Variable selection in penalized model-based clustering via regularization on grouped parameters

Benhuai Xie*
Division of Biostatistics, School of Public Health, University of Minnesota
*email: benhuaix@biostat.umn.edu

and

Wei Pan*
Division of Biostatistics, School of Public Health, University of Minnesota
*email: weip@biostat.umn.edu

and

Xiaotong Shen*
School of Statistics, University of Minnesota
*email: xshen@stat.umn.edu

Summary: Penalized model-based clustering has been proposed for high-dimensional but small sample-sized data, such as arising from genomic studies; in particular, it can be used for variable selection. A new regularization scheme is proposed to group together multiple parameters of the same variable across clusters, which is shown both analytically and numerically to be more effective than the conventional $L_1$ penalty for variable selection. In addition, we develop a strategy to combine this grouping scheme with grouping structured variables. Simulation studies and applications to microarray gene expression data for cancer subtype discovery demonstrate the advantage of the new proposal over several existing approaches.

Key words: BIC; Diagonal covariance; EM algorithm; High-dimension but low-sample size; Microarray gene expression; Mixture model; Penalized likelihood.
1. Introduction

Clustering analysis plays an important role in microarray data analysis for gene function discovery (Eisen et al 1998) and disease subtype discovery (Golub et al 1999). Among various methods, model-based clustering has been applied (Li and Hong 2001; Yeung et al 2001; Ghosh et al 2002; McLachlan et al 2002). In such a high-dimensional but low-sample-sized data setting, it is necessary to conduct variable selection (Pan and Shen 2007). For example, in clustering microarray samples for cancer subtype discovery, most of the genes in the genome are likely to be non-informative to discriminating between cancer subtypes; inclusion of many such non-informative genes may mask or distort the underlying clustering structure. A common approach in practice is a two-step procedure: first a preliminary variable selection is conducted based on some ad hoc criterion, then the selected variables are used for clustering. Pan et al (2006) and Pan and Shen (2007) gave some numerical examples demonstrating possible pitfalls of such a two-step approach to variable selection, and advocated simultaneous variable selection and model fitting. A feasible approach is through regularization in model-based clustering (Pan and Shen 2007). The key idea is that, under a finite mixture of normal distributions with a common diagonal covariance matrix, for any variable, if its cluster-specific means are all equal, then this variable is non-informative to clustering. Hence, a penalty can be added to the log-likelihood to encourage an equal estimate of the mean parameters across clusters for any variable to realize variable selection. A potential drawback of the $L_1$ penalty is that it treats the mean parameters individually and separately; for a noise variable, even if most of its cluster-specific mean parameter estimates are correctly shrunken to be equal with only few others being unequal, then this variable will be deemed incorrectly to be informative. Here we propose a more efficient scheme that penalizes all the mean parameters of the same variable together, encouraging them to be
all equal, thus realizing more effective variable selection. We first focus on grouped Lasso penalty (Yuan and Lin 2006), then extend to other grouped penalties.

Largely motivated by applications to high-dimensional genomic data, variable selection in clustering has attracted increasing research attention in the last few years. Most of the recent proposals are Bayesian (Liu et al 2003; Hoff 2005, 2006; Tadesse et al 2005; Raftery and Dean 2006; Kim et al 2006), while another class is in the framework of penalized methods for either non-model-based clustering (Mangasarian and Wild 2004) or model-based clustering (Pan and Shen 2007; Wang and Zhu 2006; Xie et al 2007). Friedman and Meulman (2004) considered a more general problem in which variables and their weights being used to form a cluster may vary over the clusters. Here we propose grouped penalization methods, built on the previous work of Pan and Shen (2007) and Xie et al (2007). The main contribution here is the concept of using a grouped penalty for multiple parameters arising from the same variable, which differs from grouping variables as studied in Xie et al (2007): our proposal here is derived from a more general heuristic while grouping variables are determined by prior knowledge. When revising this article, we noticed the work of Wang and Zhu (2006), which had a similar idea but used different penalties; an advantage of our work is explicit characterization of grouping effects for a general class of penalties. Furthermore, we extend the results to clustering with cluster-specific covariance matrices; more importantly, we study a simultaneous grouping of both multiple parameters and multiple variables, which, to our knowledge, has never been investigated in the literature, in spite of its generality, e.g., its applicability to regression and classification problems.

This article is organized as follows. Section 2 reviews some existing methods, followed by the development of our new methods. Section 3 provides numerical studies with simulated data and real gene expression data. Section 4 further generalizes our proposal to the case with cluster-specific diagonal covariance matrices. We conclude with Section 5.
2. Methods

Suppose that $x_j, j = 1, 2, \cdots, n$ are $K$-dimensional observations, which have been standardized to have sample mean 0 and sample variance 1 across all $n$ observations. In Normal mixture model-based clustering, it is assumed that each observation $x_j$ follows a mixture of $g$ multivariate normal distributions,

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

where $f_i(\cdot; \theta_i)$ is the probability density function (pdf) of a Normal distribution with parameters, including mean vector $\theta_i = (\mu_{i1}, ..., \mu_{iK})'$ and a common diagonal covariance matrix $V = \text{diag}(\sigma^2_1, \cdots, \sigma^2_K)$, for the $i$th component; $\pi_i$ is the prior probability that any observation comes from component $i$. The log-likelihood for data $\{x_1, ..., x_n\}$ is $\log L(\Theta) = \sum_j \log f(x_j; \Theta)$. Maximizing $\log L(\Theta)$ yields the maximum likelihood estimator (MLE). To compute MLE, the most commonly used algorithm is the EM (Dempster et al 1977), which works on a complete-data log-likelihood

$$\log L_c(\Theta) = \sum_i \sum_j z_{ij} \log(\pi_i f_i(x_j; \theta_i)),$$

where $z_{ij}$ is the indicator of whether $x_j$ comes from component $i$.

For regularization, a penalty $p_\lambda(\Theta)$ with penalty parameter $\lambda$ is introduced according to the goal of the analysis. It yields the corresponding penalized log-likelihood and complete-data penalized log-likelihood

$$\log L_P(\Theta) = \log L(\Theta) - p_\lambda(\Theta) \quad \text{and} \quad \log L_{c,P}(\Theta) = \log L_c(\Theta) - p_\lambda(\Theta).$$

To maximize $\log L_P(\Theta)$ to obtain the maximum penalized likelihood estimator (MPLE), an EM algorithm can be derived through $\log L_{c,P}(\Theta)$. The E-step of the EM calculates the conditional expectation of $\log L_{c,P}(\Theta)$: using $\Theta^{(r)}$ to denote the estimate at iteration $r$ and treating $z_{ij}$’s as missing data, we have

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

(1)
where $\tau_{ij}$ is the posterior probability that $x_j$ comes from component $i$, and $\tau_{ij}^{(r)}$ is its estimate as given in expression (3). The M-step maximizes $Q_P$ to update the estimated $\Theta$. In the sequel, when deriving the updating formulas in the M-step, for simplicity we may suppress dependence of the estimates on iteration $r$.

### 2.1 $L_1$ penalty

Pan and Shen (2007) proposed an $L_1$ penalty for the mean parameters:

$$p_\lambda(\Theta) = \lambda \sum_{i=1}^{g} \sum_{k=1}^{K} |\mu_{ik}|.$$  \hfill (2)

They derived the following updating formulas in the EM to obtain MPLE:

$$\hat{\tau}_{ij}^{(r)} = \frac{\hat{\pi}^{(r)}_i f_i(x_j; \hat{\theta}^{(r)}_i)}{f(x_j; \Theta^{(r)})} = \frac{\hat{\pi}^{(r)}_i f_i(x_j; \hat{\theta}^{(r)}_i)}{\sum_{i=1}^{g} \hat{\pi}^{(r)}_i f_i(x_j; \hat{\theta}^{(r)}_i)},$$  \hfill (3)

$$\hat{\pi}^{(r+1)}_i = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}}{n}, \quad \hat{\sigma}^{2(r+1)}_k = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} (x_{jk} - \hat{\mu}^{(r)}_{ik})^2}{n},$$  \hfill (4)

$$\hat{\mu}^{(r+1)}_{ik} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}} \left( 1 - \frac{\lambda \hat{\sigma}^{2(r)}_k}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}^2} \right) = \hat{\mu}^{(r+1)}_{ik} \left( 1 - \frac{\lambda \hat{\sigma}^{2(r)}_k}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}^2} \right),$$  \hfill (5)

where $\hat{\mu}^{(r+1)}_{ik}$ is the MLE at iterate $r$.

Obviously, for a sufficiently large $\lambda$, we have $\hat{\mu}_{ik} = 0$. Since each variable $k$ has been standardized to have sample mean 0, if $\hat{\mu}_{1k} = \ldots = \hat{\mu}_{gk} = 0$, then variable $k$ is noninformative in terms of clustering and can be considered as a noise variable and excluded from the clustering analysis.

### 2.2 Vertical mean grouping

Within the framework of the $L_1$ penalty, each $\mu_{ik}$ is individually penalized; on the other hand, variable $k$ is regraded as a noise variable if and only if $\mu_{1k} = \ldots = \mu_{gk} = 0$. Hence, to realize more effective variable selection, it is natural to treat $\mu_{1k}, \ldots, \mu_{gk}$ as a group of parameters, constructing a penalty that encourages all of them to be exactly 0. If we view a cluster-specific mean as a row vector, the direction of the grouping on $\mu_{1k}, \ldots, \mu_{gk}$ is vertical,
hence we call it vertical mean grouping (VMG), for which we propose the following penalty:

$$p_{\lambda}(\Theta) = \lambda \sqrt{g} \sum_{k=1}^{K} \|\mu_k\|,$$

(6)

with $\mu_k = (\mu_{1k}, \mu_{2k}, \ldots, \mu_{gk})'$ and $\|\mu_k\| = \sqrt{\sum_{i=1}^{g} \mu_{ik}^2}$ for $k = 1, 2, \ldots, K$. Note that, for any vector $v$, $\|v\|$ denotes the $L_2$ norm of $v$.

The updating formulas for $\tau_{ij}$, $\pi_i$ and $\sigma_k^2$ remain the same as (3)-(4), but that for $\mu_{ik}$ is different as shown below:

**THEOREM 1:** The sufficient and necessary conditions for $\hat{\mu} = (\hat{\mu}_k), k = 1, 2, \ldots, K$ to be the unique maximizer to (1) are

$$\text{diag} \left( \sum_{j=1}^{n} \tau_{1j}, \sum_{j=1}^{n} \tau_{2j}, \ldots, \sum_{j=1}^{n} \tau_{gj} \right) (\hat{\mu}_k - \hat{\mu}_k) = \lambda \sqrt{g} \sigma_k^2 \frac{\hat{\mu}_k}{\|\hat{\mu}_k\|} \text{ if and only if } \hat{\mu}_k \neq 0, \quad (7)$$

where $\hat{\mu}_k = \left( \frac{\sum_{j=1}^{n} \tau_{1j} x_{jk}}{\sum_{j=1}^{n} \tau_{1j}}, \frac{\sum_{j=1}^{n} \tau_{2j} x_{jk}}{\sum_{j=1}^{n} \tau_{2j}}, \ldots, \frac{\sum_{j=1}^{n} \tau_{gj} x_{jk}}{\sum_{j=1}^{n} \tau_{gj}} \right)'$ has the form of the MLE, and

$$\left( \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij} x_{jk} \right)^2 \leq \lambda \sqrt{g} \sigma_k^2 \sum_{j=1}^{n} \tau_{ij} \text{ if and only if } \hat{\mu}_k = 0. \quad (8)$$

It is clear that if the inequality in (8) is satisfied, we have $\hat{\mu}_{1k} = \ldots = \hat{\mu}_{gk} = 0$, thus variable $k$ is regarded as a noise variable. Next we highlight a key difference between (8) and (5).

From (5), we have

$$|\hat{\mu}_{ik}| \leq \frac{\lambda \sigma_k^2}{\sum_{j=1}^{n} \tau_{ij}} \text{ if and only if } \hat{\mu}_{ik} = 0,$$

while (8) can be rewritten as

$$\sqrt{\frac{\sum_{i=1}^{g} \tilde{\mu}_{ik}^2}{g}} \leq \frac{\lambda \sigma_k^2}{\sum_{j=1}^{n} \tau_{ij}} \text{ if and only if } \hat{\mu}_k = 0.$$

Hence, according to (8), if most of the components of $\tilde{\mu}_k$ are small (so that the quadratic mean of the components is less than the threshold), then we will have $\hat{\mu}_k = 0$; in contrast, by (5), some larger components of $\tilde{\mu}_k$ may remain to be non-zero. This highlights the different consequences of using the $L_1$ penalty and the grouped penalty; in particular, the effect of the grouped penalty is its tendency of realizing $\hat{\mu}_{1k} = \ldots = \hat{\mu}_{gk} = 0$ simultaneously, thus more effective variable selection.
Combining (7) and (8) yields
\[
\hat{\mu}_k = \left( \text{sign} \left( 1 - \frac{\lambda \sqrt{g\sigma_k^2}}{(\sum_{i=1}^{g}(\sum_{j=1}^{n} \tau_{ij}x_{jk})^2)^{1/2}} \right) \right) \nu_k \tilde{\mu}_k
\]
(9)
with \( \nu_k = \text{diag} \left( 1 + \frac{\lambda \sqrt{\sigma_k^2}}{\|\tilde{\mu}_k\| \sum_{j=1}^{n} \tau_{ij}}, 1 + \frac{\lambda \sqrt{\sigma_k^2}}{\|\tilde{\mu}_k\| \sum_{j=1}^{n} \tau_{ij}}, \cdots, 1 + \frac{\lambda \sqrt{\sigma_k^2}}{\|\tilde{\mu}_k\| \sum_{j=1}^{n} \tau_{ij}} \right)^{-1} \).

Equation (9) naturally suggests an iterative algorithm to update \( \hat{\mu}_k \). However, in simulations, we found that it did not work well. As an alternative, we used (8) to derive whether \( \hat{\mu}_k = 0 \); if not, we tried the following two methods. First, we rewrote (7) iteratively as
\[
\hat{\mu}_{ik}^{(r+1)} = \hat{\mu}_{ik}^{(r)} - \lambda \sqrt{g\sigma_k^2} \hat{\mu}_{ik}^{(r)} / \left( \sum_{j=1}^{n} \tau_{ij} \|\hat{\mu}_k^{(r)}\| \right),
\]
and then updated the components of \( \mu_k \) one by one. Second, we applied a Newton algorithm directly to iteratively solve (7). Although both methods worked better than using (9), the Newton method seemed best and was used in all numerical examples. Note that, if we use a damped Newton method or any method with guaranteed convergence in the M-step (e.g. by imposing a weak condition that an iterate stops if \( Q_P \) is increased), then the whole EM algorithm converges (Wu 1983).

### 2.3 Horizontal mean grouping

In some applications, we may have prior knowledge that a group of variables are likely to be all together as either informative or not; for instance, all the genes in a biological pathway are either relevant or irrelevant to a disease, depending on whether the pathway is involved in the pathology of the disease. Xie et al (2007) proposed a grouped penalty function to incorporate such prior knowledge; they focused on the case with cluster-specific covariance matrices. Here we give details of the case with a common covariance matrix.

Suppose that the variables can be grouped. Without loss of generality, we assume that \( \mu_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iK})' = (\mu_{i1}', \mu_{i2}', \cdots, \mu_{iM}')' \) with each \( \mu_{im} \) corresponding to a group of variables; \( \text{dim}(\mu_{im}) = k_m \) and \( \sum_{m=1}^{M} k_m = K \). Correspondingly, the covariance matrix is partitioned into \( V = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2) = \text{diag}(V_1, V_2, \cdots, V_M) \) with \( V_m \) as a \( k_m \times k_m \)
matrix. Other vectors are partitioned accordingly: for example, \( x_j = (x_{j1}', x_{j2}', \cdots, x_{jM}')' \). As in Xie et al (2007), we propose a grouped penalty:

\[
p_\lambda(\Theta) = \lambda \sum_{i=1}^g \sum_{m=1}^M \sqrt{k_m} \| \mu^m_i \|; \tag{10}\]

As in the vertical mean grouping, if we view a cluster-specific mean as a row vector, the direction of the grouping on the elements of \( \mu^m_i \) is horizontal, hence we call it horizontal mean grouping (HMG).

For HMG, the updating formulas for \( \tau_{ij} \), \( \pi_i \) and \( \sigma_k^2 \) remain the same as (3)-(4); we only need to derive that for \( \mu_{ik} \), as given below:

**Theorem 2**: The sufficient and necessary conditions for \( \hat{\mu} = (\hat{\mu}^m_i) \) to be the unique maximizer of (1) are

\[
(\sum_{j=1}^n \tau_{ij})V_m^{-1}(\hat{\mu}^m_i - \tilde{\mu}^m_i) = \lambda \sqrt{k_m} \hat{\mu}^m_i \| \hat{\mu}^m_i \| \text{ if and only if } \hat{\mu}^m_i \neq 0, \tag{11}\]

\[
\left\| \sum_{j=1}^n \tau_{ij} x_j^m V_m^{-1} \right\| \leq \lambda \sqrt{k_m} \text{ if and only if } \hat{\mu}^m_i = 0, \tag{12}\]

where \( \tilde{\mu}^m_i = \sum_{j=1}^n \tau_{ij} x_j^m / \sum_{j=1}^n \tau_{ij} \) has the form of the MLE.

Suppose that \( J_m \) is the index set of the variables in group \( m \). Expression (12) can be rewritten as

\[
\sqrt{\frac{1}{k_m} \sum_{k \in J_m} \frac{\hat{\mu}_{ik}^2}{\sigma_k^2}} \leq \frac{\lambda}{\sum_{j=1}^n \tau_{ij}} \text{ if and only if } \hat{\mu}^m_i = 0.
\]

Hence, if the average (based on a weighted quadratic mean) of the components of \( \hat{\mu}^m_i \) is small enough, all the components of \( \hat{\mu}^m_i \) are shrunk to be exactly zero; this is similar to that in the vertical grouping, highlighting the effect of grouping and its key difference from its counterpart with the \( L_1 \) penalty.

Easily (11) and (12) can be combined as

\[
\hat{\mu}^m_i = \left( \text{sign} \left( 1 - \frac{\lambda \sqrt{k_m}}{\sum_{j=1}^n \tau_{ij} x_j^m V_m^{-1}} \right) \right) \nu_m^m \tilde{\mu}^m_i, \tag{13}\]
where ν_m^i = \left( I + \frac{\lambda \sqrt{\tau_i}}{\sum_{j=1}^{\tau_j} \tau_i \| \mu_i \|} V_m \right)^{-1} and I is the identity matrix. Equation (13) suggests an iterative algorithm to update \hat{\mu}_m^i, which was used in all the numerical examples.

2.4 Combining horizontal and vertical mean groupings

We combine the heuristics for the vertical grouping with the prior knowledge for the horizontal grouping, resulting in a penalty for vertical and horizontal mean groupings (VHMG):

\[ p_\lambda(\Theta) = \lambda \sum_{m=1}^{M} \sqrt{gk_m} \| \mu^m \| \]

with \( \mu^m = (\mu_1^m, \mu_2^m, \ldots, \mu_g^m)' \).

Again the updating formulas for \( \tau_{ij}, \pi_i \) and \( \sigma_k^2 \) remain the same as (3)-(4), and we only need to derive that for \( \mu_{ik} \), as given below:

**Theorem 3:** The sufficient and necessary conditions for \( \hat{\mu} = (\hat{\mu}_m^i) \) to be the unique maximizer of (1) are

\[ (\sum_{j=1}^{n} \tau_{ij}) V_m^{-1} (\hat{\mu}_m^i - \mu_i^m) = \lambda \sqrt{gk_m} \| \hat{\mu}_m^i \| \]

if and only if \( \hat{\mu}_m^i \neq 0 \)

for all \( i = 1, 2, \ldots, g \), and

\[ d_m \leq \lambda \sqrt{gk_m} \]

if and only if \( \hat{\mu}_m^i = 0 \),

where \( \hat{\mu}_m^i = \frac{\sum_{j=1}^{n} \tau_{ij} x_j^m}{\sum_{j=1}^{n} \tau_{ij}} \) has the form of the MLE, and

\[ d_m = \left\| \left( \sum_{j=1}^{n} \tau_{ij} x_j^m V_m^{-1}, \sum_{j=1}^{n} \tau_{ij} x_j^m V_m^{-1}, \ldots, \sum_{j=1}^{n} \tau_{ij} x_j^m V_m^{-1} \right) \right\|. \]

Conditions (15) and (16) yield

\[ \hat{\mu}_m^i = \left( \text{sign} \left( 1 - \frac{\lambda \sqrt{gk_m}}{d_m} \right) \right) + \nu_m^i \hat{\mu}_m^i. \]

Although (17) suggests an iterative algorithm, we found that the below one performed better and was used in our numerical examples:

\[ \hat{\mu}_i^{m,(r+1)} = \left( \text{sign} \left( 1 - \frac{\lambda \sqrt{gk_m}}{d_m} \right) \right) + \left( \hat{\mu}_i^{m,(r)} - \frac{\lambda \sqrt{gk_m}}{\sum_{j=1}^{\tau_i} \tau_{ij} V_m^{-1}} \| \hat{\mu}_i^{m,(r)} \| \right). \]

2.5 Other grouped penalties

Zhao et al (2006) proposed a general class of grouped penalties called Composite Absolute
Penalties (CAP) for linear regression, which can be also used for VMG penalty function

\[
p_{\lambda}(\Theta) = \lambda \left( \sum_{k=1}^{K} g^{\gamma_0/\gamma_k} \| \mu_k \|_{\gamma_k}^{\gamma_0} \right)^{1/\gamma_0},
\]

with \( \| \mu_k \|_{\gamma_k} = (\sum_{i=1}^{g} |\mu_{ik}|^{\gamma_k})^{1/\gamma_k} \) as the \( L_{\gamma_k} \)-norm, \( 1/\gamma_k + 1/\gamma_k' = 1 \) and \( \gamma_k > 1 \) for \( k = 1, 2, \ldots, K \). It is clear that our previous grouped penalty is a special case with \( \gamma_0 = 1 \) and \( \gamma_k = 2 \), while the \( L_{\infty} \)-norm penalty of Wang and Zhu (2006) corresponds to \( \gamma_0 = 1 \) and \( \gamma_k = \infty \).

As discussed in Zhao et al (2006), \( \gamma_0 < 1 \) promotes sparsity of non-zero groups, but \( \gamma_0 < 1 \) is non-convex. On the other hand, a larger \( \gamma_0 > 1 \) encourages group similarity. The interpretation of \( \gamma_k \) is similar, only with the “group” replaced by individual parameters. Hence, typically \( \gamma_0 = 1 \) is used to maintain both sparsity and convexity, while the specific choice of \( \gamma_k \) with \( \gamma_k > 1 \) depends on the prior belief of similarity among the individual parameters. If we use the CAP with \( \gamma_0 = 1 \) while allowing \( \gamma_k \) to be any positive number, the updating formulas for \( \tau_{ij}, \pi_i \) and \( \sigma_k^2 \) remain the same as before, while that for \( \mu_{ik} \) is given in Theorem 4 of Section 4 with \( \sigma_{ik} \) replaced by \( \sigma_k \) for any \( i = 1, \ldots, g \).

Recently Choi and Zhu (2006) and Zhou and Zhu (2007) have proposed a group hierarchical penalty for linear regression. As shown by Wang and Zhu (2007), penalized model-based clustering with such a hierarchical penalty performed similarly to that with an \( L_{\infty} \)-penalty, which is a special case of CAP. We considered an application to simulated data of CAP with \( \gamma_k = 4 \), which, as illustrated in Figure 1 of Zhao et al (2006), is close to the \( L_{\infty} \)-penalty.

2.6 Model selection

Following Pan and Shen (2006) and Pan et al (2006), we adopt a modified BIC as the model selection criterion to account for regularization,

\[
BIC = -2 \log L(\hat{\Theta}) + \log(n)d_e
\]

where \( d_e = g + K + gK - 1 - q \) is the effective number of parameters with \( q = \# \{ (i, k) : \mu_{ik} = 0 \} \), the number of mean parameters shrunk to be exactly zero. The idea with \( d_e \) was
borrowed from Efron et al (2004) and Zou et al (2004), who studied the issue in the context of penalized regression. This modified BIC is used to select the number of clusters $g$ and the penalization parameter $\lambda$ jointly. Through a grid search, the optimal $(g, \hat{\lambda})$ is chosen to be the one with the minimal BIC.

For any given $(g, \lambda)$, we run an EM algorithm multiple times with random starts to obtain multiple local maxima; for our numerical examples, $K$-means results from random starts were used as inputs to the EM. From the multiple runs, we selected $\hat{\Theta}$ giving the highest values of $\log L_p(\hat{\Theta})$ as the final solution for a given pair of $(g, \lambda)$.

2.7 A supervised learning approach

A supervised learning method was discussed by Hastie et al (2001, section 14.2.4) for density estimation; it can be also used for variable selection in clustering, as suggested by the AE. Briefly, a synthetic dataset of size equal to that of the original data was generated, with each element iid from a reference distribution $g_0(x)$, which is the noise distribution $N(0,1)$ in the current context; the synthetic data were combined with the original data; a response variable $Y$ is created with value 0 for synthetic data and value 1 for the original data. Then gene selection can be conducted under a regression model by regressing $Y$ on the genes with the combined data. In our numerical examples, we used LASSO and penalized logistic regression with the $L_1$ penalty realizing variable selection.

Two main advantages of the supervised approach include its nonparametric nature and the applicability of many well-studied variable selection methods for regression. However, because the supervised approach mainly focuses on exploring the difference of the mean parameters between the reference cluster/distribution and other clusters that contributes to estimating $\beta_k$, while failing to exploit other distributional differences, it may have low efficiency; this point was confirmed by our numerical examples. Furthermore, it depends on
the correct specification of a regression model, which may not be straightforward. Finally, it also requires a parametric assumption on the reference distribution.

3. Results

3.1 Simulated data

3.1.1 Simulation set-ups. We conducted simulation studies to investigate the effectiveness of the vertical mean grouping. For comparison, we also considered the standard method without penalization and penalized methods with other forms of penalty. Only a common covariance across clusters was examined. For each simulation set-up, we generated 100 simulated datasets; each dataset contained 100 observations drawn from one or two clusters; each observation had dimension $K = 300$. For the null case with only one cluster in set-up 1, we generated each variable in each observation independently from $N(0, 1)$. For each of the other non-null set-ups, 80 observations came from one cluster while the remaining 20 observations from the other cluster; the first $K_1$ variables were informative, generated independently from $N(0, 1)$ for the first cluster and from $N(\mu_1, 1)$ with $\mu_1 \neq 0$ for the second cluster; the remaining $K - K_1$ variables were noises, all generated independently from $N(0, 1)$. Simulation set-ups 2 and 3 corresponded to $\mu_1 = 1.5$, and $K_1 = 5$ and $K_1 = 10$ respectively, while set-up 4 corresponded to $\mu_1 = 1.25$ and $K_1 = 10$. Each simulated dataset in set-up 5 three clusters with 50, 20 and 30 observations respectively; the first $K_1$ variables were informative, distributed as $N(0, 1)$, $N(1.25, 1)$ and $N(-1.25, 1)$ for the three clusters respectively, while the remaining $K - K_1$ variables were noises, generated independently from $N(0, 1)$; we used $K_1 = 10$.

For each simulated dataset, we fitted a series of models with the numbers of components $g = 1, 2$ and $3$ for the first four set-ups while up to 4 for set-up 5, and various values of penalization parameter $\lambda$. For comparison, we considered the standard method with $\lambda = 0$, the $L_1$ penalization method, the vertical mean grouping penalization method, the horizontal
mean grouping penalization method, and both vertical and horizontal groupings penalization method. In the horizontal grouping, the group size was 5 with each group consisting of either noise or informative variables only. The BIC was used to select $g$ for the standard method without penalization, while the modified BIC was used to select both $g$ and $\lambda$.

### 3.1.2 Simulation results

The results are detailed in Table 1. First, we consider selecting the correct number of the clusters. All methods correctly selected $g = 1$ in set-up 1, a null case. For the other three set-ups with $g = 2$ clusters, 1) the standard method without variable selection performed worst, indicating the necessity of variable selection in presence of a large number of noise variables; 2) the three grouping methods improved over that with the $L_1$ penalty, confirming the importance of using the heuristics of VMG and the prior knowledge in HMG; 3) overall, the vertical grouping worked best. It was somewhat surprising that VMG performed better than HMG, given that the latter used the correct and specific knowledge on the grouping of variables while the former depended only on the general heuristics. The performance of VHMG was mixed: among the three grouping methods, it could be the best for set-up 2, the second for set-up 3, and the worst for set-up 4.

In terms of variable selection for $g = 2$, HMG and VHMG were two winners with the mean selected numbers of the variables almost the same as the true values of $z_1$ and $z_2$, whereas VMG followed closely; all the three grouping methods performed better than their counterpart with the $L_1$ penalty.

[Table 1 about here.]

We reached a similar conclusion for set-up 5 with true $g = 3$. In terms of selecting the correct number of clusters, VMG performed best, followed by HMG and VHMG, then $L_1$-penalization, and finally the standard method without variable selection. For variable selection, it seemed that VHMG worked best, closely followed in the order of HMG, VMG and $L_1$-penalization. A detailed analysis revealed that, although VHMG over-selected $g$, it
did not assign any observation to one of the four clusters if \( \hat{g} = 4 \); the problem might be due to model selection with BIC, or computational challenge to finding true MPLE. Overall, VHMG gave the highest average Rand index (Rand 1971) at 0.88, while HMG, VMG, \( L_1 \)-penalization gave 0.80, 0.79 and 0.75 respectively. The details of the observation assignments and Rand index values of the methods were available from Web Tables 1-3 of Supplementary Materials.

3.1.3 Performance of other methods. We applied the supervised approach to variable selection for set-up 3. We generated a synthetic dataset with 100 samples and 300 iid \( N(0,1) \) variables. We used both LASSO (Tibshirani 1996) and penalized logistic regression (Park and Hastie 2006) implemented in R packages \textit{lars} and \textit{glmpath} respectively. Ten-fold cross-validation (CV) was used for variable selection in penalized regression; for LASSO, both squared errors and mis-classification errors were used as the criterion in CV. Finally we fitted the standard mixture model (in R package \textit{mclust02}) to the original data with only selected informative variables, and the number of clusters \( g \) was obtained based on BIC. However, the methods did not work well because they tended to delete many informative variables while keeping a large number of noise variables (Table 2).

We also considered the performance of CAP with \( \gamma_0 = 1 \) and \( \gamma_k = 4 \) for all \( k \) for VMG. It gave slightly better performance (Table 2) than that of \( \gamma_0 = 1 \) and \( \gamma_k = 2 \), though both gave almost perfect results.

[Table 2 about here.]

3.2 Real data

3.2.1 Leukemia data. A leukemia gene expression dataset (Golub et al 1999) was used to demonstrate the utility of the proposed vertical mean grouping method and its superior performance over the standard and \( L_1 \) penalized methods. The data contained 38 observations, each from a leukemia patient with his/her biological sample arrayed. Among the 38 samples,
27 were acute myeloid leukemia (AML) while the remaining 11 were acute lymphoblastic leukemia (ALL); the 27 AML samples could be further categorized into two subtypes: 8 T-cell and 19 B-cell samples. For each sample, the expression levels of $K = 7129$ genes were measured. We pre-processed the data following Dudoit et al (2002). Then we pre-selected only the top 2000 genes with the largest sample variances across the 38 samples. Finally, each array was standardized to have mean zero and standard deviation one across the genes, then each gene was standardized to have mean zero and standard deviation one across the samples.

For the horizontal grouping, the top 2000 genes were grouped according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000). About 46 percent of the 2000 genes were annotated in at least one of the 113 KEGG pathways. If a gene was annotated in two or more pathways, it was randomly assigned to one of them. The 113 KEGG pathway groups had the largest size 81, smallest size 1 and median size 4. About three quarters of the groups had sizes less than 9. Any unannotated gene formed its own group with group size 1.

Table 3 shows the clustering results. It is clear that VMG and HVMG performed best in discriminating the three subtypes of the samples, followed by HMG and $L_1$ methods; in contrast, the standard method did not work well. A problem with HMG and VHMG was that both yielded an empty cluster with no sample assigned into it. Although in theory an empty cluster might account for the non-normality of the other components, as shown in simulations, it was more likely that the two methods incorrectly selected a too large number of clusters.

In contrast to the 2000 genes used by the standard method, the penalized methods used fewer genes with variable selection: the $L_1$ penalty, VMG, HMG and VHMG methods used only 1281, 426, 504 and 54 genes. Using fewer genes not only helps uncover clustering
structures underlying the data, but also facilitates interpreting the results and shedding light on which genes are potentially involved in the underlying biology.

[Table 3 about here.]

Finally, we considered the supervised approach with LASSO and penalized logistic regression to selecting informative genes. Surprisingly, both regression methods identified all the genes to be non-informative, thus no clusters were obtained.

3.2.2 DLBCL data. We considered an application of clustering analysis to discovering molecular subtypes of diffuse large B-cell lymphoma (DLBCL) that were associated with survival outcomes. We used the data of Shipp et al (2002) with 58 patients. We pre-processed the Affymetrix gene expression data in the following steps. First, as in Shipp et al (2002), the observed expression levels were truncated below at 20 and above at 16000. Second, as a filtering step, we only kept the genes with a relative variation greater than 3 folds and an absolute variation greater than 100 units across the 58 samples. Third, we took a log transformation. Finally, we selected the top 500 genes with the largest sample variances across the 58 samples. Before clustering analysis, the data with 500 genes were standardized to have mean 0 and standard deviation 1 across the genes and then across the samples.

An application of the standard method, $L_1$-penalization, VMG, HMG and VHMG yielded 4, 6, 3, 5 and 4 clusters, with 500, 483, 423, 368 and 177 selected genes, respectively. As before we used KEGG pathways to group variables in HMG and VHMG. Based on the patient survival data, we plotted in Figure 1 the Kaplan-Meier survival estimates of the clusters for each method. Note that VHMG again gave an empty cluster with no patient assigned to it, hence only three survival curves. Using the logrank test to investigate the survival difference, we obtained the p-values of 0.20, 0.26, 0.06, 0.27 and 0.08 for the standard method, $L_1$, VMG, HMG and VHMG methods respectively. Hence, only VMG and VHMG discovered subgroups of the patients with marginally significant survival differences. In particular, for the three
groups identified by VMG, the survival probabilities at the end of year 5 were 73%, 39% and 21% respectively.

[Figure 1 about here.]

4. An extension

So far we have restricted attention to the mixture model with a common diagonal covariance matrix, which however may not hold in practice. Here we extend the method to the case with cluster-specific diagonal covariance matrices $V_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \cdots, \sigma_{iK}^2)$. For such a model, a key to realizing variable selection is to regularize variance parameters $\sigma_{ik}$, in addition to mean parameters $\mu_{ik}$ (Xie et al. 2007).

We propose applying the following penalty with two CAP components:

$$p_{\lambda_1, \lambda_2} (\Theta) = \lambda_1 \left( \sum_{k=1}^{K} g^{\gamma_k/\gamma_0} \| \mu_k \|_{\gamma_k}^{\gamma_0} \right)^{1/\gamma_0} + \lambda_2 \left( \sum_{k=1}^{K} g^{\gamma_k/\gamma_0} \| \sigma_{ik}^2 - 1 \|_{\gamma_k}^{\gamma_0} \right)^{1/\gamma_0} .$$

(19)

We only consider the common case with $\gamma_0 = 1$. The updating formulas for $\tau_{ij}$ and $\pi_i$ are the same as before; we only need to derive the updating formulas for $\sigma_{ik}^2$ and $\mu_{ik}$.

4.1 Vertical mean grouping

As shown in Web Appendix of Supplementary Materials, we can prove the below theorem:

**Theorem 4:** The sufficient and necessary conditions for any $\hat{\mu} = (\hat{\mu}_k)$ to be the unique maximizer of (1) are

$$\text{diag} \left( \sum_{j=1}^{n} \tau_{ij} \right) (\hat{\mu}_k - \mu_k) = \lambda_1 \text{diag}(\sigma_{ik}^2) g^{1/\gamma_k} \text{sign}(\hat{\mu}_k)|\hat{\mu}_k|^{\gamma_k-1} \| \hat{\mu}_k \|_{\gamma_k}^{\gamma_k-1} \text{ if and only if } \hat{\mu}_k \neq 0 .$$

(20)

where $\mu_k = (\sum_{j=1}^{n} \tau_{ij} x_{jk}, \sum_{j=1}^{n} \tau_{2j} x_{jk}, \cdots, \sum_{j=1}^{n} \tau_{gj} x_{jk})'$ has the form of the MLE, $\sum_{j=1}^{n} \tau_{ij} = \left( \sum_{j=1}^{n} \tau_{1j}, \sum_{j=1}^{n} \tau_{2j}, \cdots, \sum_{j=1}^{n} \tau_{gj} \right)$, $\sigma_k^2 = (\sigma_{1k}^2, \sigma_{2k}^2, \cdots, \sigma_{gk}^2)$, and

$$\left( \sum_{i=1}^{g} |\sigma_{ik}^{-2} \sum_{j=1}^{n} \tau_{ij} x_{jk} |^{\gamma_k} \right)^{1/\gamma_k} \leq \lambda_1 g^{1/\gamma_k} \text{ if and only if } \hat{\mu}_k = 0 .$$

(21)

4.2 Vertical variance grouping
Since the objective function (1) may not be convex in $\sigma^2_{ik}$'s, we only have a weaker result:

**Theorem 5:** The necessary condition for $\hat{\sigma}^2_k \neq 1$ to be a local maximizer of $Q_P$ is

$$
\sum_{j=1}^{n} \tau_{ij} \left( -\frac{1}{2\hat{\sigma}^2_{ik}} + \frac{(x_{jk} - \mu_{ik})^2}{2\hat{\sigma}^2_{ik}} \right) = \lambda_2 g^{1/\gamma_k} \frac{\text{sign}(\hat{\sigma}^2_k - 1) \|\hat{\sigma}^2_k - 1\|_{\gamma_k - 1}}{\|\hat{\sigma}^2_k - 1\|_{\gamma_k - 1}}
$$

for $i = 1, 2, \cdots, g$. On the other hand, the sufficient and necessary condition for $\hat{\sigma}^2_k = 1$ to be a local maximizer of $Q_P$ is

$$
\begin{cases}
\|\sum_{j=1}^{n} \tau_{j} \left( -\frac{1}{2} + \frac{(x_{jk} - \mu_{ik})^2}{2} \right)\|_{\gamma_k'} & \leq \lambda_2 g^{1/\gamma_k'}, \text{ if } \sum_{j} \tau_{j} (1/2 - (x_{jk} - \mu_{ik})^2) \geq 0; \\
\|\sum_{j=1}^{n} \tau_{j} \left( -\frac{1}{2} + \frac{(x_{jk} - \mu_{ik})^2}{2} \right)\|_{\gamma_k'} & < \lambda_2 g^{1/\gamma_k'}, \text{ otherwise}.
\end{cases}
$$

5. Discussion

In this article, we have proposed grouping the parameters of a variable and using a corresponding penalty to realize more effective variable selection in model-based clustering for high dimensional data. In addition, we combine this idea with grouping variables when there is prior knowledge that some variables work in groups. Analytical and numerical comparisons with the standard $L_1$-penalty that treats the parameters or variables individually have established superior performance of the proposed methods. The idea of grouping parameters together to realize more effective model regularization is general. For example, it can be applied to supervised or semi-supervised multi-class learning: the parameters induced by the same variable (or a group of variables) across classes can be grouped together and a corresponding grouped penalty can be used; Wang and Zhu (2007) considered grouping parameters in a shrunken centroid classifier. It would be interesting to see applications with grouped penalties for both parameters and variables, as in our proposed VHMG, to other classifiers.

**Supplementary Materials**

Proofs of the theorems and more details on simulation results are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.
Acknowledgement

The authors thank the two referees, an AE and the Editor (N. Wang) for constructive and helpful comments that led to a substantial improvement; in particular, sections 2.5, 2.7, 3.2.2 and simulation set-up 5 were added under the suggestions of the reviewers. BX and WP were partially supported by NIH grant HL65462 and a UM AHC Faculty Research Development grant, XS by NSF grants IIS-0328802 and DMS-0604394, and WP and XS by NIH grant GM081535.

References


Submitted May 24, 2007; revised October 1, 2007
Figure 1. Survival curves for the clusters identified by the methods.
### Table 1

Simulation results: the frequencies (N) of selecting $\hat{g}$ and the average numbers of selected informative variables ($\hat{z}_1$) and selected non-informative variables ($\hat{z}_2$). For set-up 1, the truths are $g = 1$, $z_1 = 10$, $z_2 = 290$ and $\mu_1 = 0.0$; for set-up 2, $g = 2$, $z_1 = 0$, $z_2 = 295$ and $\mu_1 = 1.5$; for set-up 3, $g = 2$, $z_1 = 0$, $z_2 = 290$ and $\mu_1 = 1.5$; for set-up 4, $g = 2$, $z_1 = 10$, $z_2 = 290$ and $\mu_1 = 1.25$; for set-up 5, $g = 3$, $z_1 = 0$, $z_2 = 290$.

<table>
<thead>
<tr>
<th>Set-up</th>
<th>$\hat{g}$</th>
<th>$\hat{g} = 0$</th>
<th>$\hat{g} = 1$</th>
<th>$\hat{g} = 2$</th>
<th>$\hat{g} = 3$</th>
<th>$\hat{g} = 4$</th>
<th>$\hat{g} = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\lambda = 0$</td>
<td>N</td>
<td>$L_1$</td>
<td>$L_2$</td>
<td>VMG</td>
<td>HMG</td>
<td>VHMG</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>10.0</td>
<td>290.0</td>
<td>100</td>
<td>10.0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>5.0</td>
<td>295.0</td>
<td>78</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100</td>
<td>18</td>
<td>10.0</td>
<td>290.0</td>
<td>1</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>100</td>
<td>92</td>
<td>10.0</td>
<td>290.0</td>
<td>45</td>
<td>10.0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

...
Table 2

Performance of other methods for simulation set-up 3. Two criteria, mean squared error (MSE) and mis-classifications (misclass.), were respectively used in LASSO for tuning parameter selection.

<table>
<thead>
<tr>
<th>Set-up</th>
<th>( \hat{g} )</th>
<th>( N )</th>
<th>( \hat{z}_1 )</th>
<th>( \hat{z}_2 )</th>
<th>( N )</th>
<th>( \hat{z}_1 )</th>
<th>( \hat{z}_2 )</th>
<th>( N )</th>
<th>( \hat{z}_1 )</th>
<th>( \hat{z}_2 )</th>
<th>( N )</th>
<th>( \hat{z}_1 )</th>
<th>( \hat{z}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>6.1</td>
<td>284.2</td>
<td>95</td>
<td>5.7</td>
<td>244.7</td>
<td>100</td>
<td>4.8</td>
<td>219.2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>7.5</td>
<td>273.9</td>
<td>5</td>
<td>6.8</td>
<td>287.2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>0.1</td>
<td>288.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3

*Clustering results for Golub’s data.*

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\lambda = 0$</th>
<th>$L_1$</th>
<th>VMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>#genes</td>
<td>2000</td>
<td>1281</td>
<td>426</td>
</tr>
<tr>
<td>Clusters</td>
<td>1 2</td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samples (#)</th>
<th>ALL-T (8)</th>
<th>ALL-B (19)</th>
<th>AML (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL-T (8)</td>
<td>0 8</td>
<td>3 16</td>
<td>11 0</td>
</tr>
<tr>
<td>ALL-B (19)</td>
<td>0 0 0 0 0 0 8</td>
<td>1 0 1 0 17 0</td>
<td>0 6 0 1 4 0</td>
</tr>
<tr>
<td>AML (11)</td>
<td>8 0 0 0 6 0 0 0 0</td>
<td>1 1 1 8 0</td>
<td>0 11 0 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>HMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>#genes</td>
<td>504</td>
</tr>
<tr>
<td>Clusters</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samples (#)</th>
<th>ALL-T (8)</th>
<th>ALL-B (19)</th>
<th>AML (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL-T (8)</td>
<td>0 2 0 0 6 0 0 0 0</td>
<td>8 0 0 0</td>
<td></td>
</tr>
<tr>
<td>ALL-B (19)</td>
<td>0 5 1 8 0 4 1 0 0</td>
<td>0 1 18 0</td>
<td></td>
</tr>
<tr>
<td>AML (11)</td>
<td>7 1 0 0 0 0 3 0</td>
<td>0 11 0 0</td>
<td></td>
</tr>
</tbody>
</table>