BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 4 2010, pages 501–508
doi:10.1093/bioinformatics/btp707

Gene expression

Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data

Benhuai Xie[1], Wei Pan[1,*] and Xiaotong Shen[2]

[1]Division of Biostatistics, School of Public Health and [2]School of Statistics, University of Minnesota, Minneapolis, MN, USA

Received on July 2, 2009; revised on December 11, 2009; accepted on December 18, 2009

Advance Access publication December 23, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Model-based clustering has been widely used, e.g. in microarray data analysis. Since for high-dimensional data variable selection is necessary, several penalized model-based clustering methods have been proposed tørealize simultaneous variable selection and clustering. However, the existing methods all assume that the variables are independent with the use of diagonal covariance matrices.

Results: To model non-independence of variables (e.g. correlated gene expressions) while alleviating the problem with the large number of unknown parameters associated with a general non-diagonal covariance matrix, we generalize the mixture of factor analyzers to that with penalization, which, among others, can effectively realize variable selection. We use simulated data and real microarray data to illustrate the utility and advantages of the proposed method over several existing ones.

Contact: weip@biostat.umn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Clustering is a popular tool for exploratory data analysis in many fields, including for high-dimensional microarray data. For example, Eisen *et al*. (1998) found that, for both the budding yeast and human, genes with similar functions were likely to be grouped together based on their expression profiles, suggesting that clustering genes with their expression profiles might help predict gene functions. Golub *et al*. (1999) clustered human leukemia samples with their expression profiles and discovered distinct groups corresponding to subtypes of leukemia. Thalamuthu *et al*. (2006) compared various clustering methods and found that model-based clustering (Fraley and Raftery, 2002) performed well for microarray gene expression data. On the other hand, for high-dimensional and low sample-sized data, several authors (Pan and Shen, 2007; Wang and Zhu, 2008; Xie *et al*., 2008a, b) have shown that variable selection is necessary for uncovering underlying clustering structures, and that penalized model-based clustering is effective in realizing variable selection and clustering simultaneously. However, in their penalized model-based clustering approaches, all variables are assumed to be independent with diagonal covariance matrices being used in a mixture of normals. In practice, some variables, e.g. genes, may be related to each other, leading to non-negligible correlations among them, violating the independence assumption with the use of diagonal covariance matrices.

Here we aim to generalize existing penalized model-based clustering approaches to the case with non-diagonal covariance matrices. For high dimensional and low sample sized data, if a general and unrestricted covariance matrix is used, there will be a large number of unknown parameters (i.e. its off-diagonal elements) to be estimated. In addition, there will be some computational issues in implementing the constraint that the resulting covariance matrix estimate is positive definite (Huang *et al*., 2006; Yuan and Lin, 2007). As an intermediate between a diagonal and a general covariance matrix, we model a covariance matrix using some latent variables as done in the mixture of factor analyzers (MFAs) (McLachlan and Peel, 2000). Hinton *et al*. (1997) proposed the MFA as a natural extension of a single factor analysis model, by adopting a finite mixture of single factor analysis models. Ghahramani and Hinton (1997) provided an exact EM algorithm for MFA. In clustering tissue samples with microarray gene expression data, McLachlan *et al*. (2002, 2003) first selected a subset of the genes by univariate screening, and then used a MFAs to effectively reduce the dimension of the feature space; see McLachlan *et al*. (2007), Baek and McLachlan (2008), Baek *et al*. (2009) for more recent applications and extensions. Here we extend penalized model-based clustering with diagonal covariance matrices to penalized mixtures of factor analyzers (PMFA) to capture a more general covariance structure for high-dimensional data. Variable selection and model fitting can be realized simultaneously in PMFA as proposed below.

In the next section, we first review the MFA and its EM algorithm as proposed by Ghahramani and Hinton (1997), and then propose a PMFA and derive its EM algorithm. This is followed in Section 3 by numerical results to illustrate the utility and advantages of our proposed PMFA over the MFA and the penalized mixture of normals with a diagonal covariance matrix (PMND) (Pan and Shen, 2007). We end with a short discussion in Section 4.

2 METHODS

2.1 Mixture of factor analyzers and its EM algorithm

Factor analysis can be used to explain the correlations between variables and for dimension reduction for multivariate observations. In a single-component factor analysis, a K-dimensional observation $x_j, j = 1, \ldots, n$ is modeled using

*To whom correspondence should be addressed.

a q-dimensional vector of real-valued factors U_j (latent or unobservable variables), where q is generally much smaller than K (Everitt, 1984). Each observation x_j is modeled as

$$x_j = \mu + \mathrm{B}\mathrm{U}_j + e_j,$$

where B is an unknown $K \times q$ factor loading matrix. The factors U_j are assumed to be $N(\mathbf{0}, I_q)$ distributed, and independent of the K-dimensional random variable e_j from $N(\mathbf{0}, D)$, where I_q, $D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_K^2)$ are a $q \times q$ identity matrix and a $K \times K$ diagonal matrix, respectively. According to this model, x_j therefore follows a normal distribution with mean μ and covariance matrix $BB' + D$. Note that $BB' + D$ is in general non-diagonal.

In the context of mixture modeling, Hinton *et al*. (1997) and Ghahramani and Hinton (1997) provided a local dimension reduction by assuming that the distribution of x_j can be modeled as

$$x_j = \mu_i + B_i U_{ij} + e_{ij}, \tag{1}$$

for $j = 1, 2, \ldots, n$, with prior probability $\pi_i, i = 1, \ldots, g$, where B_i is a $K \times q$ factor loading matrix. The factors U_{ij} and random variable e_{ij} are assumed to be independently distributed as $N(\mathbf{0}, I_q)$ and $N(\mathbf{0}, D)$, respectively, and U_{ij} is independent of e_{ij}.

The observations x_j's are assumed to be iid from a mixture distribution with g components: $\sum_{i=1}^{g} \pi_i H_i(x_j; \theta_i)$, where θ_i is a vector representing all unknown parameters in the distribution for component i, while π_i is the prior probability for component i. Denote

$$h_i(x_j, U_{ij}; \theta_i) = f_i(x_j | U_{ij}; \theta_i) g(U_{ij}; \theta_i),$$

where $f_i(x_j | U_{ij}; \theta_i)$ and $g(U_{ij}; \theta_i)$ are the density functions of normal distributions $N(\mu_i + B_i U_{ij}, D)$ and $N(\mathbf{0}, \mathbf{I}_q)$, respectively. According to model (1), $H_i(x_j; \theta_i)$ can be obtained by marginalizing $h_i(x_j, U_{ij}; \theta_i)$ over U_{ij}, yielding $H_i(x_j; \theta_i)$ as the density function for normal distribution $N(\mu_i, B_i B_i' + D)$.

The log-likelihood is

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

where $\Theta = \{(\theta_i, \pi_i) : i = 1, \ldots, g\}$ represents all unknown parameters. The maximum likelihood estimate (MLE) $\tilde{\Theta}$ is obtained by maximizing $\log L(\Theta)$. A commonly used algorithm is the E-M (Dempster *et al*., 1977). Denote by z_{ij} the indicator of whether x_j is from component i. Because we do not know beforehand which component an observation comes from, z_{ij}'s are regarded as missing data. If latent variables z_{ij}'s and U_{ij}'s could be observed, then the complete-data log-likelihood is

$$\log L_c(\Theta) = \sum_i \sum_j z_{ij} [\log \pi_i + \log h_i(x_j, U_{ij}; \theta_i)] \tag{2}$$

Let $X = \{x_j : j = 1, \ldots, n\}$ represent the observed data. Given the current estimate $\tilde{\Theta}^{(r)} = \{(\tilde{\theta}_i^{(r)}, \tilde{\pi}_i^{(r)}) : i = 1, \ldots, g\}$ at iteration r, the E-step of the EM calculates

$$\begin{aligned} Q(\Theta; \tilde{\Theta}^{(r)}) &= E_{\tilde{\Theta}^{(r)}}(\log L_c | X) \\ &= \sum_i \sum_j \tilde{\tau}_{ij}^{(r)} [\log \pi_i + E(\log h_i(x_j, U_{ij}; \theta_i) | X, \tilde{\theta}_i^{(r)})], \end{aligned} \tag{3}$$

where $\tilde{\tau}_{ij}^{(r)}$ is the estimated posterior probability of x_j's coming from component i:

$$\tilde{\tau}_{ij}^{(r)} = \frac{\tilde{\pi}_i^{(r)} H_i(x_j; \tilde{\theta}_i^{(r)})}{\sum_{i=1}^{g} \tilde{\pi}_i^{(r)} H_i(x_j; \tilde{\theta}_i^{(r)})}, \tag{4}$$

and

$$\begin{aligned} & E(\log h_i(x_j, U_{ij}; \theta_i) | X, \theta_i^{(r)}) \\ = \; & -\frac{1}{2} \left[\log(|D|) + (x_j - \mu_i)' D^{-1} (x_j - \mu_i) \right] \\ & + (x_j - \mu_i)' D^{-1} B_i E(U_{ij} | X, \theta_i^{(r)}) \\ & - \frac{1}{2} \left[\mathrm{tr}(B_i' D^{-1} B_i E(U_{ij} U_{ij}' | X, \theta_i^{(r)})) + \mathrm{tr}(E(U_{ij} U_{ij}' | X, \theta_i^{(r)})) \right], \end{aligned}$$

up to some additive constant, and tr() is the trace operator. The M-step maximizes Q to update Θ.

The E-step involves the calculation of $E(U_{ij} | X, \theta_i)$ and $E(U_{ij} U_{ij}' | X, \theta_i)$, which can be derived from the fact that random vector $(x_j', U_{ij}')'$ has a multivariate normal distribution with mean and covariance matrix

$$\begin{pmatrix} \mu_i \\ \mathbf{0} \end{pmatrix}, \quad \begin{pmatrix} B_i B_i' + D & B_i \\ B_i' & I_q \end{pmatrix},$$

respectively. By applying the standard results of multivariate normal distribution, the conditional expectations can be obtained as following:

$$\begin{aligned} E(U_{ij} | X, \theta_i) &= \gamma_i' (x_j - \mu_i), \\ E(U_{ij} U_{ij}' | X, \theta_i) &= I_q - \gamma_i' B_i + \gamma_i' (x_j - \mu_i)(x_j - \mu_i)' \gamma_i, \end{aligned}$$

where $\gamma_i = (B_i B_i' + D)^{-1} B_i$.

The detailed EM derivation of MFA can be found in Ghahramani and Hinton (1997). In the following, we just list the updates of Θ: for the prior probability of an observation from the i-th component H_i,

$$\tilde{\pi}_i^{(r+1)} = \sum_{j=1}^{n} \tilde{\tau}_{ij}^{(r)} / n, \tag{5}$$

for the factor loading matrix B_i,

$$\tilde{B}_i^{(r+1)} = \sum_j \tilde{\tau}_{ij}^{(r)} (x_j - \tilde{\mu}_i^{(r)}) E(U_{ij}' | X, \tilde{\theta}_i^{(r)}) \left(\sum_j \tilde{\tau}_{ij}^{(r)} E(U_{ij} U_{ij}' | X, \tilde{\theta}_i^{(r)}) \right)^{-1}, \tag{6}$$

for the diagonal variance matrix D,

$$\begin{aligned} \tilde{D}^{(r+1)} = \; & \frac{1}{n} \mathrm{Diag} \left(\sum_{i,j} \tilde{\tau}_{ij}^{(r)} (x_j - \tilde{\mu}_i^{(r)})(x_j - \tilde{\mu}_i^{(r)})' \right. \\ & \left. - \sum_{i,j} \tilde{\tau}_{ij}^{(r)} (x_j - \tilde{\mu}_i^{(r)}) E(U_{ij}' | X, \tilde{\theta}_i^{(r)}) \tilde{B}_i^{(r)'} \right), \end{aligned} \tag{7}$$

where $\mathrm{Diag}(A)$ extracts the diagonal elements of any matrix A to form a diagonal matrix, and for the mean parameter μ_i of the i-th component,

$$\tilde{\mu}_i^{(r+1)} = \sum_j \tilde{\tau}_{ij}^{(r)} [x_j - \tilde{B}_i^{(r)} E(U_{ij} | X, \tilde{\theta}_i^{(r)})] / \sum_j \tilde{\tau}_{ij}^{(r)}. \tag{8}$$

The above E- and M-steps are iterated; at the convergence, we obtain the MLE $\tilde{\Theta} = \tilde{\Theta}^{(\infty)}$.

2.2 PMFAs and its EM algorithm

Before clustering analysis, it is assumed throughout that the data have been standardized to have sample mean 0 and sample variance 1 across the n observations for each variable. As discussed by Pan and Shen (2007), with high-dimensional data, the presence of many noise variables may severely mask clustering structures, suggesting the necessity of conducting variable selection. Their study has shown that, when clustering high-dimensional data, variable selection and model fitting can be realized simultaneously by adding an L_1 penalty of mean parameters to the (complete data) log-likelihood under a common diagonal covariance matrix for each cluster. Denote the center (or mean) of cluster i as $\mu_i = (\mu_{i1}, \ldots, \mu_{iK})'$. With a common diagonal covariance matrix, variable k is irrelevant to clustering if and only if all the cluster centers are the same across the g clusters: $\mu_{1k} = \mu_{2k} = \cdots = \mu_{gk}$; by the data standardization of the grand sample mean at 0 for each variable, we then have $\mu_{1k} = \mu_{2k} = \cdots = \mu_{gk} = 0$. Thus, we can use an L_1 penalty on the mean parameters μ_{ik}'s to shrink some of them to be 0 to realize variable selection. Note that, by standardizing each variable to have variance 1, we can treat these variables in a similar scale and thus penalize their mean parameters together by an L_1 penalty. Similarly, we can realize variable selection in a MFAs by selecting a proper penalty function. In addition to that all $\mu_{ik}, i = 1, \ldots, g$ are 0, all $b_{ik\cdot} = (b_{ik1}, \ldots, b_{ikq}), i = 1, \ldots, g$

are required to be $\mathbf{0}$ to guarantee the irrelevance of variable k to all clusters. Note that our proposed approach can eliminate irrelevant variables, but not redundant variables; if the latter is desired, one can take a supervised learning approach with the discovered clusters as classes and the selected variables as candidate predictors. Alternatively, Raftery and Dean (2006) proposed a Bayesian approach to eliminate both irrelevant and redundant variables, but it is computationally too demanding for high-dimensional data.

We use L_1 penalty function $p_1(\mu)=\sum_i\sum_k|\mu_{ik}|$ for mean parameters and $p_2(B)=\sum_i\sum_k||b_{ik.}||_2$ for factor loading B_i's, where B is the set of all B_i's, and $||b_{ik.}||_2=\sqrt{\sum_l b_{ikl}^2}$. Hence the penalty is

$$\begin{aligned} p_{\lambda_1,\lambda_2}(\mu,B) &= \lambda_1 p_1(\mu)+\lambda_2 p_2(B) \\ &= \lambda_1\sum_i\sum_k|\mu_{ik}|+\lambda_2\sum_i\sum_k||b_{ik.}||_2. \end{aligned} \tag{9}$$

The L_1 norm $p_1(\mu)$, as in Pan and Shen (2007), is used to shrink a small estimate of μ_{ik} to be exactly 0, while $p_2(B)$, serving as a grouped variable penalty as in Yuan and Lin (2006) and Xie *et al*. (2008a), is used to shrink an estimate of factor loading vector $b_{ik.}$ that is close to $\mathbf{0}$ to be exactly $\mathbf{0}$. Therefore, if a variable k, having common mean 0 and common variance σ_k^2 across clusters, is independent of all other variables with $b_{ik.}=\mathbf{0}$ for any i, this variable is effectively treated as irrelevant; this can be verified in (4), where an irrelevant variable does not contribute to the posterior probability τ_{ij}, thus irrelevant to all clusters. Note that other penalty functions of the mean and factor loading parameters could be used as in Xie *et al*. (2008a) and Wang and Zhu (2008).

The penalized log-likelihood is

$$\log L_P(\Theta)=\sum_{j=1}^n\log\left[\sum_{i=1}^g\pi_i H_i(x_j;\theta_i)\right]-p_{\lambda_1,\lambda_2}(\mu,B). \tag{10}$$

In order to compute the maximum penalized likelihood estimate (MPLE) $\hat\Theta$ from (10), we derived the following EM algorithm. First, the penalized complete-data log-likelihood is

$$\log L_{c,P}(\Theta)=\sum_i\sum_j z_{ij}[\log\pi_i+\log h_i(x_j,U_{ij};\theta_i)]-p_{\lambda_1,\lambda_2}(\mu,B). \tag{11}$$

Accordingly, at iteration r, the E-step of the EM calculates

$$\begin{aligned} &Q_P(\Theta;\hat\Theta^{(r)})=E_{\hat\Theta^{(r)}}(\log L_{c,P}|X) \\ &= \sum_i\sum_j\hat\tau_{ij}^{(r)}[\log\pi_i+E(\log h_i(x_j,U_{ij};\theta_i)|X,\hat\theta_i^{(r)})] \\ &\quad -p_{\lambda_1,\lambda_2}(\mu,B), \end{aligned} \tag{12}$$

while the M-step maximizes Q_P to update Θ to $\hat\Theta^{(r+1)}$, resulting in the same updating formulas for τ_{ij}, π_i and D as given in (4), (5) and (7), respectively. Similar to that in Xie *et al*. (2008a, b), we show in Supplementary Materials the following sufficient and necessary conditions for $\hat\mu_{ik}$ to be a global maximizer of Q_P:

$$\hat\mu_{ik}^{(r+1)}=\tilde\mu_{ik}^{(r+1)}\left(1-\frac{\lambda_1\hat\sigma_k^{2,(r)}}{\sum_j\hat\tau_{ij}^{(r)}|\tilde\mu_{ik}^{(r+1)}|}\right)_+ \tag{13}$$

where $\tilde\mu_{ik}^{(r+1)}$ has the form of the MLE of μ_{ik} (without penalty) as given in (8), and $x_+=(|x|+x)/2$.

For the factor loading matrix B_i, we have the following theorem (with a proof in Supplementary Materials):

THEOREM 1. *The sufficient and necessary conditions for* $\hat b_{ik.}^{(r+1)}=(\hat b_{ik1}^{(r+1)},\ldots,\hat b_{ikq}^{(r+1)})$ *to be a global maximizer of* Q_P *are: (i) if* $\hat b_{ik.}^{(r+1)}\neq\mathbf{0}$,

$$\begin{aligned} &\sum_j\hat\tau_{ij}^{(r)}(x_j-\hat\mu_i^{(r)})E(U_{ij}'|X,\hat\theta_i^{(r)}) \\ &-\hat B_i^{(r+1)}\sum_j\hat\tau_{ij}^{(r)}E(U_{ij}U_{ij}'|X,\hat\theta_i^{(r)}) \\ &= \lambda_2\sqrt{q}\hat G^{(r+1)}\hat B_i^{(r+1)}, \end{aligned} \tag{14}$$

where $\hat G^{(r+1)}=\mathrm{diag}(\hat\sigma_1^{2,(r)}/||\hat b_{i1.}^{(r+1)}||_2,\ldots,\hat\sigma_K^{2,(r)}/||\hat b_{iK.}^{(r+1)}||_2)$, *and (ii) if* $\hat b_{ik.}^{(r+1)}=\mathbf{0}$,

$$\begin{aligned} &\left(\sum_l\left(\sum_j\hat\tau_{ij}^{(r)}(x_{jk}-\hat\mu_{ik}^{(r)})E(U_{ijl}|X,\hat\theta_i^{(r)})\right)^2\right)^{1/2} \\ &\leq \lambda_2\sqrt{q}\hat\sigma_k^{2,(r)}. \end{aligned} \tag{15}$$

If we focus on $b_{ik.}$, (14) becomes: if $\hat b_{ik.}^{(r+1)}\neq\mathbf{0}$,

$$\begin{aligned} &\sum_j\hat\tau_{ij}^{(r)}(x_{jk}-\hat\mu_{ik}^{(r)})E(U_{ij}'|X,\hat\theta_i^{(r)})- \\ &\hat b_{ik.}^{(r+1)}\sum_j\hat\tau_{ij}^{(r)}E(U_{ij}U_{ij}'|X,\hat\theta_i^{(r)}) \\ &= \frac{\lambda_2\sqrt{q}\hat\sigma_k^{2,(r)}\hat b_{ik.}^{(r+1)}}{||\hat b_{ik.}^{(r+1)}||_2}. \end{aligned} \tag{16}$$

Naturally formulas (16) and (15) suggest the following updating algorithm for $\hat B_i^{(r+1)}$:

(1) if (15) is satisfied, then $\hat b_{ik.}^{(r+1)}=\mathbf{0}$;
(2) if (15) is not satisfied, then the Newton-Raphson algorithm is used to obtain a non-zero $\hat b_{ik.}^{(r+1)}$ from (16);
(3) steps 1 and 2 are repeated for $k=1,2,\ldots,K$.

The above iterative process is continued; at the convergence, we obtain the MPLE $\tilde\Theta=\hat\Theta^{(\infty)}$.

To compare the MPLE of loading vector $b_{ik.}$ with its MLE, we consider an iteration with other parameters fixed: from (6) we have MLE

$$\tilde b_{ik.}=\sum_j\tau_{ij}(x_{jk}-\mu_{ik})E(U_{ij}'|X,\theta_i)\left(\sum_j\tau_{ij}E(U_{ij}U_{ij}'|X,\theta_i)\right)^{-1},$$

and from (16) we have MPLE

$$\hat b_{ik.}=\tilde b_{ik.}\left(I_q+\frac{\lambda_2\sqrt{q}\sigma_k^2}{||b_{ik.}||_2}\left(\sum_j\tau_{ij}E(U_{ij}U_{ij}'|X,\theta_i)\right)^{-1}\right)^{-1}$$

for $\hat b_{ik.}\neq\mathbf{0}$. Note that $\sum_j\tau_{ij}E(U_{ij}U_{ij}'|X)$ is positive definite, and $\lambda_2\sqrt{q}\sigma_k^2/||b_{ik.}||_2\geq 0$. Hence, if $\hat b_{ik.}\neq\mathbf{0}$, $\hat b_{ik.}$ is shrunken from MLE $\tilde b_{ik.}$ towards $\mathbf{0}$; $\hat b_{ik.}$ can be exactly $\mathbf{0}$ if, for example, λ_2 is sufficiently large as shown in (15).

2.3 Model selection

Commonly used model selection methods, such as cross-validation, can be used to select tuning parameters $(g,q,\lambda_1,\lambda_2)$. To save computing time, we propose using the *predictive* log-likelihood based on an independent tuning dataset as our model selection criterion. For any given $(g,q,\lambda_1,\lambda_2)$, the predictive log-likelihood for the tuning data can be obtained by plugging-in the tuning data into $\log L(\hat\theta)$, where $\hat\theta$ is the MPLE (or MLE for MFA) estimated from the training data. We propose using a grid search to estimate the optimal $(\hat g,\hat q,\hat\lambda_1,\hat\lambda_2)$ as the one with the maximum predictive log-likelihood for the tuning data.

For any given $(g,q,\lambda_1,\lambda_2)$, because of the possible existence of many local maxima for the mixture model, we have to run an EM algorithm multiple times with random starts. For our numerical examples, we randomly started the K-means and used the K-means' results for initial mean μ and variance D, and factor loading matrices B's generated from $U[0,1]$ as input to the EM. From the multiple runs, we selected the one giving the maximum penalized log-likelihood (10) as the final result for the given $(g,q,\lambda_1,\lambda_2)$.

3 RESULTS

3.1 Simulated data

We are interested in the performance of the proposed PMFA, the standard MFA as proposed by Ghahramani and Hinton (1997) and outlined in Equations (4–8), and L_1-PMND (Pan and Shen, 2007) in clustering high-dimensional data. We considered several simulation set-ups, each with 50 independent datasets. Each dataset had n observations, and each observation had K variables. For each simulated dataset, there were two clusters, with the first n_1 observations forming one cluster and the rest the other. Among all K variables, the first K_1 were informative variables, which were generated according to model (1), with $\mu_1 = 0$ for the first cluster and $\mu_2 \neq 0$ for the second, each observation having $q = 2$ loading factors (B_i for cluster i=1 or 2), and latent variables U's and error terms e's generated from $N(0, 1)$ independently; the remaining $K - K_1$ variables were noises, which were generated from $N(0, 1)$ independently across both clusters. The elements of the first 20 rows (corresponding to informative variables) of B_1 and that of B_2 were iid from $N(\sqrt{c/2}, 0.3\sqrt{c})$ and $N(\sqrt{c/4}, 0.3\sqrt{c})$, respectively, while the remaining ones (i.e. for the noise variables) were all 0. The simulation set-ups corresponded to different combinations of the values of n, n_1, K, K_1, μ_2 and c. For each training dataset, an independent tuning dataset with $n_{tu} = 100$ was generated for model selection. The predictive log-likelihood based on the tuning data was used to estimate the optimal $(\hat{g}, \hat{\lambda}_1, \hat{\lambda}_2)$ with fixed $q = 2$ for PMFA, $\hat{g}$ for MFA and $(\hat{g}, \hat{\lambda})$ for PMND, respectively.

3.1.1 Case I First we investigated the performance of the standard MFA without variable selection in clustering high-dimensional data. Three set-ups were generated with fixed $n = 100$, $n_1 = 60$, $K_1 = 20$, $\mu_2 = 6.0$ and $c = 2$, but with differing K, the total number of variables: $K = 60$, 80 and 100, respectively.

Table 1 lists the the number of datasets identified with $\hat{g}$ clusters for the three set-ups; the Rand (1971) index and adjusted Rand index (Hubert and Arabie, 1985) are used to indicate the quality of clustering results as compared with the truth. With 20 informative variables, as the number of noise variables ($K - K_1$) increased from 40 to 80, the performance of MFA deteriorated. When the number of noise variables was 40, MFA worked quite well with the Rand and adjusted Rand indices as high as 0.97 and 0.94, respectively. However, with 80 noise variables, the indices decreased dramatically to 0.60 and 0.18, respectively. This confirms the need for variable selection for high-dimensional data.

Table 1. Case I: performance of MFA for three simulation set-ups with $g = 2$ clusters and K variables, of which $K_1 = 20$ variables were informative

Cluster ($\hat{g}$)	$K = 60$	$K = 80$	$K = 100$
	N	N	N
1	2	7	41
2	48	43	9
3	0	0	0
RI/aRI	0.97/0.94	0.92/0.83	0.60/0.18

Among $n = 100$ observations, $n_1 = 60$ were in one cluster. N represents the numbers of datasets identified with $\hat{g}$ clusters; *RI* and *aRI* represent the averages of the Rand index and adjusted Rand index, respectively.

3.1.2 Case II Now we compare the performance of the proposed PMFA with that of the standard MFA and PMND, illustrating the effectiveness of penalization for variable selection and the need of using non-diagonal covariance matrices. Five simulation set-ups were explored with fixed $n = 50$, $n_1 = 30$, $K = 100$ and $K_1 = 20$, but differing μ_2 and c as follows: 1) Set-up 0: $\mu_2 = 0$ and $c = 0$; 2) Set-up 1: $\mu_2 = 4.5$ and $c = 1$; 3) Set-up 2: $\mu_2 = 4.5$ and $c = 2$; 4) Set-up 3: $\mu_2 = 6.0$ and $c = 1$; 5) Set-up 4: $\mu_2 = 6.0$ and $c = 2$. Set-up 0 was the null case with only one cluster underlying the simulated data and none of the variables was informative; Set-ups 1–4 had two clusters underlying the data and only the first 20 variables were informative.

Table 2 gives the simulation results. MFA obtained $\hat{g} = 1$ for all datasets in all set-ups, failing to uncover clustering structures for set-ups 1–4 because of no variable selection and the effects of noise variables. As expected, both PMFA and PMND correctly identified the one cluster and all noise variables in set-up 0. For set-ups 1–4, we notice that the larger the mean difference between the two clusters, or the stronger (to some extent) the correlations among variables, the more likely for the PMFA to correctly identify the two clusters. Although the simulated dataset had two true clusters, PMND tended to identify far more clusters than the truth when there were very strong correlations among variables. In addition, PMND kept much more noise variables in the final model than PMFA. For example in set-up 4, PMFA kept <8 noise variables for those datasets identified to have two or three clusters, while PMND kept $80 - 51.44 \approx 29$ noise variables.

Table 3 listed the Rand indices and adjusted Rand indices for the clusters identified by PMFA and PMND for the simulated datasets. For PMND, as c increased from 1 to 2, the adjusted Rand index decreased from 0.50 to 0.46 for $\mu_2 = 4.5$ and from 0.57 to 0.46 for $\mu_2 = 6.0$. It was reasonable since the larger the c, the larger the correlations among informative variables and thus the independence assumption (with the use of a diagonal covariance matrix) in the PMND method was more severely violated. In contrast, for PMFA, the adjusted Rand index had a different trend: as c increased from 1 to 2, the adjusted Rand index increased from 0.06 to 0.21 for $\mu_2 = 4.5$ and from 0.64 to 0.76 for $\mu_2 = 6.0$. It seems that the larger the correlations among variables, the more likely the PMFA correctly discovered underlying clustering structures, while the performance of PMND went down. In summary, the results for set-ups 1–4 demonstrated that for datasets with correlated informative variables, PMFA performed better than PMND in identifying true clustering structures.

3.1.3 Case III To investigate the effect of the sample size, we used a larger $n = 100$ with $n_1 = 60$; all other aspects were the same as in Case II. Table 4 gives the results for the five set-ups. The Rand indices and adjusted Rand indices are also provided in Table 5. The results demonstrated that PMFA worked better than PMND in identifying clustering structures with correlated variables, and that PMFA performed much better than MFA. In particular, as for Cases I & II, with the presence of many noise variables, MFA most often selected only one cluster. On the other hand, as for Case II, PMND tended to select a larger number of clusters than the truth, which can be explained by the use of a diagonal covariance matrix by PMND. Because of the independence assumption implied by the diagonal covariance matrix in PMND, the orientation (i.e. major axis) of a cluster ellipsoid identified by PMND paralleled with a coordinate axis, whereas that for the true clusters did not.

Table 2. Case II: performance of PMFA and PMND for five simulation set-ups

Method		Set-up 0			Set-up 1			Set-up 2			Set-up 3			Set-up 4		
		$\mu_2=0, c=0$			$\mu_2=4.5, c=1$			$\mu_2=4.5, c=2$			$\mu_2=6.0, c=1$			$\mu_2=6.0, c=2$		
	$\hat{g}$	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2
	1	50	20	80	37	0	34.4	33	0	38.1	15	0	44.1	10	0	60.0
	2	–	–	–	8	0	76.8	5	0	78.8	14	0	70.9	18	0	72.0
PMFA	3	–	–	–	5	0	73.0	11	0	70.9	21	0	70.9	22	0	73.6
	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1	50	20	80	–	–	–	–	–	–	–	–	–	–	–	–
	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PMND	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	4	–	–	–	5	0	59.8	–	–	–	5	0	52.8	–	–	–
	5	–	–	–	45	0	46.2	50	0	36.1	45	0	45.4	50	0	51.44

Among $K=100$ variables, $K_1=20$ were informative; among $n=50$ observations, $n_1=30$ were in one cluster. N represents the number of datasets identified with $\hat{g}$ clusters; z_1 and z_2 represent the average number of deleted informative and noise variables, respectively, among datasets identified with $\hat{g}$ clusters.

Table 3. Case II: The averages of the Rand indices and adjusted Rand indices of PMFAs and PMND for simulated datasets

Method	Set-up 1		Set-up 2		Set-up 3		Set-up 4	
	$\mu_2=4.5$ $c=1$		$\mu_2=4.5$ $c=2$		$\mu_2=6.0$ $c=1$		$\mu_2=6.0$ $c=2$	
	RI	*aRI*	*RI*	*aRI*	*RI*	*aRI*	*RI*	*aRI*
PMFA	0.54	0.06	0.61	0.21	0.82	0.64	0.88	0.76
PMND	0.75	0.50	0.73	0.46	0.79	0.57	0.73	0.46

Hence, two or more axis-parallel ellipsoids were needed in PMND to approximate a non-axis-parallel ellipsoid. Figure 1 shows the results from a representative dataset for set-up 4. Compared with Table 2, clearly both PMFA and PMND had improved performance with a larger sample size.

We also applied the penalized normal mixture model with cluster-specific diagonal covariance matrices to Set-up 3 (Xie *et al*., 2008b). As for PMND, it over-selected the number of clusters, but to a lesser degree: it selected $\hat{g}=4$ for 40 datasets with $z_1=0$ and $z_2=23.0$, while choosing $\hat{g}=5$ for the remaining 10 datasets with $z_1=0$ and $z_2=35.5$. The method not only retained more noise variables but also performed less well with a smaller average Rand index $RI=0.75$ and adjusted Rand index $aRI=0.52$.

We also considered selecting both g and q, rather than fixed $q=2$ as done before, in PMFA in two simulation set-ups. For the first one similar to set-up 2, among 50 simulated datasets, for nearly a half we correctly selected $\hat{q}=2$ (Table 6). A possible reason for incorrectly selecting q was that other incorrect $q>0$ values also led to good clustering results with high (adjusted) Rand index values. In a new set-up with more dispersed elements of the loading matrices B_1 and B_2 (simulated from two normals with SD=0.6, instead of SD=0.42 in Set-up 2 of Case III while other aspects remained the same), implying larger effects of q, we could select $\hat{q}=2$ correctly for all 50 datasets.

3.2 Real data

We applied the methods to a gene expression dataset of lung cancer patients (Beer *et al*., 2002). The original authors identified a set of genes that could predict survival in early stage lung adenocarcinoma and thus discovered a high-risk patient group who might benefit from adjuvant therapy. The data contained gene expression profiles for 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors.

To minimize the potential influence of the genes with little or no expression on any clustering algorithm, we did a preliminary gene screening by excluding any gene if the 75th percentile of its observed expression levels was <100. Then we included only the top 300 genes with the largest sample variances across the 86 samples.

We randomly divided the 86 samples into three parts for training, tuning and testing with sizes 29, 29 and 28 samples, respectively. The training dataset was used to fit the model, and the tuning dataset was used to select the tuning parameters. Finally, the selected model was applied to the test dataset to determine the cluster memberships of the test samples. For simplicity, we fixed $q=2$ for MFA and PMFA.

Clustering the data with all the 300 genes simultaneously, MFA selected only one cluster. Alternatively, we applied a two-step procedure: as in McLachlan *et al*. (2002), we first conducted a univariate gene screening before applying the MFA to the selected genes. Specifically, we applied a univariate model-based clustering on each of the 300 genes as implemented in R package `Mclust` (Fraley and Raftery, 2007): we fitted a series of normal mixture models with one to nine normal components, and selected the best model based on BIC; if a model with more than one component was selected for a gene, then the gene was retained. When applied to the training data alone, the screening yielded 190 genes; if applied to a combined training and tuning dataset, it selected 211 genes. We applied the MFA with the selected 190 or 211 genes to the training data, resulting in only one cluster in either case.

Table 4. Case III: performance of PMFA, MFA and PMND for five simulation set-ups

Method		Set-up 0			Set-up 1			Set-up 2			Set-up 3			Set-up 4		
		$\mu_2=0, c=0$			$\mu_2=4.5, c=1$			$\mu_2=4.5, c=2$			$\mu_2=6.0, c=1$			$\mu_2=6.0, c=2$		
	$\hat{g}$	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2
PMFA	1	50	20	80	4	0	34.8	–	–	–	–	–	–	–	–	–
	2	–	–	–	20	0	46.8	37	0	52.9	29	0	51.8	26	0	53.1
	3	–	–	–	26	0	57.9	13	0	58.5	21	0	59.5	24	0	67.7
	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MFA	1	50	0	0	49	0	0	49	0	0	45	0	0	41	0	0
	2	–	–	–	1	0	0	1	0	0	5	0	0	9	0	0
	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PMND	1	50	20	80	–	–	–	–	–	–	–	–	–	–	–	–
	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	5	–	–	–	50	0	35.3	50	0	30.4	50	0	28.8	50	0	53.4

Among $K=100$ variables, $K_1=20$ were informative; among $n=100$ observations, $n_1=60$ were in one cluster. N represents the number of datasets identified with $\hat{g}$ clusters; z_1 and z_2 represent the average number of deleted informative and noise variables, respectively, among datasets identified with $\hat{g}$ clusters.

Table 5. Case III: The averages of the Rand indices and adjusted Rand indices of PMFAs, MFA and PMND for simulated datasets

Method	Set-up 1		Set-up 2		Set-up 3		Set-up 4	
	$\mu_2=4.5$ $c=1$		$\mu_2=4.5$ $c=2$		$\mu_2=6.0$ $c=1$		$\mu_2=6.0$ $c=2$	
	RI	*aRI*	*RI*	*aRI*	*RI*	*aRI*	*RI*	*aRI*
PMFA	0.94	0.87	0.97	0.94	0.99	0.99	0.99	0.98
MFA	0.53	0.02	0.52	0.02	0.56	0.10	0.60	0.18
PMND	0.75	0.51	0.74	0.48	0.72	0.45	0.71	0.44

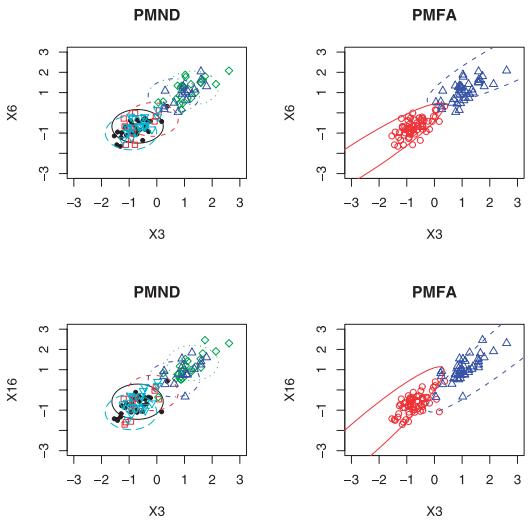

Fig. 1. Clusters identified by PMFA (right panels) and PMND (left panels) from a dataset in simulation set-up 4. Three informative variables ($X3$, $X6$ and $X16$) were plotted. The five types of the symbols represent the cluster-memberships in the five clusters identified by PMND in left panels, while the two types of the symbols represent the true cluster memberships in right panels.

We applied the normal mixture model-based clustering, as implemented in R package `Mclust`, to the training data with all the 300 genes simultaneously. Mclust only fits mixture models with various types of diagonal covariance matrices if the data dimension is larger than the sample size, as was the case here. It selected a final model with two clusters with the two diagonal covariance matrices with varying volume but equal shape (i.e. 'VEI' in Mclust notation). The final selected model was applied to the test data to yield two clusters; comparing the survival curves for the two clusters, a log-rank test gave a chi-squared statistic of 2.4 with one degree of freedom, resulting in a P-value of 0.124.

In comparison, PMFA identified two clusters with 24 and 4 samples, respectively, while PMND chose three clusters. PMFA retained 258 genes, while PMND kept 276 genes, among which 240 genes appeared in the final models of both PMFA and PMND. For PMFA, the first cluster contained four patients, two of whom died early while the other two were censored, and the remaining 24 patients consisted of cluster 2. For PMND, cluster 1 contained five patients plus the same four patients as those in cluster 1 of PMFA, the other two clusters contained 9 and 10 patients, respectively. With the patient survival data, we plotted in Figure 2 the Kaplan–Meier survival estimates of the patients in the clusters for PMFA and PMND respectively. The log-rank test was used to investigate the

Table 6. Simulation results with PMFA selecting q (and g) in two set-ups

q	Set-up 2, Case III					New set-up				
	$\#(\hat{q}=q, \hat{g}=g)$			with $(\hat{g},q)$		$\#(\hat{q}=q, \hat{g}=g)$			with $(\hat{g},q)$	
	$g=1$	2	3	RI	aRI	$g=1$	2	3	RI	aRI
0	0	0	0	0.690	0.389	0	0	0	0.774	0.552
1	0	17	9	0.979	0.958	0	0	0	0.973	0.945
2	0	18	5	0.983	0.965	0	31	19	0.994	0.989
3	0	0	1	0.970	0.939	0	0	0	0.984	0.967

RI and *aRI* were calculated with selected $\hat{g}$ and fixed q.

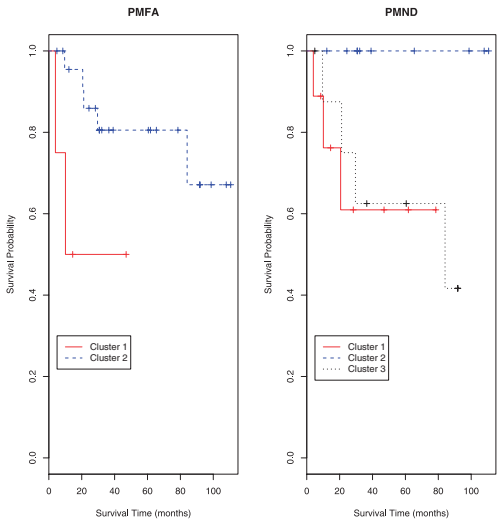

Fig. 2. Survival curves for the clusters identified by PMFA and PMND for the lung cancer data.

survival difference between/among the clusters. For PMFA, the test yielded a chi-squared statistic of 4.4 with one degree of freedom, resulting in a statistically significant P-value of 0.037. For PMND, the chi-squared test statistic was 5.5 with two degrees of freedom, leading to a P-value of 0.063, which is only marginally significant. This indicated that, compared with MFA and PMND, by accounting for possible correlations among the genes, PMFA might be more helpful to uncover the groups of cancer patients with distinct risks of mortality.

Note that a preliminary variable screening can be helpful even for a method with the capability of variable selection: in addition to saving computing time with a simple univariate variable screening, it can improve predictive performance, as theoretically shown by Fan and Lv (2008). For example, when the top 450, rather than 300, genes were used in PMFA, there was a less significant survival difference between the two clusters detected from the test data: the log-rank test gave a P-value of only 0.145.

4 DISCUSSION

We have proposed a new model-based clustering method, a PMFAs, to model non-diagonal cluster-specific covariance matrices. PMFA generalizes the usual MFAs via regularization, which can effectively realize variable selection in clustering high-dimensional data, in addition to regularizing parameter estimates and its associated benefits. Simulation studies and a microarray gene expression data application have demonstrated the utility of the proposed method and its superior performance over MFA and penalized model-based clustering with a common diagonal covariance matrix. Although the current implementation of the EM algorithm for PMFA is straightforward, it is computationally demanding, especially with the choice of the tuning parameters by a grid search. More efficient algorithms and model selection criteria will be helpful. Other possible extensions of PMFA include the following. First, although we only considered the L_1 penalization of mean parameters, other penalties, such as a grouped penalty (Wang and Zhu, 2008; Xie *et al*., 2008a) as for the loading matrix parameters considered here, can be applied. Second, it is natural to consider using general covariance matrices in the mixture model. In Zhou *et al*. (2009), we embed an unconstrained covariance matrix estimation procedure in the EM algorithm. Although the use of an unconstrained covariance matrix is more flexible than the PMFA approach, it may lose efficiency if some latent variable-induced covariance assumption holds as in the PMFA approach; numerical comparisons are needed.

ACKNOWLEDGEMENT

We thank the reviewers for helpful and constructive comments.

Funding: NIH grants (HL65462 and GM081535), and NSF grants (IIS-0328802 and DMS-0604394).

Conflict of Interest: none declared.

REFERENCES

Baek,J. and McLachlan,G.J. (2008) Mixtures of factor analyzers with common factor loadings for the clustering and visualisation of high-dimensional data. Isaac Newton Institute for Mathematical Sciences, Preprints.

. Baek,J. *et al*. (2009) Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. To appear in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Available at: http://www.maths.uq.edu.au/~gjm/bmf_pami09.pdf/

Beer,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Series B*, **39**, 1–38.

Eisen,M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Everitt,B.S. (1984) *An Introduction to Latent Variable Models.* Chapman and Hall, London.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Series B*, **70**, 849–911.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Fraley,C. and Raftery,A.E. (2007) Model-based methods of classification: using the mclust software in chemometrics. *J. Stat. Software*, **18**, paper i06. Available at http://www.jstatsoft.org/v18/i06/.

Ghahramani,Z. and Hinton,G.E. (1997) The EM algorithm for mixtures of factor analyzers. *Technical Report CRG-TR-96-1*. Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4, 1997.

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hinton,G.E. *et al*. (1997) Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Networks*, **8**, 65–74.

Huang,J.Z. *et al*. (2006) Covariance selection and estimation via penalised normal likelihood. *Biometrika*, **93**, 85–98.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

McLachlan,G.J. and Peel,D. (2000) Mixtures of factor analyzers. In Langley,P. (ed.). *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, pp. 599–606.

McLachlan,G.J. *et al*. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

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

McLachlan,G.J. *et al*. (2003) Modeling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data Analysis*, **41**, 379–388.

McLachlan,G.J. *et al*. (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Comput. Stat. Data Analysis*, **51**, 5327–5338.

Pan,W. and Shen,X. (2007) Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.*, **8**, 1145–1164.

Raftery,A.E and Dean,N. (2006) Variable selection for model-based clustering. *J. Am. Stat. Assoc.*, **101**, 168–178.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

Thalamuthu,A. *et al*. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

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

Xie,B. *et al*. (2008a) Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, **64**, 921–930.

Xie,B. *et al*. (2008b) Penalized model-based clustering with cluster-specific diagonal covariances and grouped variables. *Electron. J. Stat.*, **2**, 168–212.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B*, **68**, 49–67.

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

Zhou,H. *et al*. (2009) Penalized model-based clustering with unconstrained covariance matrices. *Electronic J. Stat.*, **3**, 1473–1496.