 independent experiments, 5-fold CV with line search is adopted to select α from $\{0, 0.1, \ldots, 1\}$.

The boxplots of selected α are shown in Fig. 4. In general, our CV scheme automatically selects the level of network similarity. That is, higher levels of joint modeling are selected given the two sub-populations are identical. We can also observe a trend that increasing n for each class leads to lower level of joint modeling, which is beneficial in detecting class-specific correlations when the graphs are not completely the same. High level of data integration may increase the risk of selecting false positives, canceling out the benefit of joint modeling provided that the network structures are not identical.

6 Real Application to TCGA Data

Cancer is a complex group of diseases, which arise from accumulation of genetic and epigenetic factors. A number of molecular features have been found to be associated with human cancer, such as gene expression, mutation, copy number variation and DNA methylation. In this part, we base our study on the Cancer Genome Atlas (TCGA) project [38], which compiles multiple types of omic data from a large number of cancer patients. Instead of the entire cancer genome, we limit our interest to the BRAF signaling pathway [45], a smaller gene module consisting of 10 most popular cancer-related genes: NRAS, RAF1, BRAF, MAP2K1, MAP2K2, MAPK1, PIK3CA, PTEN, AKT1, MTOR, regulating cell growth, migration, and proliferation. It has been shown that overexpressions, mutations and amplifications in these genes are linked to different malignancies [9, 35].

We adopt the proposed data integration technique to jointly learn the gene regulatory networks of breast invasive carcinoma (BRCA) ($n_1 = 750$ primary tumor samples) and kidney renal clear cell carcinoma (KIRC) ($n_2 = 389$ primary tumor samples). A mixture of different types of omic features, including RNA sequencing, somatic copy number variation (CNV), non-silent somatic mutation are incorporated into our study. The raw RNA-Seq count data are processed according to the procedure in Allen and Liu [2], which adjusts for sequencing depth and overdispersion [19]. Then the processed mRNA-seq data are Poisson-like, thereby modeled as Poisson variables. Gene level mutation data are considered binary. Among all these 10 genes, only 3 of them have enough variation in both cancer types, namely, MTOR,

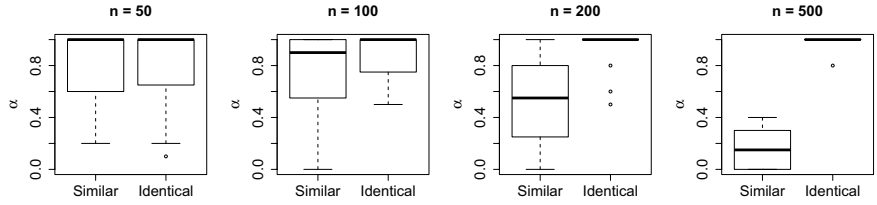

Fig. 4 Boxplots of α selected by 5-fold CV under different true network scenarios

PIK3CA and PTEN. Copy number variation (CNV), as a common type of structural variation, is a phenomenon of duplication or deletion events in the genome. The CNV data are presented as log-ratios, which are continuous and roughly symmetric. Thus, we use our proposed PLDIG approach to simultaneously learn the structures of two C–P–G mixed models.

Among $\alpha \in \{0, 0.1, \ldots, 1\}$, the 5-fold CV suggests an optimal joint modeling level at $\alpha = 0.5$, indicating moderate homogeneity among the two networks. We use bootstrap resampling to select edges getting voted by more than 95 times out of 100 bootstrapped samples, so as to find interactions with high confidence and to reduce the false discovery rate [20]. The corresponding topology estimates are demonstrated in Fig. 5.

In general, gene co-expression relationships as well as mutation-expression of CNV-expression regulatory interactions form the skeletons of the two networks. The two graphs consisting of edges with high confidence are very similar. For both graphs, we find RAF1 expression is identified as a hub node for both cancer types, which is aligned with the discovery that RAF1 and BRAF are critical upstream kinases and activators of the MAP kinase signaling pathway [37]. The CNV of RAF1 is also an important regulatory factor, which is consistent with the existing literature [30, 33]. Several breast cancer specific interactions are selected, such as the interaction between the CNV of MTOR and RAF1 expression, which is supported by the finding that PI3K/AKT/mTOR and Raf/MEK/ERK cascades are interconnected [31].

7 Conclusion and Discussion

When data are collected from different biological conditions with similar interaction mechanisms, by conducting joint analysis of multiple sub-populations, we can borrow information from different biological conditions. In this paper, we have

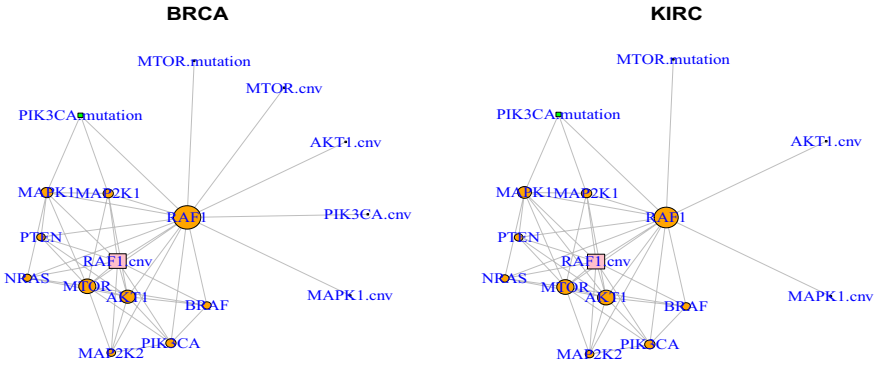

Fig. 5 Gene regulatory networks of BRCA and KIRC. Orange circles stand for gene expressions. Green squares represent point mutations. CNVs are shown in pink squares. Degrees of connectivity are indicated by relative node sizes (Color figure online)

established a joint pseudo-likelihood estimation framework named PLDIG for multiple mixed graphical models. Given that we have a limited sample size for each sub-population, but these graph structures are sparse and highly homogeneous, our proposed data integration framework shows great practical benefits.

This paper focuses on a similar joint graphical model estimation problem to Liu and Zhang [22]. However, the PL-based approach in this paper is in possession of edge selection consistency for general graph structures, whereas the approximation in ALDIG has difficulties in recovering cyclic graphs. Using a proximal gradient algorithm rather than ADMM in Liu and Zhang [22], we find that the computation burden of PLDIG is comparable to ALDIG in small to moderate graphs via simulations.

The application of our proposed method should be limited to networks with tens to hundreds of nodes due to speed restriction. For large and complex systems, a feature filtering based on existing knowledge is necessary to computational efficiency. Modules of features with high numbers of mutual correlations should be prioritized.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12561-023-09367-9.

Acknowledgements This research was partially supported by NSF grant DMS-2015481 (to YZ). We thank the anonymous reviewers for their insightful comments and suggestions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Albert R, Jeong H, Barabási A-L (200) Error and attack tolerance of complex networks. Nature 406(6794):378–382
2. Allen GI, Liu Z (2012) A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In: 2012 IEEE international conference on bioinformatics and biomedicine. IEEE, pp 1–6
3. Auslender A, Teboulle M (2006) Interior gradient and proximal methods for convex and conic optimization. SIAM J Optim 16(3):697–725
4. Beck A, Teboulle M (2009) Gradient-based algorithms with applications to signal recovery. In: Eldar Y, Palomar D (eds) Convex optimization in signal processing and communications. Cambridge University Press, Cambridge, pp 42–88
5. Becker SR, Candès EJ, Grant MC (2011) Templates for convex cone problems with applications to sparse signal recovery. Math Program Comput 3(3):165
6. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B (Methodol) 36(2):192–225
7. Chen S, Witten DM, Shojaie A (2014) Selection and estimation for mixed graphical models. Biometrika 102(1):47–64
8. Cheng J, Li T, Levina E, Zhu J (2017) High-dimensional mixed graphical models. J Comput Graph Stat 26(2):367–378
9. Corcoran RB, Dias-Santagata D, Bergethon K, Iafrate AJ, Settleman J, Engelman JA (2010) Braf gene amplification can promote acquired resistance to mek inhibitors in cancer cells harboring the BRAF V600E mutation. Sci Signal 3(149):84–84
10. Danaher P, Wang P, Witten DM (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Ser B (Methodol) 76(2):373–397

11. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
12. Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group lasso. arXiv preprint. arXiv:1001.0736
13. Guo J, Levina E, Michailidis G, Zhu J (2011) Joint estimation of multiple graphical models. Biometrika 98(1):1–15
14. Haslbeck JMB, Waldorp LJ (2020) MGM: estimating time-varying mixed graphical models in high-dimensional data. J Stat Softw 93(8):1–46. https://doi.org/10.18637/jss.v093.i08
15. Lauritzen SL (1996) Graphical models, vol 17. Clarendon Press, New York
16. Lauritzen SL, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17(1):31–57. https://doi.org/10.1214/aos/1176347003
17. Lee JD, Hastie TJ (2015) Learning the structure of mixed graphical models. J Comput Graph Stat 24(1):230–253. https://doi.org/10.1080/10618600.2014.900500
18. Lee JD, Sun Y, Taylor JE (2015) On model selection consistency of regularized M-estimators. Electron J Stat 9(1):608–642
19. Li J, Witten DM, Johnstone IM, Tibshirani R (2012) Normalization, testing, and false discovery rate estimation for RNA data. Biostatistics 13(3):523–538
20. Li S, Hsu L, Peng J, Wang P (2013) Bootstrap inference for network construction with an application to a breast cancer microarray study. Ann Appl Stat 7(1):391
21. Liu Q, Zhang Y (2020) Fast variational inference for joint mixed sparse graphical models. IEEE J Sel Areas Inf Theory 1(3):908–913. https://doi.org/10.1109/JSAIT.2020.3042124
22. Liu Q, Zhang Y (2020) Joint estimation of heterogeneous exponential markov random fields through an approximate likelihood inference. J Stat Plan Inference 209:252–266
23. Loh P-L, Wainwright MJ (2013) Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. Ann Stat 41(6):3022–3049. https://doi.org/10.1214/13-aos1162
24. Ma J, Michailidis G (2016) Joint structural estimation of multiple graphical models. J Mach Learn Res 17(1):5777–5824
25. Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34(3):1436–1462
26. Nesterov YE (1983) A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl Akad.Nauk SSSR 269:543–547
27. Nesterov Y (2013) Gradient methods for minimizing composite functions. Math Program 140(1):125–161
28. Park Y, Hallac D, Boyd SP, Leskovec J (2017) Learning the network structure of heterogeneous data via pairwise exponential markov random fields. In: AISTATS
29. Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional ising model selection using I_1-regularized logistic regression. Ann Stat 38(3):1287–1319. https://doi.org/10.1214/09-aos691
30. Ren G, Liu X, Mao X, Zhang Y, Stankiewicz E, Hylands L, Song R, Berney DM, Clark J, Cooper C (2012) Identification of frequent BRAF copy number gain and alterations of RAF genes in Chinese prostate cancer. Genes Chromosomes Cancer 51(11):1014–1023
31. Saini KS, Loi S, de Azambuja E, Metzger-Filho O, Saini ML, Ignatiadis M, Dancey JE, Piccart-Gebhart MJ (2013) Targeting the PI3K/AKT/mTOR and RAF/MEK/ERK pathways in the treatment of breast cancer. Cancer Treat Rev 39(8):935–946
32. Shaddox E, Peterson CB, Stingo FC, Hanania NA, Cruickshank-Quinn C, Kechris K, Bowler R, Vannucci M (2018) Bayesian inference of networks across multiple sample groups and data types. Biostatistics 21(3):561–576
33. Simon R, Richter J, Wagner U, Fijan A, Bruderer J, Schmid U, Ackermann D, Maurer R, Alund G, Knönagel H (2001) High-throughput tissue microarray analysis of 3p25 (RAF1) and 8p12 (FGFR1) copy number alterations in urinary bladder cancer. Cancer Res 61(11):4514–4519
34. Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. J Comput Graph Stat 22(2):231–245
35. Śmiech M, Leszczyński P, Kono H, Wardell C, Taniguchi H (2020) Emerging braf mutations in cancer progression and their possible effects on transcriptional networks. Genes 11(11):1342
36. Tansey W, Padilla OHM, Suggala AS, Ravikumar P (2015) Vector-space Markov random fields via exponential families. In: ICML

37. Varga A, Ehrenreiter K, Aschenbrenner B, Kocieniewski P, Kochanczyk M, Lipniacki T, Baccarini M (2017) RAF1/BRAF dimerization integrates the signal from RAS to ERK and ROKα. Sci Signal 10(469):8482
38. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113
39. Yang E, Allen G, Liu Z, Ravikumar PK (2012) Graphical models via generalized linear models. In: Advances in neural information processing systems, pp 1358–1366
40. Yang E, Baker Y, Ravikumar P, Allen G, Liu Z (2014) Mixed graphical models via exponential families. In: Proceedings of the 17th international conference on artificial intelligence and statistics, vol 33, pp 1042–1050
41. Yang E, Ravikumar P, Allen GI, Liu Z (2015) Graphical models via univariate exponential family distributions. J Mach Learn Res 16(1):3813–3847
42. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Methodol) 68(1):49–67
43. Yuan M, Lin Y (2007) Model selection and estimation in the Gaussian graphical model. Biometrika 94(1):19–35. https://doi.org/10.1093/biomet/asm018
44. Zhang Y, Ouyang Z, Zhao H (2017) A statistical framework for data integration through graphical models with application to cancer genomics. Ann Appl Stat 11(1):161–184
45. Zhang Y, Linder MH, Shojaie A, Ouyang Z, Shen R, Baggerly KA, Baladandayuthapani V, Zhao H (2018) Dissecting pathway disturbances using network topology and multi-platform genomics data. Stat Biosco 10(1):86–106