Electronic Journal of Statistics
Vol. 3 (2009) 1473–1496
ISSN: 1935-7524
DOI: 10.1214/09-EJS487

Penalized model-based clustering with unconstrained covariance matrices

Hui Zhou and Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota
e-mail: zhoux292@umn.edu; weip@biostat.umn.edu
url: www.biostat.umn.edu/∼weip

Xiaotong Shen

School of Statistics, University of Minnesota
e-mail: xshen@stat.umn.edu

Abstract: Clustering is one of the most useful tools for high-dimensional analysis, e.g., for microarray data. It becomes challenging in presence of a large number of noise variables, which may mask underlying clustering structures. Therefore, noise removal through variable selection is necessary. One effective way is regularization for simultaneous parameter estimation and variable selection in model-based clustering. However, existing methods focus on regularizing the mean parameters representing centers of clusters, ignoring dependencies among variables within clusters, leading to incorrect orientations or shapes of the resulting clusters. In this article, we propose a regularized Gaussian mixture model with general covariance matrices, taking various dependencies into account. At the same time, this approach shrinks the means and covariance matrices, achieving better clustering and variable selection. To overcome one technical challenge in estimating possibly large covariance matrices, we derive an E-M algorithm to utilize the graphical lasso (Friedman et al. 2007) for parameter estimation. Numerical examples, including applications to microarray gene expression data, demonstrate the utility of the proposed method.

AMS 2000 subject classifications: Primary 62H30.
Keywords and phrases: Covariance estimation, EM algorithm, Gaussian graphical models, high-dimension but low-sample size, L_1 penalization, normal mixtures, penalized likelihood, semi-supervised learning.

Received September 2009.

Contents

*Correspondence author.

1. Introduction

As an important tool of data analysis, clustering has emerged as indispensable to analyzing high-dimensional genomic data. For example, in gene function discovery, various methods of clustering have been applied to group genes with their expressions across multiple conditions (Eisen et al. 1998 [10]; Tavazoie et al. 1999 [40]); in disease subtype discovery, clustering is used to cluster patients' tissue samples with their genomic expressions (Golub et al. 1999 [1]). In this process, because of unknown identities of many relevant genes and/or experimental conditions it is necessary to select informative genes or conditions to yield meaningful clusters. Such a task of variable selection is critical not only to clustering but also to other modeling strategies such as classification. For classification, Alaiya et al. (2002) [1] studied borderline ovarian tumor classification, where classification using all 1584 protein spots is unsatisfactory but that focusing on a subset of around 200 selected spots provided more accurate results. For clustering, there have been only a limited number of studies, in contrast to a large body of literature on variable selection for classification and regression. As pointed out in Pan and Shen (2007) [31], clustering imposes many unique challenges to variable selection in that some well known model selection procedures, e.g. best subset selection with BIC (Schwarz 1978 [37]), may not be applicable to clustering, which is unlike in classification and regression. One main reason is that in general there are many true models in clustering, most of which are not useful. For example, any true noise variable may suggest the existence of only one cluster; however, this discovery, albeit true, is useless because, in clustering one would like to uncover the underlying heterogeneity and structures in the

data, such as identifying informative variables that suggest the existence of two or more clusters.

Among many clustering approaches, model-based clustering has become increasingly important due to its interpretability. In addition to good empirical performance relative to its competitors (Thalamuthu et al., 2008 [41]), model-based clustering has a solid probabilistic framework of mixture models, which facilitates model building and checking, such as selecting the number of clusters (McLachlan, 1987 [26]; Fraley and Raftery, 1998 [13]). Although Normal mixture models have been extensively studied by both frequentist and Bayesian approaches (Banfield and Raftery, 1993 [4]; Muller et al., 1996 [29], and references therein), our focus here is on high-dimensional data, for which variable selection is necessary (e.g. Pan and Shen, 2007 [31]). There are basically two categories of approaches to variable selection for high-dimensional model-based clustering: the Bayesian approaches (Liu et al., 2003 [25]; Teh et al., 2004 [42]; Hoff, 2006 [16]; Tadesse et al., 2005 [39]; Kim et al., 2006 [20]; Raftery and Dean, 2006 [33]) and penalized likelihood approaches (Pan and Shen, 2007 [31]; Xie et al., 2008a [50]; Xie et al., 2008b [51]; Wang and Zhu, 2008 [48]; Guo et al. 2009 [15]). In general, the Bayesian approaches are more flexible by allowing more general covariance matrices, but computationally are more demanding due to the use of MCMC for stochastic searches. For the penalized likelihood approaches, one common assumption is that each cluster has a diagonal covariance matrix, implying the same orientation for all clusters (Banfield and Raftery, 1993 [4]). As to be shown later, this is too stringent and can severely degrade performance in practice. Conceptually a general or unconstrained covariance matrix should be allowed for each cluster; however the challenge is how to treat means and general covariances subject to the constraint that any resulting covariance matrix estimate is positive definite. This challenge is evident in the recent literature on Gaussian graphical modeling that estimates a large covariance matrix based on a Normal sample (Huang et al., 2006 [17]; Yuan and Lin, 2007 [53]; Levina et al., 2008 [22]; Rothman et al., 2009 [35]; Fan et al., 2009 [11] and references therein). This problem continues to be even more challenging for mixture models, because it is unknown which observations are from which Normal components.

In this article, we propose a general penalized likelihood approach that permits unconstrained covariance matrices in a Normal mixture model. A major innovation here is the recognition of the connection between fitting Normal mixture models and Gaussian graphical modeling. Our approach utilizes the recent development in Gaussian graphical modeling by effectively embedding an existing penalized covariance estimation method into the E-M algorithm for Normal mixture models. In particular, we implement our method using the graphical lasso method (Friedman et al., 2007 [12]) for covariance estimation. Moreover, we generalize the proposed method to semisupervised learning, permitting partially labeled observations.

The rest of this article is organized as follows. Section 2 reviews the penalized model-based clustering method with diagonal covariance matrices, followed by a description of our proposed method that allows for a common or cluster-

specific general covariance matrices. A brief discussion of an extension to semi-supervised learning is given to permit known cluster memberships for a subset of observations. Section 3 presents simulation results, and an application to real microarray data is contained in section 4, to demonstrate the feasibility and effectiveness of the method compared against its counterpart with diagonal covariance matrices. Section 5 concludes with a summary and a discussion of future work.

2. Methods

2.1. Mixture model and its penalized likelihood

Denote by $X = \{x_1, \ldots, x_n\}$ a random sample of n K-dimensional observations. Assume that the n observations are standardized with sample mean 0 and sample variance 1 for each variable. Assume that the observations are independent and from a mixture distribution with probability density function (pdf)

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

where f_i is the pdf for component or cluster i with unknown parameter vector θ_i, and π_i is the prior probability for component i with $\sum_{i=1}^{g} \pi_i = 1$. For parameter estimation, we adopt the maximum penalized likelihood estimator (MPLE) that maximizes the penalized log-likelihood

$$\log L_P(\Theta) = \log L(\Theta) - p_\lambda(\Theta) = \sum_{j=1}^{n} \log \left[\sum_{i=1}^{g} \pi_i f_i(x_j; \theta_i) \right] - p_\lambda(\Theta), \tag{2}$$

where Θ represents all unknown parameters and $p_\lambda(\Theta)$ is a penalty function for Θ with a regularization parameter (vector) λ. In what follows, we assume a Gaussian mixture model with f_i being a multivariate Normal density, and regularize their mean vectors and possibly covariance matrices.

To obtain the MPLE, we employ the E-M algorithm (Dempster et al., 1977 [8]). Let z_{ij} denote the indicator of whether x_j belongs to component i; namely, $z_{ij} = 1$ if x_j comes from component i, and $z_{ij} = 0$ otherwise. Here z_{ij}'s are regarded as missing data simply because they are not observed. If z_{ij}'s were observed, the complete data penalized log-likelihood becomes

$$\log L_{c,P}(\Theta) = \sum_i \sum_j z_{ij} [\log \pi_i + \log f_i(x_j; \theta_i)] - p_\lambda(\Theta). \tag{3}$$

Given a current estimate $\Theta^{(r)}$ at iteration r, the E-step of the E-M calculates

$$Q_P(\Theta; \Theta^{(r)}) = E_{\Theta^{(r)}}(\log L_{c,P} | Data) = \sum_i \sum_j \tau_{ij}^{(r)} [\log \pi_i + \log f_i(x_j; \theta_i)] - p_\lambda(\Theta), \tag{4}$$

where τ_{ij} is the posterior probability of x_j's belonging to component i. The M-step maximizes Q_P to update the estimate of Θ.

2.2. Penalized clustering with diagonal covariance matrices

For comparison, we briefly review the method of Pan and Shen (2007) [31], which specifies the components f_i as multivariate Normal with a common diagonal covariance matrix $V = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$,

$$f_i(x;\theta_i) = \frac{1}{(2\pi)^{K/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{\mathrm{t}} V^{-1}(x-\mu_i)\right)$$

with $|V| = det(V) = \prod_{k=1}^{K} \sigma_k^2$. They proposed a penalty function $p_\lambda(\Theta)$ as an L_1-norm of the mean parameters

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g}\sum_{k=1}^{K} |\mu_{ik}|, \tag{5}$$

where μ_{ik} is the mean of kth variable for component i. Note that the observations are standardized to have zero mean and unit variance for each variable k. If $\mu_{1k} = ... = \mu_{gk} = 0$, then variable k cannot differentiate the components, hence deemed as noninformative (i.e. a noise variable) and automatically excluded from clustering. Variable selection is realized when small estimates of μ_{ik}'s can be shrunken to be exactly 0 by the use of the L_1 penalty (Tibshirani, 1996 [43]).

For convenience, a generic notation $\Theta^{(r)}$ is used to represent the parameter estimate at iteration r. The E-M updating formulas for maximizing the penalized likelihood (2) are as follows: at iteration $r+1$, the posterior probability of x_j belonging to component i is

$$\hat{\tau}_{ij}^{(r)} = \frac{\hat{\pi}_i^{(r)} f_i(x_j;\hat{\theta}_i^{(r)})}{f(x_j;\hat{\Theta}^{(r)})} = \frac{\hat{\pi}_i^{(r)} f_i(x_j;\hat{\theta}_i^{(r)})}{\sum_{i=1}^{g} \hat{\pi}_i^{(r)} f_i(x_j;\hat{\theta}_i^{(r)})}, \tag{6}$$

the prior probability of an observation from the i^{th} component

$$\hat{\pi}_i^{(r+1)} = \sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}/n, \tag{7}$$

the variance of variable k

$$\hat{\sigma}_k^{2,(r+1)} = \sum_{k=1}^{K}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}(x_{jk} - \hat{\mu}_{ik}^{(r)})^2/n, \tag{8}$$

and the mean of variable k in cluster i

$$\hat{\mu}_{ik}^{(r+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}} \left(1 - \frac{\lambda_1 \hat{\sigma}_k^{2,(r)}}{|\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}|}\right)_+, \tag{9}$$

for $i = 1, 2, \ldots, g$ and $k = 1, 2, \ldots, K$. For sufficiently large λ_1, we have $\hat{\mu}_{ik} = 0$; if $\hat{\mu}_{1k} = \hat{\mu}_{2k} = ... = \hat{\mu}_{gk} = 0$ for variable k, variable k is a noise variable and does not contribute to clustering as can be seen from equation (6).

If the variance parameters are not regularized, it is straightforward to extend (6)–(9) to the case with cluster-specific diagonal covariance matrices $V_i = diag(\sigma_{i,1}^2, \ldots, \sigma_{i,1}^2)$: all the updating formulas remain to be the same except replacing σ_k^2 by $\sigma_{i,k}^2$:

$$\hat{\sigma}_{i,k}^{2,(r+1)} = \frac{\sum_{j=1}^n \hat{\tau}_{i,j}^{(r)}(x_{j,k} - \hat{\mu}_{i,k}^{(r)})^2}{\sum_{j=1}^n \hat{\tau}_{i,j}^{(r)}}. \tag{10}$$

Note that the treatment here differs from Xie et al. (2008b) [51], in which the variance parameters are regularized. Throughout this article, we assume that an informative variable is defined to have cluster-specific means, no matter whether it has a common or cluster-specific variances.

2.3. Penalized clustering with a common unconstrained covariance

We now consider a general or unconstrained covariance matrix V by relaxing the diagonal covariance matrix assumption. Denote $W = V^{-1}$ the inverse covariance matrix (or precision matrix) with elements W_{kl}.

To realize variable selection, we require that a noise variable has a common mean across clusters. Since the data have been standardized to have mean 0 for each variable, a common mean implies $\mu_{1k} = ... = \mu_{gk} = 0$. As in Bayesian approaches (Tadesse et al. 2005 [39]), one can assume that any noise variable is uncorrelated with any informative variable, though this assumption is not necessary in our approach (because this assumption does not influence our estimation procedure). To facilitate estimating large and sparse covariance matrices, we propose the following penalty function:

$$p_\lambda(\Theta) = \lambda_1 \sum_{i=1}^{g}\sum_{k=1}^{K} |\mu_{ik}| + \lambda_2 \sum_{k=1}^{K}\sum_{l=1}^{K} |W_{kl}|. \tag{11}$$

Note that, the penalty on the mean parameter is mainly for variable selection, while that for the precision matrix is necessary for high-dimensional data. Since the data dimension K is larger than the sample size n, the sample covariance matrix (or the maximum likelihood estimate under the Normality) is necessarily singular. In addition, as discussed in the literature of Gaussian graphical modeling, penalization on a large covariance (or precision) matrix can yield a better estimate than the non-penalized one. Although various penalties have been proposed for a covariance (or precision) matrix, some do not yield a positive-definite covariance estimate. For the problem considered here, since we need to calculate the log-likelihood, and thus the determinant of a covariance estimate, the positive-definiteness of a covariance estimate is needed, which imposes a major technical difficulty. In Gaussian graphical modeling, one aims to estimate the covariance or precision matrix of a Normal distribution; since all the observations are known to be iid from the same Normal distribution, the problem is easier than that for mixture models, where we need to cluster the observations

into various unknown groups (each corresponding to a Normal distribution) and estimate the covariance matrix (and other parameters) for each group simultaneously. A major contribution of our work is recognition of the connection between Gaussian mixture modeling and Gaussian graphical modeling: in spite of the unknown cluster- or group-memberships of the observations in a Gaussian mixture model, the estimation of a covariance (or precision) matrix for a mixture component can be formulated to be similar to that for Gaussian graphical modeling.

2.3.1. Estimation of the non-covariance parameters

For the EM algorithm, the E-step yields Q_P as given in (4) and the M-step maximizes Q_P with respect to the unknown parameters, resulting in the same updating formulas for τ_{ij} and π_i as given in (6) and (7). The updating formula for μ_{ij} is derived from the following theorem.

Theorem 1. *The sufficient and necessary conditions for $\hat{\mu}_{ik}$ to be a (global) maximizer of Q_P (for a fixed i and k) are*

$$\sum_{j=1}^{n} \tau_{ij} \left(x_j^t W_{\cdot k}\right) - \left(\sum_{j=1}^{n} \tau_{ij}\right) \hat{\mu}_i^t W_k = \lambda_1 sign(\hat{\mu}_{ik}), \quad \text{if } \hat{\mu}_{ik} \neq 0, \tag{12}$$

and

$$\left|\sum_{j=1}^{n} \tau_{ij} \left(\sum_{s=1, s\neq k}^{K} \left(x_{js} - \hat{\mu}_{is}\right) W_{sk} + x_{jk} W_{kk}\right)\right| \leq \lambda_1, \quad \text{if } \hat{\mu}_{ik} = 0, \tag{13}$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{iK})^t$ and $W_{\cdot k} = (W_{1k}, \ldots, W_{Kk})^t$.

Hence, we have the below updating formula for the mean parameter:

$$\text{if } \left|\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} \left(\sum_{s=1, s\neq k}^{K} \left(x_{js} - \hat{\mu}_{is}^{(r)}\right) W_{sk} + x_{jk} W_{kk}\right)\right| \leq \lambda_1, \text{ then } \hat{\mu}_{ik}^{(r+1)} = 0, \tag{14}$$

otherwise,

$$\begin{aligned} &\left(\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}\right) \hat{\mu}_{ik}^{(r+1)} W_{kk} + \lambda_1 \text{sign}(\hat{\mu}_{ik}^{(r+1)}) \\ &= \sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} \left(x_j^{\text{t}} W_{\cdot k}\right) - \left(\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}\right) \left(\hat{\mu}_i^{(r)^{\text{t}}} W_{\cdot k} - \hat{\mu}_{ik}^{(r)} W_{kk}\right). \end{aligned} \tag{15}$$

Simple algebra indicates that the updating formulas (14)–(15) for μ_{ik} reduces to (9) when the covariance matrix is diagonal. The coordinate-wise updating

for μ as above converges to the global maximum in view of the results of Tseng (1988) [45] and Tseng (2001) [46], because the first term of Q_P, the conditional expectation of the complete data log-likelihood, is concave, while the second term, the L_1 penalty on the mean parameters, is separable (and concave) in μ_{ik}'s.

It remains to derive an updating formula for the covariance matrix in the E-M algorithm.

2.3.2. Estimation of the inverse covariance matrix

To derive the estimate of the covariance matrix V, we focus on the M-step of (4) with respect to the V. Replacing V with W^{-1}, we only need to find the updating formula for W. To maximize Q_p with respect to W, it suffices to maximize

$$\begin{aligned} & \frac{n}{2}\log(det(W)) - \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}(x_j-\mu_i)^{\mathrm{t}}W(x_j-\mu_i) - \lambda_2\sum_{j,l}|W_{jl}| \\ = \quad & \frac{n}{2}\log(det(W)) - \frac{n}{2}tr\big(\tilde{S}W\big) - \lambda_2\sum_{j,l}|W_{jl}|, \end{aligned} \tag{16}$$

where

$$\tilde{S} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}(x_j-\mu_i)^{\mathrm{t}}(x_j-\mu_i)}{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}(x_j-\mu_i)^{\mathrm{t}}(x_j-\mu_i)}{n}$$

is the empirical covariance matrix.

For (16), we shall use the graphical lasso algorithm of Friedman et al. (2007) [12], to maximize an objective function

$$\log(det(W)) - tr(SW) - \lambda\sum_{k=1}^{K}\sum_{l=1}^{K}|W_{kl}|$$

over all non-negative definite matrices W for a known covariance matrix S. Hence, we can apply the algorithm to maximize (16) with $\lambda = 2\lambda_2/n$ and $S = \tilde{S}$. Their algorithm is implemented in R package *glasso*.

2.4. Penalized clustering with cluster-specific covariance matrices

To permit varying cluster volumes and orientations, we now consider component-specific unconstrained covariance matrices V_i, $i = 1, \ldots, g$. We employ a slightly modified penalty:

$$p_\lambda(\Theta) = \lambda_1\sum_{i=1}^{g}\sum_{k=1}^{K}|\mu_{ik}| + \lambda_2\sum_{i=1}^{g}\sum_{j,l}|W_{i;j,l}|. \tag{17}$$

In the E-M algorithm, the updating formulas for π and τ remain the same as in (6) and (7), respectively; for μ, we only need to replace W in (14) and (15) with W_i. Thus, we only need to consider the covariance matrix estimation. Note

$$\begin{aligned} Q_p &= C + \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}\log det(W_i) - \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n}(x_j-\mu_i)^{\mathrm{t}}W_i(x_j-\mu_i) \\ &\quad - \lambda_2\sum_{i=1}^{g}\sum_{j,l}|W_{i;j,l}|, \end{aligned}$$

where C stands for a constant term unrelated to W_i. To maximize Q_p, we only need to maximize

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}\log det(W_i) - \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(r)}(x_j-\mu_i)^{\mathrm{t}}W_i(x_j-\mu_i) - \lambda_2\sum_{i=1}^{g}\sum_{j,l}|W_{i;j,l}| \\ &= \sum_{i=1}^{g}\left(\frac{1}{2}\sum_{j=1}^{n}\tau_{ij}^{(r)}\log det(W_i) - \frac{\sum_{j=1}^{n}\tau_{ij}^{(r)}}{2}tr\left(\tilde{S}_iW_i\right) - \lambda_2\sum_{j,l}|W_{i;j,l}|\right) \end{aligned}$$

with

$$\tilde{S}_i = \frac{\sum_{j=1}^{n}\tau_{ij}^{(r)}(x_j-\mu_i)^{\mathrm{t}}(x_j-\mu_i)}{\sum_{j=1}^{n}\tau_{ij}^{(r)}}.$$

Hence, we can separately maximize each of these g terms using the graphical lasso to obtain an updated estimate of W_i.

2.5. Model selection

We propose using the predictive log-likelihood based on an independent tuning dataset or cross-validation (CV) as our model selection criterion. Through this criterion, we use a grid search to estimate the optimal $(g, \lambda_1, \lambda_2)$ as the one with the maximum predictive log-likelihood.

For any given $(g, \lambda_1, \lambda_2)$, because of possibly many local maxima for the mixture model, we run the EM algorithm multiple times with random starts. For our numerical examples, we started with the K-means clustering, and used its result as initial parameter estimates for the E-M algorithm. From the multiple runs, we selected the one giving the maximal penalized log-likelihood as the final result for the given $(g, \lambda_1, \lambda_2)$.

2.6. Extension: semi-supervised learning

We further extend the proposed method to mixture model-based semi-supervised learning, in which some observations have known cluster labels while the others do not (McLachlan and Peel, 2002 [28]; Liang et al., 2007 [23]). Pan et al.

(2006) [32] developed the penalized mixture approach with diagonal covariance matrices; here we push it to the case with unconstrained covariance matrices.

Without loss of generality, assume that we have partially labeled K-dimensional data $x_1, \ldots, x_n$, in which the first n_0 observations do not have class labels while the remaining n_1 do have. Furthermore, assume the existence of g_0 classes among the first n_0 observations and g_1 known classes among the n_1 labeled observations. The log-likelihood is

$$\log L(\Theta) = \sum_{j=1}^{n_0} \log \left[\sum_{i=1}^{g} \pi_i f_i(x_j; \theta_i) \right] + \sum_{j=n_0+1}^{n} \log \left[\sum_{i=1}^{g} z_{ij} f_i(x_j; \theta_i) \right].$$

The penalized log-likelihood can be then constructed with a suitable penalty function $p_\lambda(\Theta)$. It is noted that, in the E-M algorithm, because z_{ij}'s for the last n_1 observations are indeed observed, their corresponding "posterior probabilities" are known as $\tau_{ij} = z_{ij}$, while those for the first n_0 observations are the same as (6). With the new updating formula for τ, the updating formulas for μ, π and covariance parameters in the E-M algorithm are the same as before.

Model selection can be performed as before. First, we use an independent tuning dataset or CV, including both labeled and unlabeled observations, to select the number of clusters and penalty parameters. Then we fit the selected model to the whole data set. Note that, after obtaining all parameter estimates, we calculate the posterior probabilities τ for each observation, including the n_1 observations with known class labels, and use the posterior probabilities for class assignment.

3. Simulations

3.1. Small K: Why use non-diagonal covariance matrices

To better visualize the advantage of allowing unconstrained covariance matrices, we applied the methods to two-dimensional data. We considered two simple set-ups: the first with only one true cluster while the second with two clusters. The number of observations in each set-up was 200; for set-up 2, 100 observations were generated from each cluster. We used an independent tuning dataset of an equal size as that of a training dataset to choose the number of clusters and the penalty parameters.

Figure 1 displays the true clusters, estimated clusters with cluster-specific unconstrained and diagonal covariance matrices respectively. Throughout this article, to reflect the sampling variability, for the true clusters, the parameter values were estimated based on the true cluster memberships of the observations. The correctly classified observations were represented by open circles or diamonds, while incorrect ones by filled ones. For set-up 1 (Figure 1), although there was only one cluster, due to the use of the diagonal covariance matrices, to account for the fact that the orientation of the true cluster was not parallel to either axis, two clusters with their orientation parallel to either axis were

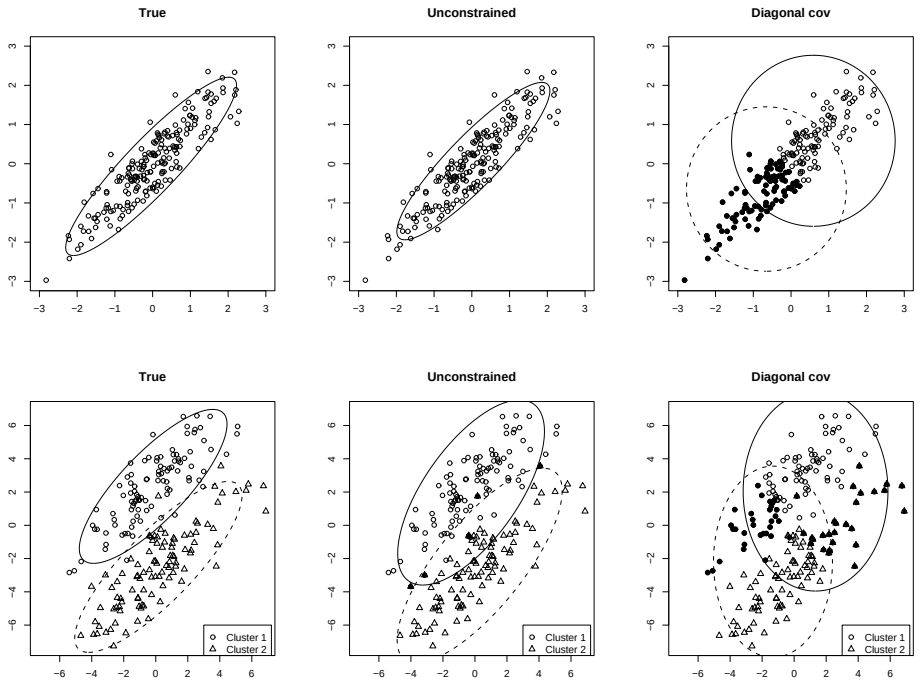

FIG 1. *Simulated data from only one cluster (top panels) and from two clusters (bottom panels) with* $K = 2$.

needed, leading to selecting an incorrect number of clusters. For set-up 2, even though a correct number of clusters was identified with the use of diagonal covariances, again due to the restriction on the orientation of the clusters imposed by the diagonal covariances, there were a large number of mis-classified observations. In contrast, the new method with unconstrained covariance matrices yielded much better results.

3.2. Simulated data with diagonal covariance matrices

We next considered three set-ups with $K > n$ and diagonal covariance matrices. The first was a null case with $g = 1$ cluster; the other two were with two clusters: one with difference only in the mean parameters, and the other in both the mean and variance parameters. For each set-up, 100 simulated datasets were generated. Each dataset contained $n = 80$ observations with dimension $K = 100$. For set-up 1, all observations belonged to the same cluster; for set-ups 2 and 3, the first 40 observations formed the first cluster while the remaining 40 observations belonged to the second cluster. Specifically, for set-up 1, all variables came from a standard normal distribution $N(0, 1)$; for set-up 2 and 3,

TABLE 1

Simulated data with diagonal covariances: frequencies of the selected numbers (g) of clusters, and mean numbers of predicted noise variables among the true informative (z_1) and noise variables (z_2). For set-up 1, the truth was $g = 1$, $z_1 = 10$ and $z_2 = 90$; for other set-ups, $g = 2$, $z_1 = 0$ and $z_2 = 90$.

		common covariance			cluster-specific covariance			diagonal covariance		
Set-up	g	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2
1	1	100	10	90	59	10	89.8	100	10	90
	2	0	-	-	25	9.4	84.3	0	-	-
	3	0	-	-	16	9.9	89.5	0	-	-
2	1	0	-	-	0	-	-	1	10	90
	2	74	0	85.6	58	0	82.7	83	0	84.5
	3	26	0	83.5	42	0	81.9	16	0	72.6
3	1	0	-	-	0	-	-	9	10	90
	2	68	0	76.0	87	0	83.1	91	0	82.0
	3	32	0	73.5	13	0	76.5	0	-	-

90 variables were noises generated from $N(0, 1)$, and the remaining 10 variables were informative. For set-up 2, the informative variables for the first cluster came from $N(0, 1)$ and from $N(1.5, 1)$ for the second cluster. For set-up 3, the informative variables were from $N(0, 1)$ and $N(1.5, 2)$ for the two clusters respectively.

We applied three methods: a common unconstrained covariance, cluster-specific unconstrained covariance matrices, and diagonal covariances. For the one with diagonal covariances, we used either a common diagonal covariance or cluster-specific diagonal covariances according to the truth to ideally optimize its performance. For each simulated dataset, we fitted a series of models with the number of components $g = 1$, 2 and 3, and performed a grid search to choose the penalty parameters. We show the frequencies of selecting various numbers of clusters, the average number of informative variables incorrectly selected to be noninformative (z_1), the average number of noninformative variables correctly selected (z_2), the number of observations correctly classified to cluster i (C_i), and the number of observations mis-classified from cluster i (IC_i). We also report the Rand index (RI) or adjusted Rand index (aRI) to summarize the quality of clustering.

As shown in Table 1, for the null case, we could select the correct number of clusters using the proposed method with a common covariance matrix, while the proposed method with cluster-specific covariance matrices did not perform so well. For set-up 2 with the true model with a common covariance matrix, the proposed method with a common unconstrained covariance matrix correctly selected $g = 2$ most often, and had a comparable performance to the method with a diagonal covariance in terms of the sample assignments, as summarized by the Rand or adjusted Rand index (Table 2). For set-up 3 with the true model with a cluster-dependent diagonal covariance matrices, the proposed method with cluster-dependent covariances correctly selected $g = 2$ nearly as often as using cluster-dependent diagonal covariance matrices. In terms of sample assignments, the proposed method also performed comparably. For these three

TABLE 2
Simulated data with diagonal covariance matrices: sample assignments for $\hat{g} = 2$ and (adjusted) Rand index (RI/aRI) values.

$\#C_i$ represents the average number of the samples correctly assigned to cluster i, $i = 1, 2$;
$\#I - C_i$ represents the average number of the samples incorrectly assigned to cluster i that # arise from the other cluster, $i = 1, 2$.

		Sample assignments, $\hat{g} = 2$				Rand Index							
		Cluster 1 ($n = 40$)		Cluster 2 ($n = 40$)		$\hat{g} = 1$		$\hat{g} = 2$		$\hat{g} = 3$		Overall	
Set-up	Methods	$\#C_1$	$\#IC_1$	$\#C_2$	$\#IC_2$	RI	aRI	RI	aRI	RI	aRI	RI	aRI
2	Common	39.8	0.1	40.0	0.0	-	-	0.99	0.99	0.98	0.96	0.99	0.98
	Cluster-spec	37.9	2.1	38.5	1.5	-	-	0.93	0.85	0.83	0.66	0.89	0.77
	Diagonal	39.6	0.4	39.5	0.5	0.49	0	0.98	0.95	0.75	0.51	0.94	0.88
3	Common	38.1	1.9	38.5	1.5	-	-	0.94	0.88	0.92	0.84	0.94	0.87
	Cluster-spec	37.9	2.1	38.6	1.4	-	-	0.92	0.85	0.87	0.75	0.91	0.84
	Diagonal	38.4	1.6	36.8	3.2	0.49	0	0.89	0.78	-	-	0.85	0.71

set-ups, a close examination indicated that for the proposed method the highest (predictive) log-likelihood was achieved when we had a sufficiently large penalty on the off-diagonal elements of the inverse covariance matrices, leading to the estimated covariance matrices close to being diagonal as were the truth.

3.3. Simulated data with non-diagonal covariance matrices

We considered some true models with non-diagonal covariance matrices, while other aspects of simulation remained the same as in section 3.2; in particular, among 100 variables, ten of which were informative. We used two non-diagonal covariance matrices for the 10 informative variables: a compound symmetry (CS) and an AR-1 with $\rho = 0.6$; the noise variables were independent of each other and of the informative variables. The resulting two covariance matrices are denoted as $V_{0,1}$ and $V_{0,2}$.

The following four set-ups were considered: set-up 1 was for a null case with only one cluster; for set-up 2 or3, we had two clusters sharing the same covariance matrix $V_{0,1}$ or $V_{0,2}$, but with mean parameters differing by 1.5 in each informative variable; for set-up 4, the two clusters differed in both the mean vectors (by 1.5) and covariance matrices as $V_{0,1}$ and $V_{0,2}$ respectively. As before, a training dataset contained 80 observations, 40 of which came from the first cluster (if there were two clusters); we used an independent tuning dataset of size 80. We applied the three methods; again according to the truth, we used a common diagonal covariance matrix for set-ups 1-3, but cluster-specific diagonal covariance matrices for set-up 4.

The frequencies of the selected numbers of clusters based on 100 simulated datasets are shown in Table 3. For each case, the proposed methods performed better than the diagonal matrix method; between the two proposed methods, depending on the truth (i.e. whether there was a common covariance matrix), one of them performed better than the other. The same conclusion can be drawn on the performance of the methods for sample classification (Table 4).

TABLE 3

Simulated data with non-diagonal covariance matrices: frequencies of the selected numbers (g) of clusters, and mean numbers of predicted noise variables among the true informative (z_1) and noise variables (z_2). For set-up 1, the truth was $g = 1$, $z_1 = 10$ and $z_2 = 90$; for others, $g = 2$, $z_1 = 0$ and $z_2 = 90$.

		common covariance			cluster-specific covariance			diagonal covariance		
Set-up	g	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2
1	1	43	10	90.0	36	10	90.0	0	-	-
	2	25	0	72.6	23	0	81.2	0	-	-
	3	32	0	88.7	41	0	90.0	100	0	77.5
2	1	0	-	-	7	10	90.0	0	-	-
	2	72	0	89.5	65	0	89.1	0	-	-
	3	28	0	85.0	28	0	81.4	100	0	75.6
3	1	0	-	-	0	-	-	0	-	-
	2	93	0	85.9	80	0	83.0	24	0	79.2
	3	7	0	79.0	20	0.7	80.4	76	0	78.1
4	1	0	-	-	0	-	-	0	-	-
	2	88	0	89.3	100	0	88.9	79	0	81.3
	3	12	0	86.1	0	-	-	21	0	82.0

TABLE 4

Simulated data with non-diagonal covariance matrices: sample assignments for $\hat{g} = 2$ and (adjusted) Rand index (RI/aRI) values.

$\#C_i$ represents the average number of the samples correctly assigned to cluster i, $i = 1, 2$; # $\#I - C_i$ represents the average number of the samples incorrectly assigned to cluster i that arise from the other cluster, $i = 1, 2$.

		Sample assignments, $\hat{g} = 2$				Rand Index							
		Cluster 1 ($n = 40$)		Cluster 2 ($n = 40$)		$\hat{g} = 1$		$\hat{g} = 2$		$\hat{g} = 3$		Overall	
Set-up	Methods	$\#C_1$	$\#IC_1$	$\#C_2$	$\#IC_2$	RI	aRI	RI	aRI	RI	aRI	RI	aRI
2	common	34.7	5.3	34.2	5.8	-	-	0.79	0.55	0.68	0.36	0.76	0.49
	cluster-spec	34.2	5.8	34.1	5.9	0.49	0	0.73	0.46	0.66	0.32	0.69	0.39
	diagonal	-	-	-	-	-	-	-	-	0.64	0.29	0.64	0.29
3	common	37.4	2.6	38.2	1.8	-	-	0.90	0.81	0.82	0.65	0.90	0.81
	cluster-spec	37.1	2.9	37.7	2.3	-	-	0.88	0.74	0.78	0.57	0.86	0.71
	diagonal	36.3	3.7	35.3	4.7	-	-	0.81	0.62	0.75	0.51	0.76	0.54
4	common	37.5	2.5	33.9	6.1	-	-	0.86	0.69	0.78	0.58	0.85	0.68
	cluster-spec	38.4	1.6	37.8	2.2	-	-	0.92	0.84	-	-	0.92	0.84
	diagonal	36.8	3.2	32.9	7.1	-	-	0.75	0.49	0.57	0.15	0.71	0.42

4. Examples

4.1. Leukemia gene expression data

4.1.1. Data and a clustering analysis

We first applied the methods to a well-known leukemia gene expression dataset of Golub et al. (1999) [14] to compare their performance. The (training) data contained 38 patient samples, among which 11 were acute myeloid leukemia (AML) while the remaining were acute lymphoblastic leukemia (ALL) samples. The ALL samples consisted of two subtypes: 8 T-cell and 19 B-cell samples. For each sample, the expression levels of 7129 genes were measured by an Affymetrix

TABLE 5
Clustering results for the leukemia gene expression data with $K = 300$ genes.

	Unconstrained covariances			Diagonal covariances		
Clusters (# Samples)	1	2	3	1	2	3
ALL-T (8)	7	1	0	8	0	0
ALL-B (19)	0	19	0	2	16	1
AML (11)	0	1	10	0	3	8

microarray. As in Dudoit et al. (2002) [9], we pre-processed the data in the following steps: 1) truncation: any expression level x_{jk} was truncated below at 1 if $x_{jk} < 1$, and above at 16,000 if $x_{jk} > 16,000$; 2) filtering: any gene was excluded if its $max/min \leq 5$ and $max - min \leq 500$, where max and min were the maximum and minimum expression levels of the gene across all the samples. Finally, as preliminary gene screening, we selected the top 300 genes with the largest sample variances across the 38 samples. Because there were only a small number of samples, we took stratified sampling by cell types in 3-fold cross validation (CV). We fitted the models with $g = 1, 2, 3$ and 4. The first local maximum of the predictive log-likelihood was achieved at $g = 3$.

The clustering results for $g = 3$ are shown in Table 5. Although all the 300 genes were selected as informative, there was evidence for a large number of genes with differential expression between the leukemia subtypes (e.g. Pan and Shen 2007 [31]). It is clear that the new method with cluster-specific unconstrained covariance matrices gave fewer errors in sample classification. To see why, we examined a few genes in more details. Genes CST3 (cystatin c, M23197) and ZYX (zyxin, X95735) were in the top 50 genes ranked by Golub et al. (1999) [14], and were two of the 17 genes selected by Antonov et al. (2004) [2] to distinguish the AML and ALL subtypes. CST3 was also regarded as a suitable marker by Bardi et al. (2004) [5]. Baker et al. (2006) [3] and Wang et al. (2005) [47] further identified ZYX as the most significant gene to discriminate AML/ALL subtypes. These two genes were also identified among the top 20 genes used in the classifier by Liao et al. (2007) [24]. In addition, we included two genes with gene accession number HG613-HT613 and M38591. The expression levels of gene pairs (HG613-HT613, M23197), and (X95735,M38591) are shown in Figure 2, with the true and estimated clusters. Clearly the true clusters did not necessarily have orientations parallel with either axis, which was captured by the proposed method, leading to more accurate sample classifications with more homogeneous clusters. We also examined element-wise differences between the covariance matrices resulting from the true clusters and estimated clusters: the unconstrained covariance matrices were much closer to the true covariance matrices than the diagonal covariance matrices were (results not shown).

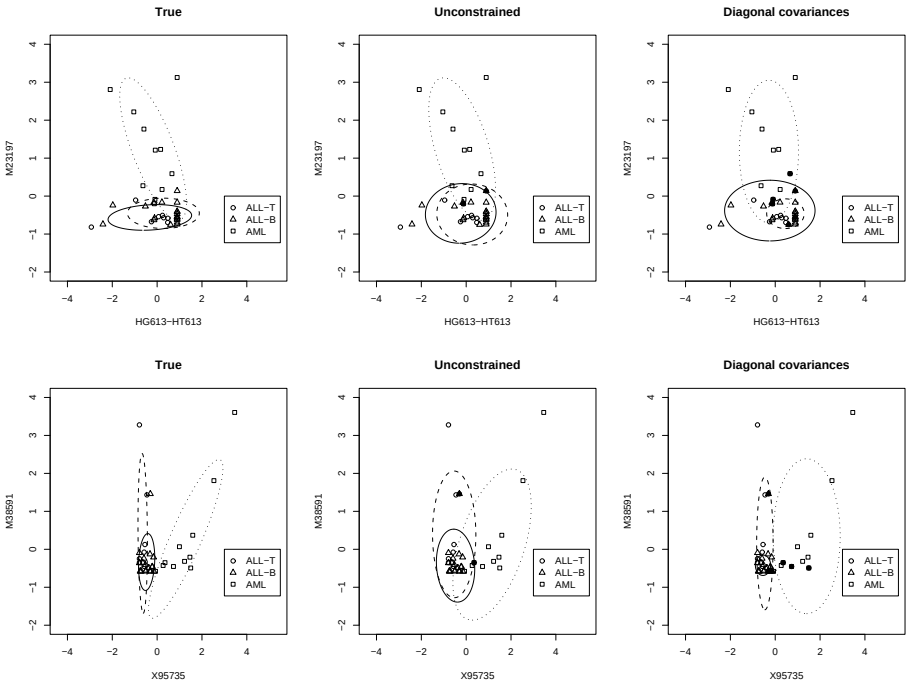

FIG 2. *Expression levels of gene pairs (HG613-HT613, M23197) and (X95735, M38591), and the corresponding clusters for the leukemia data.*

4.1.2. A comparison with a Bayesian method

Kim et al. (2006) [20] proposed a mixture of multivariate normal distributions via Dirichlet process (DP) mixtures for model-based clustering with the capability of variable selection for high-dimensional data. In particular, their method uses non-diagonal covariance matrices. Among others, a concentration parameter α for the DP prior and some data-dependent priors have to be specified; note that, as pointed out by other authors (Richardson and Green, 1997 [36]; Wasserman, 2000 [49]), it is not possible to obtain proper posterior distributions with fully noninformative priors in mixture models. Kim et al. (2006) [20] applied their method to the leukemia gene expression data of Golub et al. (1999) [14]. The data preprocessing was the same as before except that, rather than using the top 300 genes with the largest sample variances, all the 3571 genes were used. They considered two values of α. For $\alpha = 38$, the MCMC sampler visited models with 4 to 7 components. The sample allocations based on the maximum *a posteriori* probability (MAP) and on the least-squares clustering (LSC) algorithm were respectively,

$$\hat{c}_{MAP} = (\underbrace{1,2,1,\ldots,1,1,1,\ldots,1,1,1,1}_{\text{ALL}},\underbrace{2,1,4,5,3,2,3,4,2,2,7}_{\text{AML}}),$$

TABLE 6
Clustering results with cluster-specific unconstrained covariance matrices for the leukemia gene expression data with $K = 3571$ genes.

	$g = 3$			$g = 7$						
Clusters (# Samples)	1	2	3	1	2	3	4	5	6	7
ALL-T (8)	7	1	0	6	1	0	0	0	0	1
ALL-B (19)	1	15	3	0	11	0	7	0	1	0
AML (11)	0	0	11	0	0	9	0	2	0	0

$$\hat{c}_{LSC} = (\underbrace{1,2,1,\ldots,1,2,1,\ldots,1,2,1,1}_{\text{ALL}},\underbrace{2,1,4,5,3,2,3,6,2,2,7}_{\text{AML}}).$$

On the other hand, with $\alpha = 1$, the sampler visited models with 3 to 6 components, and the clustering results were

$$\hat{c}_{MAP} = (\underbrace{1,\ldots,1,1,1,\ldots,1,1,1,1}_{\text{ALL}},\underbrace{2,2,4,3,3,2,3,6,2,2,5}_{\text{AML}}),$$

$$\hat{c}_{LSC} = (\underbrace{1,\ldots,1,2,1,\ldots,1,2,1,1}_{\text{ALL}},\underbrace{2,2,4,5,3,2,3,6,2,2,5}_{\text{AML}}).$$

In each case, about 120 genes were selected to be informative.

We applied our proposed method with cluster-specific unconstrained covariance matrices. For the number of cluster $g = 1$ to 7, the predictive log-likelihood values based on a 3-fold CV were -28160, -24836, -23117, -23551, -23896, -24432 and -25218, respectively. Hence, we would select $g = 3$. For comparison, we showed the clustering results for both $g = 3$ and $g = 7$ in Table 6.

Comparing our results with that of Kim et al. (2006) [20], we see some significant differences. First, our method could distinguish the two subtypes of ALL, while that of Kim et al. (2006) [20] failed. Second, in contrast to that of Kim et al. (2006) [20], our results did not show the heterogeneous subgroups of AML even when we forced to have 7 clusters. Third, our method selected 3 clusters, corresponding to the three subtypes of the leukemia, while the Bayesian method chose a larger number of clusters. Finally, the method of Kim et al. (2006) [20] selected about 120 informative genes, in contrast to 1372 genes selected by our method for $g = 3$. We also note that the results of the Bayesian method depended on the choices of the prior and sample allocation method.

Currently we implemented our method in R, in which the glasso function is called to estimate a covariance matrix. For analysis of Golub's data as considered here, it took about two days to complete running our program on a laptop.

4.2. BOEC gene expression data

4.2.1. Data and a clustering analysis

One biologically interesting issue is whether human blood outgrowth endothelial cells (BOECs) belong to or are closer to either large vessel endothelial cells

TABLE 7
Clustering results for the BOEC data.

	Unconstrained cov			Diagonal cov		
Clusters (# Samples)	1	2	3	1	2	3
BOEC (27)	26	0	1	26	0	1
LVEC (28)	0	28	0	2	23	3
MVEC (25)	0	3	22	0	4	21

(LVECs) or microvascular endothelial cells (MVECs) based on global expression profiling. To address this question, as in Pan et al. (2006) [32], we combined the data from two separate studies: 28 LVEC and 25 MVEC samples in Chi et al. (2003) [7], and 27 BOEC samples in Jiang (2007) [18], and normalized the data as in Jiang (2007) [18]. Jiang (2007) [18] identified 37 genes that would discriminate the LVEC and MVEC samples, and we applied the proposed method using these 37 genes to cluster the 80 samples. Here, we used four-fold CV for tuning. We fitted the models with $g = 1$, 2, 3 and $g = 4$; the first local maximum of the predictive log-likelihood was reached at $g = 3$. At $g = 3$, 30 genes out of 37 were selected as informative.

The clustering results for three clusters are shown in Table 7. It is confirmed that each type of the samples clustered closely together, as observed by Jiang (2007) [18]. Nevertheless, it is noted that using unconstrained covariance matrices led to more accurate (i.e. more homogeneous) clustering than using diagonal covariance matrices. For illustration, two pairs of genes were chosen to have their expression levels and the corresponding true and estimated clusters plotted in Figure 3. Again, based on the true clusters, it is evident that non-diagonal covariance structures were captured better by our proposed method, leading to better (i.e. more homogeneous) clustering results; a direct examination of the estimated covariance matrices also confirmed this point (results not shown).

4.2.2. Semi-supervised learning

To address the biological question of whether the BOEC samples belong to either or neither of the LVEC and MVEC classes, we fully utilize the known memberships of the LVEC and MVEC samples while allowing the memberships of the BOEC samples to be determined. This is a semi-supervised learning problem with some observations with known cluster memberships (McLachlan and Peel 2002 [28]; Pan et al. 2006 [32]). Below we illustrate the application of the methods to the BOEC data.

For the BOEC data, treating the LVEC and MVEC samples with known class labels (i.e. $g_1 = 2$), we obtained the following results using the 37 genes as used before. First, if we did not allow the BOEC samples to form their own class with $g_0 = 0$, the BOEC samples (largely) fell to the class of MVEC, as shown before (Pan et al. 2006). Second, if we allowed the possibility of the BOEC samples to form their own class with $g_0 = 1$, they indeed formed a separate class. As shown

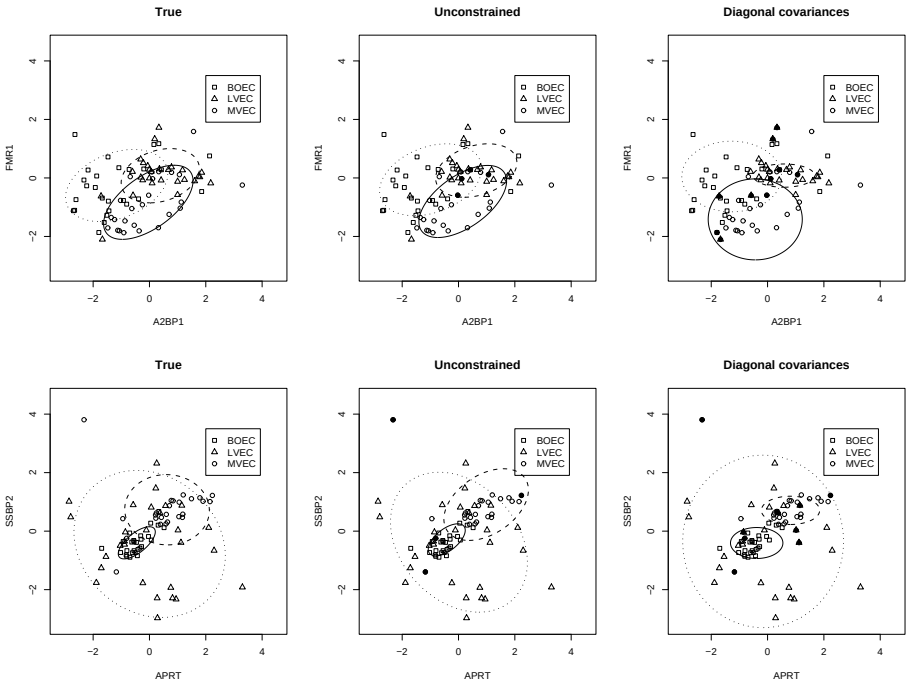

FIG 3. *Expression levels of gene pairs (A2BP1, FMR1) and (APRT, SSBP2), and the corresponding clusters for the BOEC data.*

TABLE 8
Semi-supervised learning results for the BOEC data.

	$g_0 = 0$, $g_1 = 2$			
	Unconstrained cov		Diagonal cov	
Clusters (# Samples)	1	2	1	2
BOEC (27)	0	27	2	25
LVEC (28)	28	0	24	4
MVEC (25)	0	25	3	22

	$g_0 = 1$, $g_1 = 2$					
	Unconstrained cov			Diagonal cov		
Clusters (# Samples)	1	2	3	1	2	3
BOEC (27)	22	3	2	22	2	3
LVEC (28)	0	28	0	0	27	1
MVEC (25)	0	0	25	0	4	21

in Table 8, in either case, the proposed method with unconstrained covariance matrices performed better than using diagonal covariance matrices with fewer mixed sample assignments. The proposed method selected 26 and 29 informative genes for the two cases respectively.

5. Discussion

We have proposed a new approach to penalized model-based clustering with unconstrained covariance matrices. We have shown its better empirical performance than that using diagonal covariance matrices when some informative variables are correlated, which is often the case for high-dimensional genomic data, as supported by co-expressions of functionally-related genes for microarray gene expression data. One key technical challenge is in estimating a possibly large covariance matrix. By taking advantage of the recent development in Gaussian graphical models, we have implemented our approach with the use of the graphical lasso algorithm (Friedman et al. 2007 [12]), largely due to its fast speed. Nevertheless, in principle, other covariance estimation methods, either frequentist (Huang et al. 2006 [17]; Yuan and Lin 2007 [53]; Levina et al. 2008 [22]; Rothman et al. 2009 [35]; Fan et al. 2009 [11], and references therein), or Bayesian (Jones et al. 2005 [19]; Scott and Carvalho 2009 [38]; Carvalho and Scott 2009 [6], and references therein), could be used, though computational speed is an important factor. Alternatively, some non-diagonal structural assumptions may be imposed on covariance matrices. Xie et al. (2009) [52] proposed such an approach based on the mixture of factor analyzers: some latent variables are used to model the covariance structure in each cluster, which however is computationally demanding if the number of the latent variables needs to be chosen data-adaptively. We comment on that, although we have focused on its application in variable selection, penalized model-based clustering may be useful in its own right, such as in providing better parameter estimates (due to regularization) and thus better clustering results for small samples. For future improvement, we may follow Xie et al. (2008b) [51] to impose a penalty on variance parameters to account for clustering structures in varying variances across clusters, in addition to that in locations or means. It may be of interest to extend the proposed method to deal with other issues in high-dimensional genomic data, such as prior biological knowledge and outliers (Pan 2006 [30]; Tseng 2007 [44]). More investigations are necessary, especially in further evaluating the proposed method with real-world applications.

Free software will be posted on our web site.

Appendix A: Appendix: Proof of Theorem 1

For simplicity of notation, we drop the "hat" from any parameter estimate.

Since Q_P is differentiable with respect to μ_{ik} when $\mu_{ik} \neq 0$, while non-differentiable at $\mu_{ik} = 0$, we consider the following two cases:

i) If $\mu_{ik} \neq 0$ is a maximum, given that Q_P is concave and differentiable, then the sufficient and necessary condition for μ_{ik} to be the global maximum of Q_P is

$$\partial Q_P/\partial \mu_{ik} = 0 \Longleftrightarrow \sum_{j=1}^{n} \tau_{ij} \left(\sum_{l=1}^{K} (x_{jl} - \mu_{il}) W_{lk} \right) - \lambda_1 \text{sign}(\mu_{ik}) = 0,$$

from which (12) can be easily derived if we separate μ_{ik} from other components of μ_i.

ii) If $\mu_{ik} = 0$ is a maximum, we compare $Q_P(0, .)$ with $Q_P(\Delta\mu_{ik}, .)$, the values of Q_P at $\mu_{ik} = 0$ and $\mu_{ik} = \Delta\mu_{ik}$ respectively (while other components of μ_i are fixed at its maximum). By definition, we have

$$
\begin{aligned}
&Q_P(0, .) \geq Q_P(\Delta\mu_{ik}, .) \text{ for any } \Delta\mu_{ik} \text{ near } 0 \\
&\Longleftrightarrow \\
&\sum_{j=1}^{n} \tau_{ij}^{(r)} \left[(x_j - \mu_i)^{\mathrm{t}} W (x_j - \mu_i)|_{\mu_{ik}=\Delta\mu_{ik}} - (x_j - \mu_i)^{\mathrm{t}} W (x_j - \mu_i)|_{\mu_{ik}=0}\right] \\
&\quad \geq -2\lambda_1 |\Delta\mu_{ik}| \\
&\Longleftrightarrow \\
&\sum_{j=1}^{n} \tau_{ij}^{(r)} \left[2\Delta\mu_{ik} \sum_{s=1, s\neq k}^{K} (x_{js} - \mu_{is}) W_{ks} + W_{kk} \left(-\Delta\mu_{ik}^2 + 2x_{jk}\Delta\mu_{ik}\right)\right] \\
&\quad \leq 2\lambda_1 |\Delta\mu_{ik}| \\
&\Longleftrightarrow \\
&\left| \sum_{j=1}^{n} \tau_{ij}^{(r)} \left(\sum_{s=1, s\neq k}^{K} (x_{js} - \mu_{is}) W_{sk} + x_{jk} W_{kk} \right) \right| \leq \lambda_1, \text{ as } \Delta\mu_{ik} \to 0.
\end{aligned}
$$

This yields (12) and (13).

Acknowledgements

This research was partially supported by NIH grant RO1-GM081535; in addition, HZ and WP by NIH grant RO1-HL65462, and XS by NSF grants DMS-0604394 and DMS-0906616.

References

[1] Alaiya, A.A. et al. (2002). Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles. *Int. J. Cancer*, **98**, 895–899.

[2] Antonov, A.V., Tetko, I.V., Mader, M.T., Budczies, J. and Mewes, H.W. (2004). Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644–652.

[3] Baker, Stuart G. and Kramer, Barnett S. (2006). Identifying genes that contribute most to good classification in microarray. *BMC Bioinformatics*, Sep 7; 7:407.

[4] Banfield, J.D. and Raftery, A.E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**, 803–821. MR1243494

[5] Bardi, E., Bobok, I., Olah, A.V., Olah, E., Kappelmayer, J. and Kiss, C. (2004). Cystatin C is a suitable marker of glomerular function in children with cancer. *Pediatric Nephrology*, **19**, 1145–1147.
[6] Carvalho, C.M. and Scott, J.G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, **96**, 497–512.
[7] Chi, J-T. et al. (2003). Endothelial cell diversity revealed by global expression profiling. *PNAS*, **100**, 10623–10628.
[8] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *JRSS-B*, **39**, 1–38. MR0501537
[9] Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using expression data. *J. Am. Stat. Assoc.*, **97**, 77–87. MR1963389
[10] Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14863–14868.
[11] Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Stat.*, **3**, 521–541.
[12] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **0**, 1–10.
[13] Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *Computer J.*, **41**, 578–588.
[14] Golub, T. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
[15] Guo, F.J., Levina, E., Michailidis, G. and Zhu, J. (2009). Pairwise Variable Selection for High-dimensional Model-based Clustering. To appear in *Biometrics*.
[16] Hoff. P.D. (2006). Model-based subspace clustering. *Bayesian Analysis*, **1**, 321–344. MR2221267
[17] Huang, J.Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance selection and estimation via penalised normal likelihood. *Biometrika*, **93**, 85–98. MR2277742
[18] Jiang, A., Pan, W., Yu, S. and Robert, P.H. (2007). A practical question based on cross-platform microarray data normalization: are BOEC more like large vessel or microvascular endothelial cells or neither of them? *Journal of Bioinformatics and Computational Biology* **5** 875–893.
[19] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, **20**, 388–400. MR2210226
[20] Kim, S., Tadesse, M.G. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**, 877–893. MR2285077
[21] Lau, J.W. and Green, P.J. (2007) Bayesian model based clustering procedure. *Journal of Computational and Graphical Statistics*, **16**, 526–558. MR2351079

[22] Levina, L., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, **2**, 245–263. MR2415602

[23] Liang, F., Mkherjee, S. and West, M. (2007). The use of unlabeled data in predictive modeling. *Statistical Science*, **22**, 189–205. MR2408958

[24] Liao, J.G. and Chin, K.V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, **23**, 1945–1951.

[25] Liu, J.S., Zhang, J.L., Palumbom M.J. and Lawrencem C.E. (2003). Bayesian clustering with variable and transformation selection (with discussion). *Bayesian Statistics* **7**, 249–275. MR2003177

[26] McLachlan, G. (1987). On bootstrapping likelihood ratio test statistics for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

[27] McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413 - 422.

[28] McLachlan, G.J. and Peel, D. (2002). *Finite Mixture Model.* New York, John Wiley & Sons, Inc. MR1789474

[29] Muller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79. MR1399156

[30] Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, **22**, 795–801.

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

[32] Pan, W., Shen, X., Jiang, A. and Hebbel, R.P. (2006). Semi-supervised learning via penalized mixture model with application to microarray sample classification. *Bioinformatics* **22**, 2388–2395.

[33] Raftery, A.E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**, 168–178. MR2268036

[34] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *JASA*, **66**, 846–850.

[35] Rothman, A., Levina, L. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *JASA*, 2009, 104(485): 177–186. MR2504372

[36] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixture models (with Discussion). *J R Statist Soc B*, **59**, 731–792. MR1483213

[37] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. MR0468014

[38] Scott, J.G. and Carvalho, C.M. (2009). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Stat.*, **17**, 790–808.

[39] Tadesse, M.G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**, 602–617. MR2160563

[40] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and

Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet*, **22**, 281–285.

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

[42] Teh, Y.W., Jordan, M.I., Beal, M.J. and Beal, M.J. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. *NIPS*.

[43] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *JRSS-B*, **58**, 267–288. MR1379242

[44] Tseng, G.C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data *Bioinformatics*, **23**, 2247–2255.

[45] Tseng, P. (1988) Coordinate ascent for maximizing nondifferentiable concave functions. Technical report LIDS-P; 1840, Massachusetts Institute of Technology. Laboratory for Information and Decision Systems.

[46] Tseng, P. (2001) Convergence of block coordinate descent method for nondifferentiable maximization. *J. Opt. Theory and Applications*, **109**, 474–494. MR1835069

[47] Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X. and Mewes, H.W. (2005). Gene selection from microarray data for cancer classification - a machine learning approach. *Comput Biol Chem*, **29**, 37–46.

[48] 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

[49] Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J R Statist Soc B*, **62**, 159–180. MR1747402

[50] Xie, B., Pan, W. and Shen, X. (2008a). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, **64**, 921–930.

[51] Xie, B., Pan, W. and Shen, X. (2008b). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electron. J. Statist.*, **2**, 168–212. MR2386092

[52] Xie, B., Pan, W. and Shen, X. (2009). Penalized mixtures of factor analyzers with application to clustering high dimensional microarray data. To appear *Bioinformatics*.
Available at `http://www.biostat.umn.edu./rrs.php` as Research Report 2009-019, Division of Biostatistics, University of Minnesota.

[53] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35. MR2367824