Electronic Journal of Statistics
Vol. 10 (2016) 1133–1154
ISSN: 1935-7524
DOI: 10.1214/16-EJS1135

Estimation of multiple networks in Gaussian mixture models

Chen Gao

Division of Biostatistics, School of Public Health, University of Minnesota

Yunzhang Zhu

Department of Statistics, Ohio State University

Xiaotong Shen

School of Statistics, University of Minnesota

and

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota
e-mail: weip@biostat.umn.edu

Abstract: We aim to estimate multiple networks in the presence of sample heterogeneity, where the independent samples (i.e. observations) may come from different and unknown populations or distributions. Specifically, we consider penalized estimation of multiple precision matrices in the framework of a Gaussian mixture model. A major innovation is to take advantage of the commonalities across the multiple precision matrices through possibly nonconvex fusion regularization, which for example makes it possible to achieve simultaneous discovery of unknown disease subtypes and detection of differential gene (dys)regulations in functional genomics. We embed in the EM algorithm one of two recently proposed methods for estimating multiple precision matrices in Gaussian graphical models. We demonstrate the feasibility and potential usefulness of the proposed methods in an application to glioblastoma subtype discovery and differential gene network analysis with a microarray gene expression data set. We also conduct realistic simulation studies to evaluate and compare the performance of various methods.

Keywords and phrases: Disease subtype discovery, Gaussian graphical model, model-based clustering, non-convex penalty, glioblastoma, gene expression.

Received March 2015.

1. Introduction

We consider the problem of estimating multiple networks in the presence of sample heterogeneity; that is, the samples come from several populations with

different Gaussian distributions, however it is unknown which samples are from which distributions. The precision matrix of each distribution corresponds to a network. This is related to but differs from the usual task of inferring and contrasting multiple networks in Gaussian graphical models, where it is known which samples are from which distributions (Guo et al. 2011; Danaher et al. 2014; Zhu et al. 2014 [11, 4, 38]). Although Gaussian mixture models are widely used for model-based clustering, our primary goal is for estimation and comparison of cluster-specific precision matrices, for which existing model-based clustering methods (McLachlan and Peel 2001; Fraley and Raftery 2006; Zhou et al. 2009 [18, 8, 37]) are not suitable. The existing model-based clustering methods either specify a common precision matrix or estimate multiple unconstrained cluster-specific precision matrices; due to the lack of a fusion penalty or other mechanisms, the cluster-specific precision matrix estimates are either exactly the same or completely different. On the other hand, in many applications one would expect both commonalities and differences among the cluster-specific precision matrices. Accounting for their commonalities not only improves statistical estimation efficiency through information borrowing, but also enhances the ability of interpretation with a focus on few possible changes across the cluster-specific precision matrices.

Our proposed methods were motivated by genomic applications to disease subtype discovery while accounting for differential gene expression and/or differential gene regulations across (unknown) disease subtypes. This is in contrast to existing methods allowing for only differential gene expression in disease subtype discovery (Verhaak et al. 2010 [33]). Arguably, a biologically more interesting problem is not only in detecting differential gene expression, but also in discovering gene dysregulations, across to-be-discovered disease subtypes, which will facilitate understanding disease mechanisms and thus developing individualized treatments.

Our approach is in the framework of multivariate Gaussian mixture modeling (McLachlan and Peel 2001 [18]). The majority of the existing literature on mixture modeling focus on regularizing only the mean parameters with diagonal covariance matrices (Pan and Shen 2007; Wang and Zhu 2008; Xie et al. 2008 [23, 34, 36]), though some (Zhou et al. 2009; Hill and Mukerjee 2013; Wu et al. 2013 [37, 12, 35]) have started considering regularization of the covariance parameters too, all of which, however, do not touch on the key issue of identifying both common and varying substructures of the precision matrices across the components of a mixture model. Since these methods always give different networks for different populations unless a common network is assumed, they do not address the question of interest here: which parts of the networks change with the populations. To address this question, we propose embedding one of the current methods of estimating multiple Gaussian graphical models (Danaher et al. 2014; Zhu et al. 2014 [4, 38]) in the EM algorithm (Dempster et al. 1977 [6]) for the Gaussian mixture model, for which existing algorithms can be effectively used in the M-step of an EM algorithm for a Gaussian mixture model.

Since these methods apply a fusion penalty to shrink multiple networks towards each other, they not only are statistically more efficient with information

borrowing, but also facilitate interpretation in identifying differential network substructures. In particular, due to the use of a non-convex penalty, the method of Zhu et al. (2014) [38] strives to uncover the commonalities among multiple networks while maintaining their unique substructures too.

Due to the connections to and differences from our current problem, we briefly review the literature on Gaussian graphical models *without sample heterogeneity*; that is, it is known that the samples come from the same Gaussian distribution. Gaussian graphical models are commonly used to describe conditional dependence relationships between interacting variables for continuous multivariate data. They are widely applied to reveal the structures in gene regulatory networks ([9, 7]), protein interaction networks ([10, 30, 15]) and brain functional connectivity ([13, 15]). Each network or graph consists of a set of nodes representing variables (e.g. genes) and edges; each edge between two nodes indicates the conditional dependency of the two nodes, given all other nodes. In Gaussian graphical models, the edges between nodes are determined by the non-zero off-diagonal elements in the precision matrix (the inverse of the covariance matrix). Therefore, reconstruction of the graph is equivalent to estimating the precision matrix in the Gaussian graphical model. Friedman et al. (2008) [10] proposed the graphical lasso method to estimate the (inverse) covariance matrices, where they provided an efficient algorithm to directly maximize the L_1-penalized log-likelihood. While the graphical lasso is fast, it only focuses on estimating a single graph. It ignores the structural similarities of multiple graphs when graphical lasso is applied to estimate each graph separately. Recent works aim to recognize possible commonalities among multiple graphs. Peterson et al. (2015) [24] proposed a Bayesian approach to estimating multiple Gaussian graphs by placing a Markov random field prior on the edges and a spike-and-slab prior to control the similarity between graphs. Qiu et al. (2015) [25] proposed a kernel method for joint estimation of multiple Gaussian graphs. Guo et al. [11] proposed to control the sparsity of the off-diagonal elements of the precision matrices and to use the L_1 penalty to control the differences between the off-diagonal elements for each pair of precision matrices. Danaher et al. (2014) [4] proposed the joint graphical lasso algorithm, which uses the L_1 penalty to regularize both the sparsity and the differences between the corresponding off-diagonal elements for each pair of precision matrices. Mohan et al. (2014) [21] extended the joint graphical lasso by taking a node-based approach for estimation of multiple Gaussian graphs. Recently, Zhu et al. (2014) [38] proposed a regularized maximum likelihood method for estimation of multiple precision matrices, In addition to seeking sparseness with a non-convex penalty to regularize the off-diagonal elements in each precision matrix, it also imposes a non-convex fusion penalty on the differences between each pair of some related precision matrices that can be flexibly specified.

The rest of this paper is organized as follows. In Section 2 we introduce our proposed new methods for estimating component-wise precision matrices in the framework of a Gaussian mixture model. Section 3 presents simulation studies to demonstrate the promising performance of our proposed methods, followed in Section 4 for an application to a glioblastoma gene expression data set. We conclude in Section 5 with a summary of our findings.

2. Methods

2.1. Gaussian mixture model

We assume that each of n iid p-dimensional observations, $x_1, x_2, \ldots, x_n$, comes from a Gaussian mixture distribution with probability density function

$$f(x_j) = \sum_{i=1}^{g} \pi_i f_i(x_j; \theta_i),$$

where g is the number of components (or populations), π_i is the prior probability for component i with $\sum_{i=1}^{g} \pi_i = 1$, $\theta_i = \{\mu_i, V_i\}$ is the set of the mean and covariance matrix parameters for cluster i, and f_i is a multivariate Normal density (with a component-specific mean μ_i and covariance matrix V_i),

$$f_i(x; \theta_i) = \frac{1}{(2\pi)^{p/2}|V_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)' V_i^{-1}(x-\mu_i)\right).$$

Since each component corresponds to a cluster, we will refer to component and cluster exchangeably. The primary goal here is to estimate the cluster-specific precision matrices $W_i = V_i^{-1}$, though identifying the clusters is often of interest either as a direct or side product.

Given the data, the log-likelihood is

$$\log L(\Theta) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i(x_j; \theta_i)\right), \tag{1}$$

where $\Theta = \{(\pi_i, \theta_i) : i = 1, 2, ..., g\}$ denotes the set of all unknown parameters. An Expectation-Maximization (EM) algorithm [6] is often used to obtain the maximum likelihood estimates. For high-dimensional data, it is often beneficial to use the maximum penalized likelihood estimator based on a penalized log-likelihood

$$\log L_P(\Theta) = \log L(\Theta) - p_\lambda(\Theta), \tag{2}$$

where $p_\lambda(\Theta)$ is to be specified as a penalty on all or a subset of the parameters. Various penalties have been proposed to achieve better performance in different contexts.

2.2. New methods

2.2.1. New method 1: With a convex penalty

A zero entry $W_{i;kl}$, the (k,l)th entry of W_i, indicates conditional independence between the kth and lth variables in cluster i given other variables. Estimating multiple cluster-specific precision matrices can reveal changes of dependency structures across multiple clusters. To facilitate detecting structural changes, a

penalty is imposed on the differences between the corresponding entries across multiple precision matrices. We propose using a joint lasso and fused graphical lasso (FGL) penalty of Danaher et al. (2014) [4] on each precision matrices W_i's:

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g} \sum_{k \neq l} |W_{i;kl}| + \lambda_2 \sum_{i<i'} \sum_{k,l} |W_{i;kl} - W_{i';kl}|, \tag{3}$$

where λ_1 and λ_2 are nonnegative tuning parameters. In addition to achieving sparseness as in graphical lasso, FGL also encourages identical entries across cluster-specific precision matrices. This feature helps to reveal both commonalities and cluster-specific network structures, in addition to improving statistical estimation efficiency through borrowing information across the multiple networks.

Note that Danaher et al. (2014) used the above penalty in the context of Gaussian graphical modeling, knowing which observations are from which Gaussian distribution, differing from our Gaussian mixture modeling. Nevertheless, we will show how to apply their proposed ADMM algorithm ([1]) (as implemented in the R package JGL) in the M-step of an EM algorithm in the current context.

We denote the new method that incorporates the use of the joint lasso and fused graphical lasso (JGL) in our Gaussian mixture modeling as **New-JGL**.

2.2.2. New method 2: With a non-convex penalty

In the context of Gaussian graphical modeling, Zhu et al. (2014) [38] proposed the following non-convex penalty function for W_i,

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g} \sum_{k \neq l} J_\tau(|W_{i;kl}|) + \lambda_2 \sum_{i<i'} \sum_{k \neq l} J_\tau(|W_{i;kl} - W_{i';kl}|), \tag{4}$$

where λ_1, λ_2 and τ are nonnegative tuning parameters, and $J_\tau(z) = \min(|z|, \tau)$ is the truncated Lasso penalty (TLP) (Shen et al. 2012 [28]). The two penalties serve the corresponding sparseness and fusion roles as in JGL. However, in contrast to FGL in (3), only non-diagonal elements, but not diagonal elements, are penalized for their differences in (4).

The non-convex TLP reduces the bias induced by the lasso penalty because no more penalty is imposed if $|z| > \tau$ in $J_\tau(z)$. In the current context, the TLP can do better in maintaining the magnitudes of non-zero entries or differences between two unequal entries. The scaled TLP, $J_\tau(z)/\tau$, approximates the L_0-function, $I(z \neq 0)$, as τ tends to 0^+. Like FGL, this method is able to detect possible element-wise heterogeneity across multiple networks, for example in identifying signaling network changes across distinct cancer subtypes.

We propose using the same non-convex penalty (4) in our current context of Gaussian mixture modeling, and will demonstrate that the algorithm of Zhu et al. (2014) can be applied in the M-step of an EM algorithm for our purpose.

We denote the new method that incorporates the use of structural pursuit (SP) penalty (4) in our Gaussian mixture modeling as **New-SP** (New-Structural-Pursuit).

2.3. Computing

We develop an EM algorithm to obtain the maximum penalized likelihood estimates (MPLEs). In particular, we will demonstrate how to use an existing Gaussian graphical modeling algorithm in the M-step of the EM algorithm for a penalized Gaussian mixture model.

We introduce z_{ij} as the indicator of whether x_j belongs to component i, so $z_{ij} = 1$ if x_j comes from component i and $z_{ij} = 0$ otherwise. Here z_{ij}'s are treated as missing data. If z_{ij}'s are observed, the complete data penalized log-likelihood is

$$\log L_{c,P}(\Theta) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}[\log \pi_i + \log f_i(x_j;\theta_i)] - p_\lambda(\Theta), \tag{5}$$

where $p_\lambda(\Theta)$ is a penalty on the parameters; typically only the mean parameters μ_i's and/or covariance matrices V_i's are penalized, which is assumed throughout.

Define the posterior probability of x_j's belonging to component i as $\rho_{ij} = P(z_{ij} = 1|x_j;\Theta)$, then the E-step calculates the following with the current estimate $\Theta^{(r)}$ at iteration r,

$$Q_P(\Theta;\Theta^{(r)}) = E_{\Theta^{(r)}}(\log L_{c,P}|X) = \sum_{i=1}^{g}\sum_{j=1}^{n} \rho_{ij}^{(r)}[\log \pi_i + \log f_i(x_j;\theta_i)] - p_\lambda(\Theta) \tag{6}$$

with

$$\hat{\rho}_{ij}^{(r)} = P(z_{ij} = 1|x_j;\Theta^{(r)}) = \frac{\hat{\pi}_i^{(r)} f_i(x_j';\theta_i^{(r)})}{\sum_{i=1}^{g}\hat{\pi}_i^{(r)} f_i(x_j';\theta_i^{(r)})}. \tag{7}$$

In the M-step, we find $\hat{\pi}_i^{(r+1)}$, $\hat{\mu}_i^{(r+1)}$ and $\hat{W}_i^{(r+1)}$ that maximize Q_P. Using the Lagrange multiplier η to constrain $\sum_{i=1}^{g}\pi_i = 1$, we omit the terms without π_i's and rewrite Q_P as

$$L(\pi,\eta) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\rho}_{ij}^{(r)} \log \pi_i + \eta(\sum_{i=1}^{g}\pi_i - 1). \tag{8}$$

Taking the partial derivative of $L(\pi,\eta)$ with respect to π_i and set it to 0, we arrive at the updating formula for $\hat{\pi}_i$

$$\hat{\pi}_i^{(r+1)} = \sum_{j=1}^{n} \hat{\rho}_{ij}^{(r)}/n. \tag{9}$$

To update μ_i, if there is no penalty on μ_i, we take the derivative of Q_P with respect to μ_i and set it to 0,

$$\frac{\partial Q_P}{\partial \mu_i} = \sum_j^n \hat{\rho}_{ij}^{(r)}(x_j - \mu_i)'\hat{W}_i = 0, \tag{10}$$

obtaining the updating formula for $\hat{\mu}_i$ as

$$\hat{\mu}_i^{(r+1)} = \frac{\sum_{j=1}^n \hat{\rho}_{ij}^{(r)} x_j}{\sum_{j=1}^n \hat{\rho}_{ij}^{(r)}}. \tag{11}$$

On the other hand, if the Lasso penalty is imposed on μ_i, then its updating formula involves a soft-thresholding on the above quantity (e.g., Pan and Shen 2007 [23]).

Finally, to update V_i or equivalently, $W_i = V_i^{-1}$, we only need to consider the terms related to W_i in Q_P:

$$\begin{aligned} Q_P &= \frac{1}{2}\sum_{i=1}^g\sum_{j=1}^n \hat{\rho}_{ij}^{(r)} \log|W_i| - \frac{1}{2}\sum_{i=1}^g\sum_{j=1}^n \hat{\rho}_{ij}^{(r)}(x_j - \hat{\mu}_i^{(r)})'W_i(x_j - \hat{\mu}_i^{(r)}) - p_\lambda(\Theta) \\ &= \frac{1}{2}\sum_{i=1}^g\sum_{j=1}^n \hat{\rho}_{ij}^{(r)} \left(\log|W_i| - tr(\tilde{S}_i^{(r)}W_i)\right) - p_\lambda(\Theta) \end{aligned} \tag{12}$$

with

$$\tilde{S}_i^{(r)} = \frac{\sum_{j=1}^n \hat{\rho}_{ij}^{(r)}(x_j - \hat{\mu}_i^{(r)})(x_j - \hat{\mu}_i^{(r)})'}{\sum_{j=1}^n \hat{\rho}_{ij}^{(r)}} \tag{13}$$

as a weighted sample covariance matrix.

Typically there is no closed-form solution to update W_i or V_i when one of them is penalized. However, we can take advantage of the existing methods for penalized Gaussian graphical models. Below we point out their connection.

If we know the cluster label for each observation x_j, as in Gaussian graphical modeling, then the penalized log-likelihood for W_i is

$$\frac{1}{2}\sum_{i=1}^g \left[n_i\left(\log|W_i| - tr(S_iW_i)\right) - p_\lambda(W_i)\right], \tag{14}$$

where n_i is the sample size for cluster i, and S_i is the sample covariance matrix for cluster i. Correspondingly, in the current context of Gaussian mixture modeling, the Q_P function in the EM algorithm with a penalty on W_i is

$$Q_P = \frac{1}{2}\sum_{i=1}^g \left[\sum_{j=1}^n \hat{\rho}_{ij}^{(r)} \left(\log|W_i| - tr(\tilde{S}_i^{(r)}W_i)\right) - p_\lambda(W_i)\right]. \tag{15}$$

To maximize Q_P, we use the soft assignment, instead of hard assignment, of each observation x_j into a cluster. Specifically, setting $n_i = \sum_{j=1}^n \hat{\rho}_{ij}^{(r)}$ and $S_i = \tilde{S}_i^{(r)}$,

then maximizing expression (15) will be equivalent to maximizing (14). Since there are already efficient computational algorithms to maximize the penalized log-likelihood (14) in the Gaussian graphical model, we can incorporate one of them into the M-step in our EM algorithm to obtain an update for W_i. Zhou et al. (2009) used this idea in applying graphical Lasso (Friedman et al. 2008 [10]) in their penalized model-based clustering with unconstrained covariance matrices. We applied the R functions of Danaher et al. (2014) and Zhu et al. (2014) in the M-step for the proposed two new methods respectively.

2.4. Review: two existing methods

Different choice of $p_\lambda(W_i)$ will lead to different penalized maximum likelihood estimates of W_i and corresponding algorithms. For comparison, we briefly review two existing penalized mixture modeling methods (Pan and Shen (2007); Zhou et al. (2009) [23, 37]). The method of Pan and Shen (2007) specifies each component in the Gaussian mixture model as a multivariate normal with a common diagonal covariance matrix $V_i = V = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. They proposed an L_1-penalty for the mean parameters,

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g} \sum_{k=1}^{p} |\mu_{ik}|, \tag{16}$$

where μ_{ik} is the mean of kth variable for component i. Using the L_1 penalty, small estimates of the mean parameters will be shrunken to be exactly zero. If for a given variable k, $\mu_{ik} = 0$ for all components i, then this variable will have no effect on clustering. Hence this penalty is used for variable selection, but not for inferring the cluster-specific networks.

Zhou et al. (2009) [37] relaxed the diagonal covariance matrix assumption and adopted unconstrained covariance/precision matrices. To regularize the parameters in the precision matrices, they proposed a penalty function of the form

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g} \sum_{k=1}^{p} |\mu_{ik}| + \lambda_2 \sum_{k \neq l} |W_{kl}|, \tag{17}$$

where $W = V^{-1}$ is the common precision matrix (or inverse covariance matrix). The first term in the above penalty function aims at variable selection as in Pan and Shen (2007), while the second term uses the L_1-penalty to promote the sparseness of the precision matrix. For penalized covariance matrix estimation, they used the graphical lasso algorithm of Friedman et al. (2008) and maximized the following objective function

$$\log |W| - tr(\tilde{S}^{(r)} W) - \lambda \sum_{k \neq l} |W_{kl}|, \tag{18}$$

where $\lambda = 2\lambda_2/n$ and

$$\tilde{S}^{(r)} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{\rho}_{ij}^{(r)} (x_j - \hat{\mu}_i^{(r)})(x_j - \hat{\mu}_i^{(r)})'}{n}$$

is a weighted sample covariance matrix based on the soft assignments of all the samples as for $\tilde{S}_i^{(r)}$.

Zhou et al. (2009) also considered the case where each component i in the mixture model has an unconstrained covariance matrix V_i. Then they proposed the following penalty function to regularize the means and cluster-specific covariance matrices,

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g} \sum_{k=1}^{p} |\mu_{ik}| + \lambda_2 \sum_{i=1}^{g} \sum_{j,l}^{p} |W_{i;j,l}|, \tag{19}$$

where $W_{i;j,l}$ is the (j,l)th entry of W_i. They again used the graphical lasso algorithm to obtain the estimate of the cluster-specific precision matrix $W_i = V_i^{-1}$ in the M-step of the EM algorithm.

2.5. Implementation

By default our EM algorithm starts with some initial values given by the K-means method, though other (random or fixed) and/or multiple starting values can be equally applied.

We first use the L_1-penalty, then try τ at each of the quantiles of the L_1-penalized estimates of $|W_{i;kl}|$ and $|W_{i;kl} - W_{i';kl}|$. By default the tuning parameters λ_1 and λ_2 are chosen from $\lambda_1 \in \{\log(p) \times (1.5, 1, 0.8, 0.3, 0.1, 0.05, 0.01, 0.001)\}$ and $\lambda_2 \in \{\log(p) \times (10^8, 1000, 500, 100, 50, 10, 5, 1, 0.8, 0.5, 0.3, 0.1, 0.01, 0.001)\}$. A grid search is used to find a combination of the penalty parameter values (λ_1, λ_2, τ) and a cluster number g that lead to the highest predictive log-likelihood as calculated by 5-fold cross-validation.

Our methods are implemented in an R package called pGMM that will be freely downloadable on CRAN.

3. Simulations

Due to the unknown truth for real data, it is difficult to draw definitive conclusions on the relative performance of various methods. As an alternative, we conducted simulations to evaluate and compare the performance of the methods in both clustering (i.e. the assignments of the samples to clusters) and precision matrix estimation.

To mimic real data, we used the fitted models to the glioblastoma gene expression data by Zhou et al. (2009) and our proposed new methods as the true model to generate simulated data; in this way, we avoided possible biases in using only one true model to generate simulated data that might favor one of the methods. In each case, there were 4 clusters with $n = 173$ or $n = 346$ observations with $p = 20$. We then applied the usual non-penalized model-based clustering as implemented in R package mclust (Fraley and Raftery 2006 [8]), the methods of Pan and Shen (2007) [23] and Zhou et al. (2009) [37], and our proposed two new methods. To measure the accuracy of parameter estimation for

precision matrices, we used the average entropy loss (EL) and average quadratic loss (QL),

$$EL = \frac{1}{g}\sum_{i=1}^{g}\left(tr(V_i\hat{W}_i) - \log\det(V_i\hat{W}_i)\right)$$
$$QL = \frac{1}{g}\sum_{i=1}^{g} tr\left((V_i\hat{W}_i - I)^2\right).$$

To measure the accuracy of estimating zero or non-zero entries and grouping structures in precision matrices, following Zhu et al. (2014) [38], we used the average number of false positives for sparseness pursuit (FPV), average number of false negatives for sparseness pursuit (FNV), average number of false positives for grouping (FPG), and average number of false negatives for grouping (FNG):

$$FPV = \frac{1}{g}\sum_{i=1}^{g}\frac{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}=0,\hat{W}_{i;jj'}\neq 0)}{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}=0)}\times\left(1-\mathbb{I}(W_{i,\text{off}}\neq 0)\right)$$
$$FNV = \frac{1}{g}\sum_{i=1}^{g}\frac{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}\neq 0,\hat{W}_{i;jj'}=0)}{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}\neq 0)}\mathbb{I}(W_{i,\text{off}}\neq 0)$$
$$FPG = \frac{1}{C(g,2)}\sum_{i<i'}\frac{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}=W_{i';jj'},\hat{W}_{i;jj'}\neq\hat{W}_{i';jj'})}{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}=W_{i';jj'})}$$
$$\times\left(1-\mathbb{I}(W_{i,\text{off}}\neq W_{i',\text{off}})\right)$$
$$FNG = \frac{1}{C(g,2)}\sum_{i<i'}\frac{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}\neq W_{i';jj'},\hat{W}_{i;jj'}=\hat{W}_{i';jj'})}{\sum_{1\le j\le j'\le K}\mathbb{I}(W_{i;jj'}\neq W_{i';jj'})}$$
$$\times\mathbb{I}(W_{i,\text{off}}\neq W_{i',\text{off}}),$$

where $C(g,2)$ is the combinatorial number of choosing 2 from g.

Table 1 shows the results for $n = 173$ based on 50 simulations for each set-up. With the true model as the fitted model by the method of Zhou et al. (2009), the method of Zhou et al. (2009) itself gave the highest Rand index, suggesting the best accuracy for clustering. However, It did not give the lowest average entropy loss (EL) and quadratic loss (QL) for precision matrix estimation, though the differences were not large. Recall that the true model here was based on four largely differing cluster-specific precision matrices, which might not favor fusing the cluster-specific precision matrices. Impressively our method New-SP gave the second highest Rand index that was quite close to that of Zhou et al. (2009), and more importantly, New-SP gave the most or second most accurate estimates of the cluster-specific precision matrices with the lowest average EL and second lowest QL. In addition, it also gave low false positive rates of sparseness and grouping, but high false negative rates. It is noted that New-JGL also performed well.

TABLE 1

Simulation results with $n = 173$ and the true model being that estimated byone of the three methods based on the glioblastoma dataset. The means (standard deviations) of the Rand Index (RI), adjusted Rand Index (aRI), average entropy loss (EL) average quadratic loss (QL), average false positive for sparseness pursuit (FPV), average false negative for sparseness pursuit (FNV), average false positive for grouping (FPG) and average false negative for grouping (FNG) are shown for 50 simulations.

Truth	Method	RI	aRI	EL	QL	FPV	FNV	FPG	FNG
Zhou09	Zhou09	**0.714**	**0.309**	30.495	61.121	0.749	0.204	0.875	0.110
		(0.036)	(0.087)	(1.330)	(21.484)	(0.026)	(0.021)	(0.081)	(0.084)
	Pan07	0.648	0.164	29.569	66.893	0.000	1.000	0.000	1.000
		(0.033)	(0.054)	(0.208)	(4.349)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.632	0.159	30.054	70.533	0.000	1.000	0.000	1.000
		(0.027)	(0.044)	(0.620)	(11.465)	(0.000)	(0.000)	(0.000)	(0.000)
	New-JGL	0.660	0.189	29.049	**59.005**	0.127	0.817	0.000	1.000
		(0.026)	(0.054)	(0.673)	(11.801)	(0.174)	(0.182)	(0.000)	(0.000)
	New-SP	0.689	0.260	**28.726**	59.695	0.020	0.952	0.022	0.957
		(0.034)	(0.071)	(0.881)	(11.112)	(0.042)	(0.086)	(0.043)	(0.075)
New-SP	Zhou09	0.632	0.024	31.609	61.775	0.649	0.256	0.881	0.000
		(0.024)	(0.010)	(1.138)	(17.245)	(0.047)	(0.023)	(0.058)	(0.000)
	Pan07	0.804	0.501	33.785	49.081	0.000	1.000	0.000	0.000
		(0.043)	(0.107)	(0.612)	(8.787)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.862	0.647	26.179	72.099	1.000	0.000	0.319	0.000
		(0.047)	(0.114)	(5.333)	(209.858)	(0.000)	(0.000)	(0.470)	(0.000)
	New-JGL	0.856	0.629	23.853	11.255	0.376	0.292	0.000	0.000
		(0.038)	(0.098)	(0.362)	(2.275)	(0.124)	(0.069)	(0.000)	(0.000)
	New-SP	**0.917**	**0.785**	**23.265**	**10.242**	0.130	0.558	0.000	0.000
		(0.053)	(0.136)	(0.977)	(4.657)	(0.072)	(0.106)	(0.000)	(0.000)
New-JGL	Zhou09	0.664	0.063	30.174	46.403	0.692	0.245	0.911	0.090
		(0.027)	(0.017)	(0.769)	(8.676)	(0.029)	(0.018)	(0.037)	(0.040)
	Pan07	0.886	0.717	30.283	31.125	0.000	1.000	0.000	1.000
		(0.036)	(0.087)	(0.624)	(5.356)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.897	0.744	25.875	24.760	0.020	0.980	0.000	1.000
		(0.030)	(0.072)	(0.392)	(3.434)	(0.141)	(0.141)	(0.000)	(0.000)
	New-JGL	0.926	0.815	**22.533**	**7.615**	0.361	0.341	0.000	1.000
		(0.043	(0.109)	(0.258)	(1.237)	(0.047)	(0.048)	(0.000)	(0.000)
	New-SP	**0.930**	**0.823**	23.493	12.694	0.056	0.759	0.000	1.000
		(0.042)	(0.107)	(0.786)	(5.237)	(0.049)	(0.159)	(0.000)	(0.000)

On the other hand, if the true model was the fitted one by New-SP, then New-SP was the clear winner for both clustering and precision matrix estimation, followed by New-JGL. This was the case when the cluster-specific precision matrices differed but sharing some commonalities. Finally, if the true model was the fitted one from New-JGL, the winners were New-JGL and New-SP, followed by mclust and the method of Pan and Shen (2007) (where a common diagonal precision matrix was assumed).

We also investigated the sensitivity of the EM algorithm to its starting values. For the set-up with the true model as the one fitted by New-JGL, instead of using the K-means output as the starting value for New-JGL and New-SP, we used some randomly generated numbers as the starting value. The resulting Rand index values for New-JGL and New-SP decreased from 0.926 and 0.930 to 0.635 and 0.757 respectively, confirming the importance of using good starting

TABLE 2

Simulation results with $n = 346$ and the true model being that estimated byone of the three methods based on the glioblastoma dataset. The means (standard deviations) of the Rand Index (RI), adjusted Rand Index (aRI), average entropy loss (EL) average quadratic loss (QL), average false positive for sparseness pursuit (FPV), average false negative for sparseness pursuit (FNV), average false positive for grouping (FPG) and average false negative for grouping (FNG) are shown for 50 simulations.

Truth	Method	RI	aRI	EL	QL	FPV	FNV	FPG	FNG
Zhou09	Zhou09	0.775	0.455	26.072	24.435	0.825	0.115	0.936	0.049
		(0.040)	(0.097)	(1.185)	(12.397)	(0.021)	(0.022)	(0.065)	(0.062)
	Pan07	0.608	0.123	29.191	56.655	0.000	1.000	0.000	1.000
		(0.036)	(0.049)	(0.140)	(2.808)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.675	0.239	29.647	72.100	0.000	1.000	0.000	1.000
		(0.040)	(0.079)	(1.543)	(28.488)	(0.000)	(0.000)	(0.000)	(0.000)
	New-JGL	0.661	0.200	27.316	34.188	0.583	0.329	0.007	0.992
		(0.022)	(0.043)	(0.356)	(4.787)	(0.028)	(0.033)	(0.050)	(0.057)
	New-SP	**0.842**	**0.624**	**24.502**	**17.738**	0.128	0.605	0.206	0.541
		(0.052)	(0.129)	(1.163)	(7.946)	(0.032)	(0.067)	(0.054)	(0.074)
New-SP	Zhou09	0.893	0.726	29.645	42.974	0.743	0.156	0.956	0.000
		(0.032)	(0.083)	(0.587)	(7.665)	(0.041)	(0.016)	(0.033)	(0.000)
	Pan07	0.830	0.570	34.003	49.082	0.000	1.000	0.000	0.000
		(0.025)	(0.061)	(0.354)	(5.454)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.922	0.802	24.722	23.761	1.000	0.000	0.950	0.000
		(0.037)	(0.094)	(1.704)	(18.492)	(0.000)	(0.000)	(0.198)	(0.000)
	New-JGL	0.883	0.701	22.864	7.019	0.426	0.185	0.000	0.000
		(0.037)	(0.095)	(0.234)	(1.33)8	(0.103)	(0.040)	(0.000)	(0.000)
	New-SP	**0.962**	**0.902**	**21.035**	**2.534**	0.544	0.106	0.000	0.000
		(0.018)	(0.048)	(0.318)	(0.827)	(0.230)	(0.116)	(0.001)	(0.000)
New-JGL	Zhou09	0.931	0.827	28.631	35.451	0.788	0.153	0.963	0.036
		(0.025)	(0.063)	(0.618)	(6.236)	(0.022)	(0.013)	(0.007)	(0.011)
	Pan07	0.898	0.747	30.255	30.215	0.000	1.000	0.000	1.000
		(0.023)	(0.056)	(0.507)	(3.997)	(0.000)	(0.000)	(0.000)	(0.000)
	Mclust	0.941	0.853	22.358	7.122	1.000	0.000	0.275	0.725
		(0.021)	(0.052)	(0.610)	(2.899)	(0.002)	(0.000)	(0.446)	(0.446)
	New-JGL	0.938	0.845	21.844	4.899	0.458	0.216	0.104	0.869
		(0.024)	(0.060)	(0.130)	(0.650)	(0.069)	(0.055)	(0.241)	(0.304)
	New-SP	**0.961**	**0.903**	**21.637**	**4.359**	0.109	0.540	0.000	0.998
		(0.011)	(0.028)	(0.171)	(0.698)	(0.116)	0.105)	(0.001)	(0.006)

values for the EM algorithm. However, the estimation errors for the precision matrices were less influenced: for example, the mean EL for New-JGL and New-SP increased from 22.533 and 23.493 to only 23.964 and 25.743, respectively, still lower than those of the other methods.

Next we doubled the sample size in simulations. With the increased sample size, the proposed method New-SP became the clear overall winner, followed by New-JGL (Table 2). Although mclust performed well in the last two set-ups (with the true model being that fitted by New-SP or New-JGL), it did not work well in the first set-up. Again it is noted that the two new methods largely outperformed the method of Zhou et al. (2009) [37] for estimating the cluster-specific precision matrices, perhaps due to the former two's use of the fusion penalties for information borrowing across multiple cluster-specific precision matrices.

4. Example

4.1. Glioblastoma gene expression data

Verhaak et al. (2010) [33] studied a gene expression data set of glioblastoma tumor and normal samples. They used a consensus hierarchical clustering method to identify four disease subtypes. It is noted that, due to the limitation of the clustering method, the conditional dependencies between genes in each cluster were ignored and thus not revealed. This leaves room for our new and other methods to explore possible dependency relationships among the genes. Furthermore, the identified four clusters, albeit biologically reasonable, are in no way to be perfect, which bears importance when one uses their sample assignments as a reference to compare various methods.

To be practically focused, we restricted our analysis to the gene expression data from the 173 core samples as used by Verhaak et al.(2010) [33], and we selected only 20 genes that are related to cell signaling pathways. Some of these genes were demonstrated to be altered in Figure 4B in Brennan et al. (2013) [2] and Figure 5 in Mclendon et al. (2008) [19]. Specifically, genes EGFR, PDGFRA, FGFR3 are members of the RTK signaling pathway. RASGRP3 and RRAS are downstream targets of the RTK signaling pathway. PIK3C2B, PIK3R1, PIK3R3, PIK3IP1 and AKTIP are components of the PI3K/AKT signaling pathway. NFIB is the downstream target of RTK and PI3K/AKT signaling pathways. CDKN3, CDK4, CDKN1A, CDKN2C, CCND2 are involved in RB signaling pathways and they play important roles in cell cycle regulation. CASP1 and CASP4 are important genes in cell apoptosis.

4.2. Estimated networks

We applied our method New-SP to the glioblastoma gene expression data set. Trying with $g = 1, 2, 3, 4, 5$ clusters, it reached four clusters/subtypes. Each cluster showed a distinct conditional dependency structure among the genes, though their overall structures were similar (Figure 1).

This suggests distinct cell signalling network changes across the disease subtypes. A closer examination of the estimated precision matrices reveals that the conditional dependencies among the receptor kinases and the downstream target genes were altered. The PI3K/Akt signaling pathway plays an important role in cell survival and proliferation in glioblastoma ([3, 22]). One of the estimated networks shows that the AKTIP gene was conditionally correlated with CDKN2C, a gene encoding a cyclin-dependent kinase inhibitor that regulates cell growth. However, this link was lost in all other three estimated networks. Similarly, PIK3IP and AKTIP were not conditionally correlated with CDKN1A, CDKN2C and CDKN3 in one or more estimated networks, while the network in bottom left of Figure 1 preserved most of the connections. The PI3K/Akt signaling pathway is reported to be upstream of CCND2, a gene encoding the cell cycle regulating protein Cyclin D2 ([20]). Only one out of four subtypes demonstrated a conditional dependence between AKTIP and CCND2. The changes

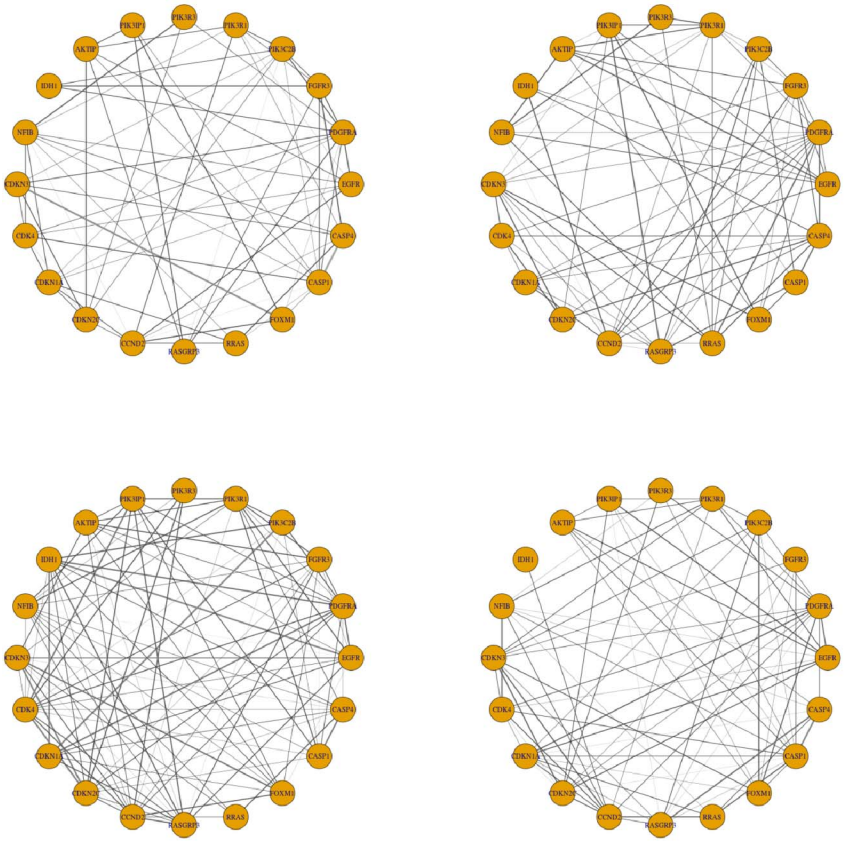

FIG 1. *Estimated cluster-specific networks based on 173 core samples using the new method New-SP.*

between these links collectively suggested dysregulation of cell growth by the PI3K/Akt signaling pathway in some subtypes of glioblastoma.

Gene IDH1 is known to have a higher mutation frequency in some glioblastoma subtypes, and here it exhibited cluster-specific associations with FGFR3, which was also reported to have mutations in glioblastoma subtypes classified by Verhaak et al. (2010) [33]. We found that gene IDH1's expression was positively correlated with that of FGFR3 in only one cluster, suggesting possibly altered co-expressions in other clusters. IDH1 mutation is reported to cause widespread changes in histone and DNA methylation and potentially promoting tumorigenesis ([32, 16]). CCND2 was found to be amplified in IDH1 mutant medulloblastoma subtypes ([29]). Therefore, the abnormal IDH1 gene level and its disconnection with CCND2 observed in the estimated network pointed to possible roles of IDH1 in oncogenesis in certain subtypes of glioblastoma.

For comparison, we applied Zhou et al.'s method to the glioblastoma gene expression data set with cluster-specific covariance matrices. Among $g = 1, 2, 3, 4, 5$

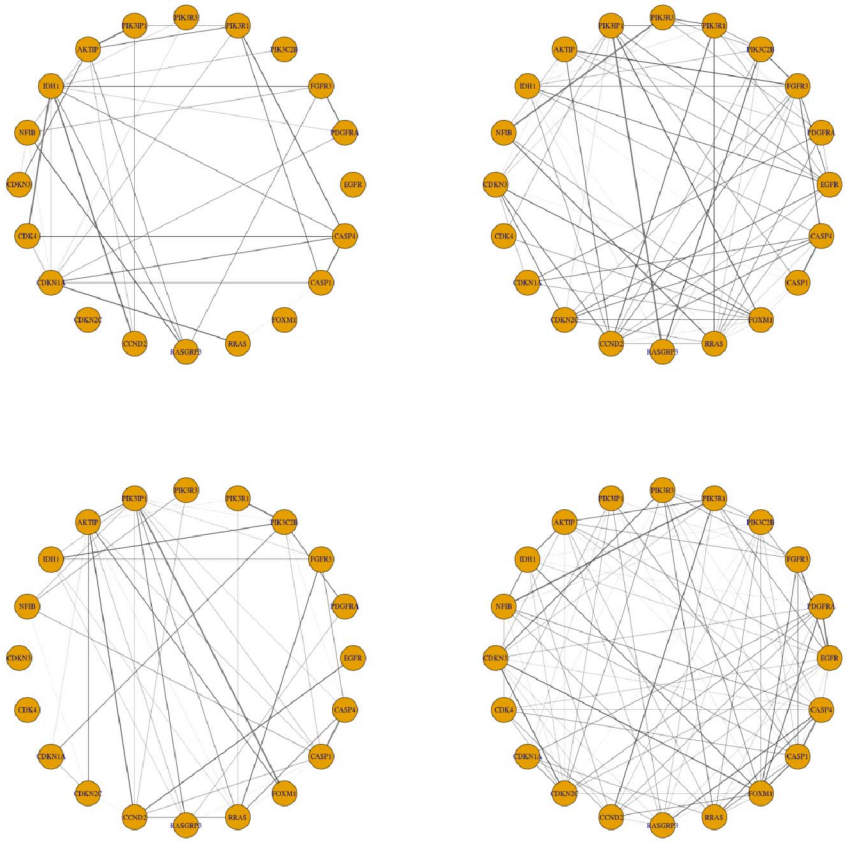

FIG 2. *Estimated cluster-specific networks based on 173 core samples using the method of Zhou et al. (2009).*

clusters, it selected four clusters. The estimated cluster-specific precision matrices demonstrated cluster-specific dependencies among the genes (Figure 2). The estimated networks using Zhou et al.'s method confirmed that the conditional correlation between IDH1 and CCND2 was lost in one network estimated by the New-SP method. The conditional correlation between AKTIP and CCND2 was present in three subtypes, though the correlation was weak in one subtype. Compared to the networks estimated by the method of New-SP, the dependency changes across the clusters estimated by Zhou et al.'s method were much more dramatic, reflecting possibly large variations of the estimates without borrowing information across clusters.

The New-JGL method also yielded four clusters (Figure 3). Like the networks estimated by the New-SP method, the networks estimated by the New-JGL method shared some structural similarity. The AKTIP and CCND2 correlation was found in two out of four subtypes, although the correlation in one subtype was weak. This agreed with the correlation in the networks estimated by

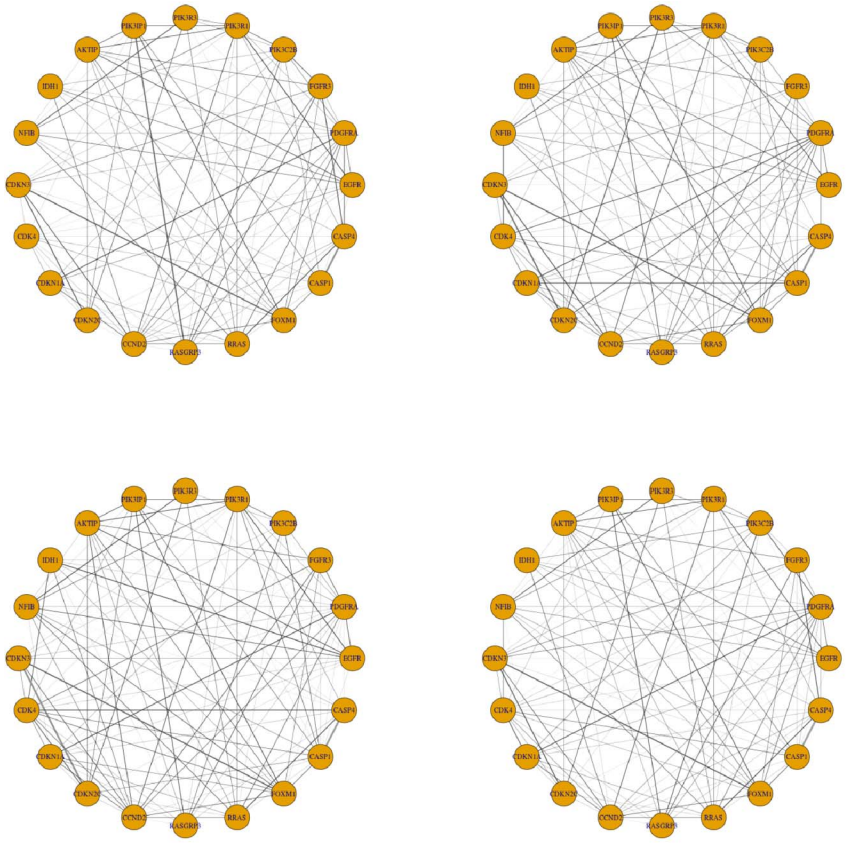

FIG 3. *Estimated cluster-specific networks based on 173 core samples using the method of New-JGL.*

the New-SP method. Unlike in Figure 1, the conditional correlation between IDH2 and CCND2 was present in all four subtypes, though the magnitude of correlation was small.

The non-penalized mclust yielded four clusters with a common covariance matrix, suggesting that the differences among the four cluster-specific covariance matrices were possibly subtle. This was also reflected from the overall similarity across the four cluster-specific estimates of the two new methods.

4.3. Sample cluster assignments

Using the cluster assignments in Verhaak et al. (2010) [33] as the reference, the agreement between our New-SP method and the reference as measured by the Rand index was 0.747 and by the adjusted Rand index was 0.354 (Table 3). Its performance was compared with several other methods. First, the method of Pan and Shen (2007) [23] based on a common diagonal covariance matrix

TABLE 3
Rand Index (RI) and adjusted Rand Index (aRI) for the glioblastoma gene expression data with 20 genes by various methods. The class assignments given in [33] are used as the reference.

		New-SP	mclust	Pan07	Zhou09	New-JGL
$p = 20$	RI	0.747	0.688	0.780	0.713	0.746
	aRI	0.354	0.222	0.439	0.305	0.355

yielded 5 clusters with a Rand index of 0.749 and the adjusted Rand index of 0.358. For the purpose of comparison, we also examined the clustering results of the method by forcing 4 clusters, which led to a Rand index of 0.780 and the adjusted Rand index of 0.439 (Table 3). Although the method of Pan and Shen yielded a slightly higher Rand index, it was possibly due to the bias of the reference clustering method (that ignored varying within-cluster dependencies that would in turn favor the results of Pan and Shen (2007)). More importantly, a common diagonal covariance matrix assumed and estimated by the method cannot be used to examine possibly varying within-cluster dependency structures. Finally, the two new methods seemed to perform better than the two other methods.

4.4. Model assessment

To check the goodness-of-fit of a final model, we propose using the parametric bootstrap, which was used by McLachlan and others to select the number of components in a Gaussian mixture model ([17, 26]). For example, for our real data, the New-SP method selected a final model with four components, each with a component-specific precision matrix, which is called an alternative model here; it may be of interest to compare this alternative model with a null (or reduced) model with four components but a common precision matrix, which could be achieved by forcing a large λ_2 value (while other tuning parameters were selected as before). We generated 50 bootstrap samples from the fitted null and alternative models respectively, then fitted the two models respectively to the bootstrap samples; finally, we compared their corresponding CV log-likelihood values, as shown in Figure 4. For the bootstrap samples, in both cases fitting the alternative model seemed to yield a higher mean value of the CV log-likelihood; however, the difference between the two fitted models was larger when the bootstrap samples were generated from the alternative model, as expected. Since the CV log-likelihood value difference between the two fitted models based on the original data was larger than that from the bootstrap samples generated from the alternative model, there was some evidence to support the use of the alternative model. Nevertheless, perhaps due to the relatively small sample sizes and shrinkage effects of the four component-specific precision matrices towards each other (as imposed by the fusion penalty even in the alternative model), the difference between the two models was not overwhelming, and cautions must be taken in not over-interpreting their differences.

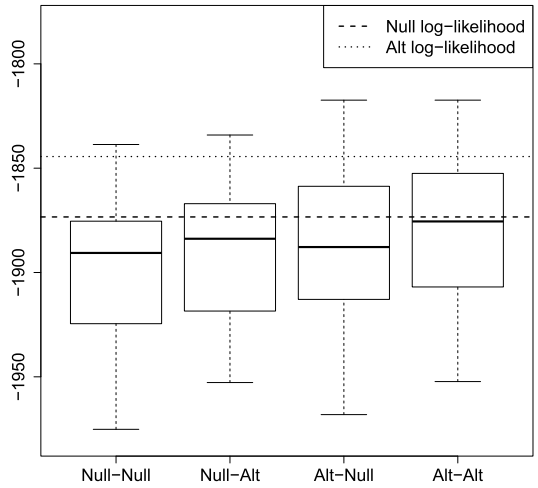

FIG 4. *Distributions of the CV log-likelihood values of various fitted models based on bootstrap samples. Null-Null, bootstrap samples were generated from the null model, to which the null model was fitted; Null-Alt, bootstrap samples were generated from the null model, to which the alternative model was fitted; Alt-Null, bootstrap samples were generated from the alternative model, to which the null model was fitted; Alt-Alt, bootstrap samples were generated from the alternative model, to which the alternative model was fitted. The two horizontal lines are the CV log-likelihood values for the two fitted models to the original data.*

5. Discussion

We have presented a new approach to estimation of multiple networks in the context of a penalized Gaussian mixture model. The primary goal is for estimating and comparing cluster-specific network changes, though automatic cluster discovery is often of interest too. For the primary goal, it is necessary to encourage the equalities of the entries across the cluster-specific precision matrices while maintaining their differences if any, which is best accomplished by fusion with a non-convex penalty such as TLP as adopted in our proposed method New-SP ([28], [38]). Note that standard and existing penalized model-based clustering methods are not suitable for our primary goal: due to the lack of fusion penalties, the existing methods cannot highlight few major differneces across multiple precision matrix estimates, in addition to their loss of estimation efficiency without information borrowing. Both our proposed methods pursue both sparseness and fusion for multiple precision matrices in the framework of Gaussian mixture modeling. Our approach takes advantage of the existing methods using convex or non-convex penalties to regularize the parameters in the unconstrained precision matrices based on Gaussian graphical models, which assumes that it is known that which samples are from which Gaussian distributions, differing from our current context with unknown sample heterogeneity.

We applied the methods to a real data set containing gene expression profiles of glioblastoma patients. Using the New-SP method, the samples were parti-

tioned into four disease subtypes, as reported in Verhaak et al. (2010) but based on only differential gene expression. Importantly, our method reconstructed disease subtype-specific gene networks, suggesting candidates for possibly subtype-specific gene dysregulations that can be followed up in further biological experiments. Since the truth is unknown for the real data, we recoursed to realistic simulations mimicking the real data to evaluate the methods; it was demonstrated that our method New-SP based on the non-convex TLP gave the best overall performance in both clustering (i.e. subtype discovery) and network estimation when the sample size was at least moderately large, followed by the other proposed method New-JGL based on the convex (fused) Lasso penalty. The better performance of New-SP over New-JGL is likely due to the non-convex TLP adopted in the former, as demonstrated in Shen et al. (2012) [28] for regression and single precision matrix estimation and Zhu et al. (2014) [38] for estimating multiple precision matrices in Gaussian graphical models. On the other hand, New-JGL is simpler and faster than New-SP, and thus can be used for larger problems and/or to provide a quick preliminary solution; in particular, we advocate using the results of New-JGL (or any other method with a convex penalty) as a good starting value for New-SP, thus the latter can be regarded as a refinement of the former. We also note that partition rules discussed in Zhu et al. (2014) can be used to speed up the new methods for high-dimensional data.

We emphasize that the existing methods for estimation of multiple networks, including the two used here (Danaher et al. 2014; Zhu et al. 2014 [4, 38]), are based on Gaussian graphical models without sample heterogeneity; that is, each sample is assumed to be known from a given Gaussian distribution. In our target applications and other settings, this sample homogeneity assumption may not hold. For example, in clinical genomic studies, due to disease heterogeneity, the assumption that all the gene expression profiles of cancer patients come from the same Gaussian distribution is not practical. To discover unknown disease subtypes, clustering or unsupervised learning becomes useful, which will facilitate personalized medicine. To our knowledge, existing clustering methods of gene expression have focused on detecting differential mean expression levels across clusters or disease subtypes, as demonstrated in Verhaak et al. (2010) [33]. However, in addition to differential gene expression, there are possibly differential gene regulations or dysregulations across disease subtypes. If disease subtypes are known, then differential gene regulations can be treated as estimating multiple precision matrices in Gaussian graphical models, as handled by many existing methods; otherwise, as discussed here, both disease subtypes and possibly differential precision matrices must be inferred simultaneously based on a Gaussian mixture model.

Our methods are implemented in an R package pGMM that will be available on CRAN.

Acknowledgment

The authors are grateful to the Editor and reviewers for constructive comments. This research was supported by NIH grants R01GM081535, R01GM113250,

R01HL105397 and R01HL116720, and by the Minnesota Supercomputing Institute.

References

[1] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**(1), 1–122.

[2] Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., *et al.* (2013). The somatic genomic landscape of glioblastoma. *Cell*, **155**(2), 462–477.

[3] Cantley, L. C. and Neel, B. G. (1999). New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway. *Proceedings of the National Academy of Sciences*, **96**(8), 4240–4245.

[4] Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, **76**(2), 373–397. MR3164871

[5] de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., and Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**(1), 497.

[6] Dempster, A. P., Laird, N. M., Rubin, D. B., *et al.* (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38. MR0501537

[7] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**(1), 196–212. MR2064941

[8] Fraley, C. and Raftery, A.E. (2006). MCLUST version 3 for R: normal mixture modeling and model-based clustering. Technical Report no. 504, Department of Statistics, University of Washington.

[9] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **305**(5659), 799–805.

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

[11] Guo, J., Levina, E., Michailidis, G., Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, **98**, 1–15. MR2804206

[12] Hill, S.M., and Mukherjee, S. (2013). Network-based clustering with mixtures of L1-penalized Gaussian graphical models: an empirical investigation. http://arxiv.org/abs/1301.2194.

[13] Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., Reiman, E. and Alzheimer's Disease NeuroImaging Initiative (2010). Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *Neuroimage*, **50**(3), 935–949.

[14] Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, **38**(3), 283–293.

[15] Kolar, M., Liu, H. and Xing, E. P. (2014). Graph estimation from multi-attribute data. *Journal of Machine Learning Research*, **15**(1), 1713–1750. MR3225245

[16] Liu, X. and Ling, Z. Q. (2015). Role of isocitrate dehydrogenase 1/2 (IDH 1/2) gene mutations in human tumors *Histology and Histopathology*, **30**(10), 1155–1160.

[17] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 318–324.

[18] McLachlan, G., and Peel, D. (2001). *Finite Mixture Models*, Wiley. MR1789474

[19] McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., *et al.* (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216), 1061–1068.

[20] Mirzaa, G., Parry, D. A., Fry, A. E., Giamanco, K. A., Schwartzentruber, J., Vanstone, M., Logan, C. V., Roberts, N., Johnson, C. A., Singh, S. and Kholmanskikh, S. S. (2014). *De novo* CCND2 mutations leading to stabilization of cyclin D2 cause megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome. *Nature Genetics*, **46**(5), 510.

[21] Mohan, K., London, P., Fazel, M., Witten, D., and Lee, S. I. (2014). Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, **15**(1), 445–488. MR3190845

[22] Narita, Y., Nagane, M., Mishima, K., Huang, H. S., Furnari, F. B. and Cavenee, W. K. (2002). Mutant epidermal growth factor receptor signaling down-regulates p27 through activation of the phosphatidylinositol 3-kinase/Akt pathway in glioblastomas. *Cancer Research*, **62**(22), 6764–6769.

[23] Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**, 1145–1164.

[24] Peterson, C., Stingo, F. C. and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, **110**(509), 159–174. MR3338494

[25] Qiu, H., Han, F., Liu, H. and Caffo, B. (2015). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**(2), 487–504.

[26] Reynolds, J. H. and Templin, W. D. (2004). Comparing mixture estimates by parametric bootstrapping likelihood ratios. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**(1), 57–74.

[27] Rozenblatt-Rosen, O., Mosonego-Ornan, E., Sadot, E., Madar-Shapiro, L., Sheinin, Y., Ginsberg, D., and Yayon, A. (2002). Induction of chondrocyte growth arrest by fgf: transcriptional and cytoskeletal alterations. *Journal of Cell Science*, **115**(3), 553–562.

[28] Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, **107**(497), 223–232. MR2949354

[29] Snuderl, M., Triscott, J., Northcott, P. A., Shih, H. A., Kong, E., Robinson, H., Dunn, S. E., Iafrate, A. J. and Yip, S. (2015). Deep sequencing identifies IDH1 R132S mutation in adult medulloblastoma. *Journal of Clinical Oncology*, **33**(6), 27–31.

[30] Telesca, D., Müller, P., Kornblau, S. M., Suchard, M. A. and Ji, Y., 2012. (2012). Modeling protein expression and protein signaling pathways. *Journal of the American Statistical Association*, **107**(500), 1372–1384. MR3036401

[31] Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**(19), 2405–2412.

[32] Turkalp, Z., Karamchandani, J. and Das, S. (2014). IDH mutation in glioma: new insights and promises for the future. *JAMA neurology*. **71**(10), 1319–1325.

[33] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., *et al.* (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*. *Cancer Cell*, **17**(1), 98–110.

[34] Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, **64**, 440–448. MR2432414

[35] Wu M-Y, Dai D-Q, Zhang X-F, Zhu Y (2013). Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm. *PLoS ONE*, **8**(6), e66256.

[36] Xie, B., Pan, W., Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, **2**, 168–212. MR2386092

[37] Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, **3**, 1473. MR2578834

[38] Zhu, Y., Shen, X., and Pan, W. (2014). Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, **109**, 1683–1696. MR3293620