May 2007

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 an application to microarray gene expression data for cancer subtype discovery demonstrate the advantage of the new proposal over existing approaches.

Key words: BIC; Diagonal covariance; EM algorithm; High-dimension but low-sample size; L_1 penalization; 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 taken 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 penalizing all the mean parameters of the same variable together, encouraging them to be *all* equal, thus realizing more effective variable selection. The proposed penalty is similar to that for grouped variables in Lasso

(Yuan and Lin 2006) and in penalized model-based clustering (Xie et al 2007). However, we emphasize that the former is for parameters of the same variable across clusters (or classes in supervised learning), not parameters for grouped variables as in Yuan and Lin (2006) and Xie et al (2007). In fact, as to be shown, these two grouping schemes can be combined.

Analogous to that for naive Bayes and recent developments in supervised learning (e.g., Tibshirani et al 2003; Bickel and Levina 2004), it has been argued that it is more effective to work with an independence model involving diagonal covariance matrices in model-based clustering analysis for high-dimensional data (Fraley and Raftery 2006; Pan and Shen 2007). Penalized model-based clustering with a common diagonal covariance matrix (Pan and Shen 2007) and that with cluster-specific diagonal covariance matrices (Xie et al 2007) have been investigated. Here, following the same line of arguments, we restrict our attention to a common diagonal covariance matrix; an extension to cluster-specific diagonal covariance matrices will be discussed at the end.

This article is organized as follows. Section 2 first reviews the L_1 penalization method of Pan and Shen (2007) and one for grouped variables of Xie et al (2007), then develops a new method that groups the multiple mean parameters of the same variable together. In addition, the two grouping schemes on parameters and variables respectively are proposed to be combined. Section 3 provides simulation studies and a gene expression data analysis for leukemia subtype discovery, demonstrating the utility and advantage of our proposed methods over existing approaches. Section 4 further generalizes the proposed method to the case with cluster-specific diagonal covariance matrices. Section 5 summarizes the main points and outlines some future work.

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 $f_i(.;\theta_i)$ is the probability density function (pdf) of the Normal distribution with parameters θ_i, including mean vector μ_i and covariance matrix V_i, for the ith component; π_i is the prior probability that any observation comes from component i. The log-likelihood for data $\{x_1,...,x_n\}$ is

$$\log L(\Theta)=\sum_j \log\left(\sum_i \pi_i f_i(x_j;\theta_i)\right).$$

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). To implement the EM, one starts with the 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. See McLachlan and Peel (2002) and Fraley and Raftery (2002) for details.

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

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

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

$$Q_P(\Theta;\Theta^{(r)})=E_{\Theta^{(r)}}(\log L_{c,P}|X)=\sum_i\sum_j \tau_{ij}^{(r)}[\log\pi_i+\log f_i(x_j;\theta_i)]-p_\lambda(\Theta), \qquad (1)$$

where τ_{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 Θ. In the sequel, when deriving the updating formulas in the M-step, for simplicity we may suppress dependence of the estimates on iteration r.

In what follows, unless specified otherwise, we assume that all clusters share a common diagonal covariance matrix. Specifically, we assume

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

with $V = diag(\sigma_1^2, \sigma_2^1, \cdots, \sigma_K^2)$, and $|V| = \prod_{k=1}^K \sigma_k^2$.

2.1 L_1 *penalty*

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

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

where μ_{ik}'s are the components of μ_i, the mean vector of cluster i.

Pan and Shen (2007) derived the following updating formulas for the EM algorithm (Dempster *et al* 1977) to obtain MPLE:

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

the posterior probability that the jth observation comes from component i,

$$\hat{\pi}_i^{(r+1)} = \sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}/n, \tag{4}$$

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

$$\hat{\sigma}_k^{2,(r+1)} = \sum_{i=1}^{g} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} (x_{jk} - \hat{\mu}_{ik}^{(r)})^2/n, \tag{5}$$

the variance for the kth variable, and

$$\hat{\mu}_{ik}^{(r+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)}} \left(1 - \frac{\lambda \hat{\sigma}_k^{2,(r)}}{|\sum_{j=1}^{n} \hat{\tau}_{ij}^{(r)} x_{jk}|}\right)_+, \tag{6}$$

the mean for the kth variable in cluster i, with $i = 1, 2, \cdots, g$ and $k = 1, 2, \cdots, K$.

Obviously, for a sufficiently large λ, we have $\hat{\mu}_{ik} = 0$. Since each variable k has been standardized to have sample mean 0, if $\hat{\mu}_{1k} = ... = \hat{\mu}_{gk} = 0$, 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 μ_{ik} is individually penalized; on the other hand, variable k is regraded as a noise variable if and only if $\mu_{1k} = ... = \mu_{gk} = 0$. Hence, to realize more effective variable selection, it is natural to treat μ_{1k}, ..., μ_{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 μ_{1k}, ..., μ_{gk} is vertical, hence we call it vertical mean grouping (VMG), for which we propose the following penalty:

$$p_\lambda(\Theta) = \lambda\sqrt{g}\sum_{k=1}^{K} \|\mu_{.k}\|, \tag{7}$$

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

The updating formulas for τ_{ij}, π_i and σ_k^2 remain the same as (3)-(5), but that for μ_{ik} is different. We derive in Appendix A.1 the following:

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

$$diag\left(\sum_{j=1}^{n}\tau_{1j}, \sum_{j=1}^{n}\tau_{2j}, \cdots, \sum_{j=1}^{n}\tau_{gj}\right)(\tilde{\mu}_{.k} - \hat{\mu}_{.k}) = \lambda\sqrt{g}\sigma_k^2\frac{\hat{\mu}_{.k}}{\|\hat{\mu}_{.k}\|} \quad \textit{if and only if } \hat{\mu}_{.k} \neq \mathbf{0}, \tag{8}$$

where $\tilde{\mu}_{.k} = \left(\frac{\sum_{j=1}^{n}\tau_{1j}x_{jk}}{\sum_{j=1}^{n}\tau_{1j}}, \frac{\sum_{j=1}^{n}\tau_{2j}x_{jk}}{\sum_{j=1}^{n}\tau_{2j}}, \cdots, \frac{\sum_{j=1}^{n}\tau_{gj}x_{jk}}{\sum_{j=1}^{n}\tau_{gj}}\right)'$ *has the form of the MLE, and*

$$\left(\sum_{i=1}^{g}(\sum_{j=1}^{n}\tau_{ij}x_{jk})^2\right)^{1/2} \leqslant \lambda\sqrt{g}\sigma_k^2 \quad \textit{if and only if } \hat{\mu}_{.k} = \mathbf{0}. \tag{9}$$

It is clear that if the inequality in (9) is satisfied, we have $\hat{\mu}_{1k} = ... = \hat{\mu}_{gk} = 0$, thus variable k is regarded as a noise variable. Next we highlight a key difference between (9) and (6).

From (6), we have

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

while (9) can be rewritten as

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

Hence, according to (9), 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 $\hat{\mu}_{.k} = \mathbf{0}$; in contrast, by (6), some larger components of $\hat{\mu}_{.k}$ may remain to be non-zero. This highlight 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 $\hat{\mu}_{1k} = ... = \hat{\mu}_{gk} = 0$ simultaneously, thus more effective variable selection.

Combining (8) and (9) yields

$$\hat{\mu}_{.k} = \left(sign\left(1 - \frac{\lambda\sqrt{g}\sigma_k^2}{(\sum_{i=1}^{g}(\sum_{j=1}^{n}\tau_{ij}x_{jk})^2)^{1/2}}\right)\right)_+ \nu_k\tilde{\mu}_{.k} \tag{10}$$

with $\nu_k = diag\left(1 + \frac{\lambda\sqrt{g}\sigma_k^2}{||\hat{\mu}_{.k}||\sum_{j=1}^{n}\tau_{1j}}, 1 + \frac{\lambda\sqrt{g}\sigma_k^2}{||\hat{\mu}_{.k}||\sum_{j=1}^{n}\tau_{2j}}, \cdots, 1 + \frac{\lambda\sqrt{g}\sigma_k^2}{||\hat{\mu}_{.k}||\sum_{j=1}^{n}\tau_{gj}}\right)^{-1}$.

Equation (10) naturally suggests an iterative algorithm to update $\hat{\mu}_{.k}$. However, in simulation studies, we found that it did not work well. As alternative, first, we used (9) to derive whether $\hat{\mu}_{.k} = \mathbf{0}$; if not, we tried the following two methods. First, we rewrote (8) iteratively as

$$\hat{\mu}_{ik}^{(r+1)} = \tilde{\mu}_{ik}^{(r)} - \lambda\sqrt{g}\sigma_k^2\hat{\mu}_{ik}^{(r)}/(\sum_{j=1}^{n}\tau_{ij}||\hat{\mu}_{.k}^{(r)}||),$$

and then updated the components of $\mu_{.k}$ one by one. Second, we applied a Newton algorithm directly to solve (8). Although both methods worked better than using (10), the Newton method seemed best and was used in all numerical examples.

2.3 *Horizontal mean grouping*

In some applications, we may have prior knowledge that a group of variables are likely to be either informative or noise variables all together; 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 on the case with a common covariance matrix.

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

$$p_\lambda(\Theta) = \lambda \sum_{i=1}^{g} \sum_{m=1}^{M} \sqrt{k_m} \|\mu_i^m\|, \tag{11}$$

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 μ_i^m is horizontal, hence we call it horizontal mean grouping (HMG).

For HMG, the updating formulas for τ_{ij}, π_i and σ_k^2 remain the same as (3)-(5); we only need to derive that for μ_{ik}. After some algebra given in Appendix A.2, we obtain the following result:

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

$$(\sum_{j=1}^{n} \tau_{ij}) V_m^{-1} (\tilde{\mu}_i^m - \hat{\mu}_i^m) = \lambda \sqrt{k_m} \frac{\hat{\mu}_i^m}{\|\hat{\mu}_i^m\|} \quad \textit{if and only if } \hat{\mu}_i^m \neq \mathbf{0}, \tag{12}$$

$$\left\| \sum_{j=1}^{n} \tau_{ij} x_j^{m\prime} V_m^{-1} \right\| \leqslant \lambda \sqrt{k_m} \quad \textit{if and only if } \hat{\mu}_i^m = \mathbf{0}, \tag{13}$$

where $\tilde{\mu}_i^m = \sum_{j=1}^n \tau_{ij} x_j^m / \sum_{j=1}^n \tau_{ij}$ *has the form of the MLE.*

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

$$\sqrt{\frac{1}{k_m} \sum_{k \in J_m} \frac{\tilde{\mu}_{ik}^2}{\sigma_k^2}} \leqslant \frac{\lambda}{\sum_{j=1}^n \tau_{ij}} \quad \text{if and only if } \hat{\mu}_i^m = \mathbf{0}.$$

Hence, if the average (based on a weighted quadratic mean) of the components of $\tilde{\mu}_i^m$ is small enough, all the components of $\hat{\mu}_i^m$ are shrunken 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 (12) and (13) can be rewritten as

$$\hat{\mu}_i^m = \left(sign \left(1 - \frac{\lambda\sqrt{k_m}}{\| \sum_{j=1}^n \tau_{ij} x_j^m V_m^{-1} \|} \right) \right)_+ \nu_i^m \tilde{\mu}_i^m, \tag{14}$$

where $\nu_i^m = \left(\mathbf{I} + \frac{\lambda\sqrt{k_m}}{\sum_{j=1}^n \tau_{ij} \|\hat{\mu}_i^m\|} V_m \right)^{-1}$ and $\mathbf{I}$ is the identity matrix. (14) suggests an iterative algorithm to update $\hat{\mu}_i^m$, 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):

$$p_\lambda(\Theta) = \lambda \sum_{m=1}^M \sqrt{gk_m} \|\mu^m\| \tag{15}$$

with $\mu^m = (\mu_1^{m'}, \mu_2^{m'}, \cdots, \mu_g^{m'})'$.

Again the updating formulas for τ_{ij}, π_i and σ_k^2 remain the same as (3)-(5), and we only need to derive that for μ_{ik}. In Appendix A.3, we derive the following theorem:

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

$$(\sum_{j=1}^n \tau_{ij}) V_m^{-1} (\tilde{\mu}_i^m - \hat{\mu}_i^m) = \lambda \sqrt{gk_m} \frac{\hat{\mu}_i^m}{\|\hat{\mu}^m\|} \quad \textit{if and only if } \hat{\mu}^m \neq \mathbf{0} \tag{16}$$

for all $i = 1, 2, \cdots, g$, *and*

$$d_m \leqslant \lambda\sqrt{gk_m} \quad \textit{if and only if } \hat{\mu}^m = \mathbf{0}, \tag{17}$$

where $\tilde{\mu}_i^m = \sum_{j=1}^n \tau_{ij} x_j^m / \sum_{j=1}^n \tau_{ij}$ *has the form of the MLE, and*
$d_m = \left\| \left(\sum_{j=1}^n \tau_{1j} x_j^m{}' V_m^{-1}, \sum_{j=1}^n \tau_{2j} x_j^m{}' V_m^{-1}, \cdots, \sum_{j=1}^n \tau_{gj} x_j^m{}' V_m^{-1} \right) \right\|$.

Conditions (16) and (17) yield

$$\hat{\mu}_i^m = \left(sign \left(1 - \frac{\lambda\sqrt{gk_m}}{d_m} \right) \right)_+ \nu_i^m \tilde{\mu}_i^m. \tag{18}$$

Although (18) suggests an iterative algorithm, we found that the below one performed better and was used in our numerical examples:

$$\hat{\mu}_i^{m,(r+1)} = \left(sign\left(1 - \frac{\lambda\sqrt{gk_m}}{d_m}\right)\right)_+ \left(\tilde{\mu}_i^{m,(r)} - \frac{\lambda\sqrt{gk_m}}{\sum_{j=1}^n \tau_{ij}} V_m^{-1} \frac{\hat{\mu}_i^{m,(r)}}{||\hat{\mu}_i^{m,(r)}||}\right).$$

2.5 *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 shrunken to be exactly zero. The idea 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 λ jointly. Through a grid search, the optimal $(g, \hat{\lambda})$ is chosen to be the one with the minimal BIC.

For any given (g, λ), 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, λ).

3. Results

3.1 *Simulated data*

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 $N(0,1)$, the standard Normal distribution with mean 0 and variance 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 $K_1 = 10$.

For each simulated dataset, we fitted a series of models with the numbers of components $g = 1$, 2 and 3, and various values of penalization parameter λ. 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 λ.

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) in 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 L_1 penalty.

[Table 1 about here.]

3.2 *Real 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. Following Dudoit et al (2002), 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 1600 if $x_{jk} > 1600$; 2) filtering: any gene was excluded if its $max/min \leqslant 5$ and $max - min \leqslant 500$, where max and min were the maximum and minimum expression levels of the gene across all the samples; 3) transformation: the natural logarithms of the expression levels were used. Next, as in Pan and Shen (2007), 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.

Tables 2-3 show the clustering results. The vertical grouping (VMG) method could discriminate not only between AML and ALL samples, but also between two ALL subtypes: the T-cell and B-cell ALL samples were perfectly separated, while only one B-cell ALL sample was mis-allocated to the cluster with AML samples. In contrast, the standard clustering method without penalization yielded only two clusters: while most B-cell ALL samples were mixed with all the T-cell ALL samples in a cluster, three B-cell samples were misclassified into the cluster of the ALL samples. With 4 clusters, the L_1 penalty method performed better than the standard method in discriminating between ALL and AML samples, and between the two ALL subtypes, but four AML samples were mis-allocated into a cluster with 17 B-cell ALL samples.

The horizontal grouping (HMG) gave 9 clusters, though one cluster was empty. It worked well except that one cluster contained five B-cell ALL samples, two T-cell ALL samples and one AML samples. VHMG also worked well, yielding 4 clusters, one of which was empty. The empty cluster or component in HMG or VHMG could account for the non-normality of the other components, such as caused by the existence of outliers.

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 2 about here.]

[Table 3 about here.]

4. An extension

We have studied the vertical mean grouping with a common diagonal covariance across clusters. In practice, the assumption of a common covariance matrix may not hold. Here we extend the method to the case with cluster-specific diagonal covariance matrices, in which the probability density function of component i is

$$f_i(x;\theta_i) = \frac{1}{(2\pi)^{K/2}|V_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)\prime V_i^{-1}(x-\mu_i)\right) \tag{19}$$

with $V_i = diag(\sigma_{i1}^2, \sigma_{i2}^2, \cdots, \sigma_{iK}^2)$ and $|V_i| = \prod_{k=1}^K \sigma_{ik}^2$. For such a model, a key to realizing variable selection is to regularize variance parameters σ_{ik}^2, in addition to mean parameters μ_{ik} (Xie et al 2007).

We propose the following penalty

$$p_{\lambda_1,\lambda_2}(\Theta) = \lambda_1\sqrt{g}\sum_{k=1}^K \|\mu_{.k}\| + \lambda_2\sqrt{g}\sum_{k=1}^K \|\sigma_{.k}^2 - \mathbf{1}\| \tag{20}$$

with $\mu_{.k} = (\mu_{1k}, \mu_{2k}, \cdots, \mu_{gk})'$ and $\sigma_{.k}^2 = (\sigma_{1k}^2, \sigma_{2k}^2, \cdots, \sigma_{gk}^2)'$. The updating formulas for τ_{ij} and π_i are the same as (3)-(4); we only need to derive the updating formulas for σ_{ik}^2 and μ_{ik}.

4.1 *Vertical mean grouping*

Using a similar argument as in Appendix A.1, 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 maximizer of* (1) *are*

$$diag\left(\sum_{j=1}^n \tau_{1j}, \sum_{j=1}^n \tau_{2j}, \cdots, \sum_{j=1}^n \tau_{gj}\right)(\tilde{\mu}_{.k} - \hat{\mu}_{.k}) = \lambda_1\sqrt{g}diag(\sigma_{1k}^2, \sigma_{2k}^2, \cdots, \sigma_{gk}^2)\frac{\hat{\mu}_{.k}}{\|\hat{\mu}_{.k}\|}$$

$$\textit{if and only if } \hat{\mu}_{.k} \neq \mathbf{0}, \tag{21}$$

where $\tilde{\mu}_{.k} = \left(\frac{\sum_{j=1}^n \tau_{1j} x_{jk}}{\sum_{j=1}^n \tau_{1j}}, \frac{\sum_{j=1}^n \tau_{2j} x_{jk}}{\sum_{j=1}^n \tau_{2j}}, \cdots, \frac{\sum_{j=1}^n \tau_{gj} x_{jk}}{\sum_{j=1}^n \tau_{gj}}\right)'$ *has the form of the MLE, and*

$$\left(\sum_{i=1}^g \sigma_{ik}^{-2} \left(\sum_{j=1}^n \tau_{ij} x_{jk}\right)^2\right)^{1/2} \leqslant \lambda_1 \sqrt{g} \quad \textit{if and only if } \hat{\mu}_{.k} = \mathbf{0}. \tag{22}$$

As before, (22) clearly shows the effect of grouping: whether the whole vector $\hat{\mu}_{.k}$ is thresholded to $\mathbf{0}$ is determined by the average (i.e. weighted quadratic mean) of the components of $\tilde{\mu}_{.k}$, which is in the form of the MLE of $\mu_{.k}$. Combining (21) and (22) leads to

$$\hat{\mu}_{.k} = \left(sign \left(1 - \frac{\lambda_1 \sqrt{g}}{\left(\sum_{i=1}^g \sigma_{ik}^{-2} (\sum_{j=1}^n \tau_{ij} x_{jm})^2\right)^{1/2}} \right) \right)_+ \nu_k \tilde{\mu}_{.k} \tag{23}$$

with $\nu_k = diag \left(1 + \frac{\lambda_1 \sqrt{g} \sigma_{1k}^2}{\|\hat{\mu}_{.k}\| \sum_{j=1}^n \tau_{1j}}, 1 + \frac{\lambda_1 \sqrt{g} \sigma_{2k}^2}{\|\hat{\mu}_{.k}\| \sum_{j=1}^n \tau_{2j}}, \cdots, 1 + \frac{\lambda_1 \sqrt{g} \sigma_{gk}^2}{\|\hat{\mu}_{.k}\| \sum_{j=1}^n \tau_{gj}}\right)^{-1}$.

4.2 *Vertical variance grouping*

Because the objective function (1) may not be convex in σ_{ik}^2's, we can only obtain a weaker result:

THEOREM 5: *The necessary condition for* $\hat{\sigma}_{.k}^2 \neq \mathbf{1}$ *to be a local maximizer of* Q_P *is*

$$\sum_{j=1}^n \tau_{ij} \left(-\frac{1}{2\hat{\sigma}_{ik}^2} + \frac{(x_{jk} - \mu_{ik})^2}{2\hat{\sigma}_{ik}^4} \right) = \lambda_2 \sqrt{g} \frac{\hat{\sigma}_{ik}^2 - 1}{\|\hat{\sigma}_{.k}^2 - \mathbf{1}\|} \tag{24}$$

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

$$\begin{cases} \left\|\sum_{j=1}^n \tau_{.j} \left(-\frac{1}{2}\mathbf{1} + \frac{(x_{jk} - \mu_{.k})^2}{2}\right)\right\| \leqslant \lambda_2 \sqrt{g}, & \textit{if } \sum_j \tau_{.j}(1/2 - (x_{jk} - \mu_{.k})^2) > \mathbf{0}; \\ \left\|\sum_{j=1}^n \tau_{.j} \left(-\frac{1}{2}\mathbf{1} + \frac{(x_{jk} - \mu_{.k})^2}{2}\right)\right\| < \lambda_2 \sqrt{g}, & \textit{otherwise.} \end{cases} \tag{25}$$

The proof is given in Appendix A.4. According to the above result, a computational algorithm can be implemented as follows: we first check whether the sufficient and necessary condition (25) for $\hat{\sigma}_{.k}^2 = \mathbf{1}$ is satisfied; if yes, we have $\hat{\sigma}_{.k}^2 = \mathbf{1}$ and stop; otherwise, an iterative algorithm, e.g. the Newton algorithm, is developed to solve equation (24) to obtain $\hat{\sigma}_{.k}^2 \neq \mathbf{1}$.

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 that of grouping variables when there is prior knowledge that some variables work in groups. Analytical and numerical comparisons with the standard L_1-norm that treat 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. It can be used in other model-based clustering methods, such as the one that incorporates prior knowledge on variables as prior probabilities (Pan 2006). Furthermore, it can be applied to supervised and semi-supervised learning for more than two classes: 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 such an application in the shrunken centroid classifier, though a different penalty with the L_∞-norm, not L_2-norm, on groups of parameters was used.

For high-dimensional data, such as arising in genomic studies, we advocate the use of diagonal covariance matrices, following the same line of arguments as in supervised learning for such data (e.g. Tibshirani et al 2003; Bickel and Levina 2004). In particular, model-based clustering with diagonal covariance matrices is more general than the popular K-means clustering because the latter assumes not only a common diagonal covariance matrix across clusters, but also all equal diagonal elements. Nevertheless, it is worth extending the proposed methods to cases with more general covariance matrices, as discussed in McLachlan et al (2003).

Acknowledgement

BX and WP were partially supported by NIH grant HL65462 and a UM AHC Faculty Research Development grant, and XS by NSF grants IIS-0328802 and DMS-0604394.

APPENDIX

A.1 Proof of Theorem 1.

Using the vertical mean grouping penalty (7), we have

$$Q_P(\Theta;\Theta^{(r)}) = E_{\Theta^{(r)}}(\log L_{c,P}|X) = \sum_i\sum_j \tau_{ij}^{(r)}[\log\pi_i + \log f_i(x_j;\theta_i)] - \lambda\sum_{k=1}^{K}\sqrt{g}||\mu_{.k}||. \quad (26)$$

Since $Q_P(\Theta;\Theta^{(r)})$ is differentiable with respect to $\mu_{.k}$ when $\mu_{.k}\neq\mathbf{0}$, the solution $\mu_{.k}$ must satisfy the following equation

$$\begin{cases} \frac{\partial}{\partial\mu_{.k}}Q_P(\Theta;\Theta^{(r)}) = \mathbf{0} & \text{for all } \mu_{.k}\neq\mathbf{0}; \\ Q_P(\mathbf{0},.) \geqslant Q_P(\Delta\mu_{.k},.) & \text{for all } \Delta\mu_{.k} \text{ near } \mathbf{0}. \end{cases} \quad (27)$$

where . in $Q_P(\mathbf{0},.)$ represents all parameters in $Q_P(\Theta;\Theta^{(r)})$ except $\mu_{.k}$.

Note that $Q_P(\Theta;\Theta^{(r)}) = \sum_i\sum_j \tau_{ij}\left[-\frac{1}{2}(x_{jk}-\mu_{ik})^2\sigma_k^{-2}\right] - \lambda\sqrt{g}||\mu_{.k}|| + C$, where C is a constant w.r.t. $\mu_{.k}$. After taking derivatives, the first equation of (27) reduces to (8). Since both minus the first term in Q_P and the L_2-norm penalty are convex, (8) is the sufficient and necessary condition by the Karush-Kuhn-Tucker (KKT) condition.

From the second equation of (27), we have

$$\begin{aligned} \text{LHS} &= \sum_i\sum_j \tau_{ij}\left[-\frac{1}{2}x_{jk}^2\sigma_k^{-2}\right] + C \\ \text{RHS} &= \sum_i\sum_j \tau_{ij}\left[-\frac{1}{2}(x_{jk}-\Delta\mu_{ik})^2\sigma_k^{-2}\right] - \lambda\sqrt{g}||\Delta\mu_{.k}|| + C \end{aligned}$$

Thus

$$\begin{aligned}
& \mu_{.k} = \mathbf{0} \\
\Longleftrightarrow\quad & \sum_i \sum_j \tau_{ij} \left[-\frac{1}{2}(x_{jk})^2 \sigma_k^{-2} \right] \geqslant \sum_i \sum_j \tau_{ij} \left[-\frac{1}{2}(x_{jk} - \Delta\mu_{ik})^2 \sigma_k^{-2} \right] - \lambda\sqrt{g}||\Delta\mu_{.k}|| \\
\Longleftrightarrow\quad & \lambda\sqrt{g}||\Delta\mu_{.k}|| \geqslant -\frac{1}{2}\sum_i \sum_j \tau_{ij}(-2x_{jk}\Delta\mu_{ik} + (\Delta\mu_{ik})^2)\sigma_k^{-2} \\
\Longleftrightarrow\quad & \lambda\sqrt{g}\sigma_k^2 \geqslant \sum_i \sum_j \tau_{ij} x_{jk} \Delta\mu_{ik}/||\Delta\mu_{.k}|| - \frac{1}{2}\sum_i \sum_j \tau_{ij}(\Delta\mu_{ik})^2/||\Delta\mu_{.k}||. \qquad (28)
\end{aligned}$$

Note that $\frac{1}{2}\sum_i \sum_j \tau_{ij}(\Delta\mu_{ik})^2)/||\Delta\mu_{.k}|| \to 0^+$ as $\Delta\mu_{.k} \to \mathbf{0}$. By the Cauchy-Schwarz inequality, we have $\sum_i \sum_j \tau_{ij} x_{jk} \Delta\mu_{ik}/||\Delta\mu_{.k}|| \leqslant ||\sum_j \tau_{.j} x_{jk}||$, and the equality can be attained. Therefore, (28) is equivalent to (9), a sufficient and necessary condition for $\mu_{.k} = \mathbf{0}$.

A.2 Proof of Theorem 2

Consider two cases:

i) $\mu_i^m \neq \mathbf{0}$. We can treat Q_P as the Lagrange multiplier for a constrained optimization problem with the penalty as the inequality constraint, and considering that both minus the objective function and the L_2 norm penalty function are convex, by the Karush-Kuhn-Tucker (KKT) condition, we have the following sufficient and necessary condition

$$\partial Q_P/\partial \mu_i^m = \mathbf{0} \Longleftrightarrow \sum_j \tau_{ij} V_m^{-1}(x_j^m - \mu_i^m) - \lambda\sqrt{k_m}\mu_i^m/||\mu_i^m|| = \mathbf{0},$$

leading to (12).

ii) $\mu_i^m = \mathbf{0}$. By definition, we have

$$\begin{aligned}
& Q_P(\mathbf{0}, .) \geqslant Q_P(\Delta\mu_i^m, .) \text{ for any } \Delta\mu_i^m \text{ close to } \mathbf{0} \\
\Longleftrightarrow\quad & -\sum_j \tau_{ij} \frac{1}{2}(x_j^m)' V_m^{-1} x_j^m + C_1 \geqslant \\
& -\sum_j \tau_{ij} \frac{1}{2}(x_j^m - \Delta\mu_i^m)' V_m^{-1}(x_j^m - \Delta\mu_i^m) - \lambda\sqrt{k_m}||\Delta\mu_i^m|| + C_1 \\
\Longleftrightarrow\quad & \sum_j \tau_{ij} x_j^m{}' V_m^{-1} \Delta\mu_i^m/||\Delta\mu_i^m|| - \sum_j \tau_{ij}(\Delta\mu_i^m)' V_m^{-1} \Delta\mu_i^m/(2||\Delta\mu_i^m||) \leqslant \lambda\sqrt{k_m}.
\end{aligned}$$

Plugging-in $\Delta\mu_i^m = \alpha \sum_j \tau_{ij} V_m^{-1} x_j^m$ and letting $\alpha \to 0$, we obtain (13) from the above inequality. On the other hand, by the Cauchy-Schwarz inequality, we have $\sum_j \tau_{ij} x_j^m{}' V_m^{-1} \Delta\mu_i^m/||\Delta\mu_i^m||$

$\leqslant ||\sum_j \tau_{ij} x_j^m{}' V_m^{-1}||$, and because V_m^{-1} is positive definite, we obtain (13) from the above inequality.

A.3 Proof of Theorem 3

Using the penalty (15), we have

$$Q_P(\Theta; \Theta^{(r)}) = \sum_i \sum_j \tau_{ij}^{(r)} [\log \pi_i + \log f_i(x_j; \theta_i)] - \lambda \sum_{m=1}^{M} \sqrt{g \cdot k_m} \|\mu^m\|. \tag{29}$$

Since $Q_P(\Theta; \Theta^{(r)})$ is differentiable with respect to μ^m when $\mu^m \neq \mathbf{0}$, the solution μ^m must satisfy the following equation

$$\begin{cases} \frac{\partial}{\partial \mu^m} Q_P(\Theta; \Theta^{(r)}) = \mathbf{0} & \text{if } \mu^m \neq \mathbf{0} \\ Q_P(\mathbf{0}, .) \geqslant Q_P(\Delta\mu^m, .) & \text{if } \mu^m = \mathbf{0} \text{ and } \Delta\mu^m \text{ near } \mathbf{0}. \end{cases} \tag{30}$$

where . in $Q_P(\mathbf{0}, .)$ represents all parameters in $Q_P(\Theta; \Theta^{(r)})$ except μ^m.

Notice that $Q_P(\Theta; \Theta^{(r)}) = -\frac{1}{2} \sum_i \sum_j \tau_{ij} [(x_j - \mu_i^m)' V_m^{-1} (x_j - \mu_i^m)] - \lambda \sqrt{g \cdot k_m} ||\mu^m|| + C$, where C is a constant w.r.t. μ^m. After taking the derivative w.r.t. μ_i^m, the first equation of (30) becomes (16). Again according to the KKT condition, it is a sufficient and necessary condition for $\mu^m \neq \mathbf{0}$ to be the unique maximizer of (29).

For the second equation of (30),

$$\begin{aligned} \text{LHS} &= -\frac{1}{2} \sum_i \sum_j \tau_{ij} \left[(x_j)' V_m^{-1} (x_j) \right] + C, \\ \text{RHS} &= \sum_i \sum_j \tau_{ij} \left[-\frac{1}{2} (x_j - \Delta\mu_i^m)' V_m^{-1} (x_j - \Delta\mu_i^m) \right] - \lambda \sqrt{g \cdot k_m} ||\Delta\mu^m|| + C. \end{aligned}$$

Thus

$$
\begin{aligned}
& \mu^m = \mathbf{0} \\
\Longleftrightarrow \quad & -\frac{1}{2}\sum_i \sum_j \tau_{ij}\left[(x_j)' V_m^{-1}(x_j)\right] \geqslant \\
& -\frac{1}{2}\sum_i \sum_j \tau_{ij}\left[(x_j - \Delta\mu_i^m)' V_m^{-1}(x_j - \Delta\mu_i^m)\right] - \lambda\sqrt{g\cdot k_m}||\Delta\mu^m|| \\
\Longleftrightarrow \quad & \lambda\sqrt{g\cdot k_m}||\Delta\mu^m|| \geqslant \sum_i \sum_j \tau_{ij}\left[(x_j)' V_m^{-1}(\mu_i^m) - \frac{1}{2}(\mu_i^m)' V_m^{-1}(\mu_i^m)\right] \\
\Longleftrightarrow \quad & \lambda\sqrt{g\cdot k_m} \geqslant \sum_i \sum_j \tau_{ij} x_j' V_m^{-1}\frac{\mu_i^m}{||\Delta\mu^m||} - \frac{1}{2}\sum_i \sum_j \tau_{ij}\mu_i^{m'} V_m^{-1}\frac{\mu_i^m}{||\Delta\mu^m||}. \qquad (31)
\end{aligned}
$$

Notice that $||\frac{1}{2}\sum_i \sum_j \tau_{ij}\mu_i^{m'} V_m^{-1}\mu_i^m/||\Delta\mu^m|||| \to 0^+$ as $\Delta\mu^m \to \mathbf{0}$, and by the Cauchy-Schwarz inequality, we have $\sum_i \sum_j \tau_{ij} x_j' V_m^{-1}\mu_i^m/||\Delta\mu^m|| \leqslant d_m$, and the equality can be attained. Therefore, (31) is equivalent to (17).

A.4 Proof of Theorem 5

We have

$$
Q_P(\Theta;\Theta^{(r)}) = E_{\Theta^{(r)}}(\log L_{c,P}|X) = \sum_i \sum_j \tau_{ij}^{(r)}[\log \pi_i + \log f_i(x_j;\theta_i)] - \lambda_2 \sum_{k=1}^{K} \sqrt{g}||\sigma_{.k}^2 - \mathbf{1}|| \qquad (32)
$$

Since $Q_P(\Theta;\Theta^{(r)})$ is differentiable with respect to $\sigma_{.k}^2$ when $\sigma_{.k}^2 \neq \mathbf{1}$ for $k = 1, 2, \cdots, K$, a local maximum of (32) must satisfy the following conditions

$$
\begin{cases}
\frac{\partial}{\partial\sigma_{.k}^2} Q_P(\Theta;\Theta^{(r)}) = \mathbf{0} & \text{for all } \sigma_{.k}^2 \neq \mathbf{0}; \\
Q_P(\mathbf{1}, .) \geqslant Q_P(\mathbf{1} + \Delta\sigma_{.k}^2, .) & \text{for all } \Delta\sigma_{.k}^2 \text{ near } \mathbf{0}.
\end{cases} \qquad (33)
$$

where . in $Q_P(\mathbf{1}, .)$ represents all parameters in $Q_P(\Theta;\Theta^{(r)})$ except $\sigma_{.k}^2$.

Since $Q_P(\Theta;\Theta^{(r)}) = -\frac{1}{2}\sum_i \sum_j \tau_{ij}\left[\log(\sigma_{ik}^2) + (x_{jk} - \mu_{ik})'\sigma_{ik}^{-2}(x_{jk} - \mu_{ik})\right] - \lambda_2\sqrt{g}||\sigma_{.k}^2 - \mathbf{1}|| + C$, where C is a constant w.r.t. $\sigma_{.k}^2$, by taking the derivative, we obtain (24) from the first equation of (33).

If $\sigma^2_{.k} = \mathbf{1}$, the second equation of (33) gives

$$\begin{aligned}\text{LHS} &= -\frac{1}{2}\sum_i\sum_j \tau_{ij}\left[(x_{jk}-\mu_{ik})'(x_{jk}-\mu_{ik})\right] + C,\\ \text{RHS} &= -\frac{1}{2}\sum_i\sum_j \tau_{ij}\left[\log(1+\Delta\sigma^2_{ik}) + (x_{jk}-\mu_{ik})'(1+\Delta\sigma^2_{ik})^{-1}(x_{jk}-\mu_{ik})\right] - \lambda_2\sqrt{g}||\Delta\sigma^2_{.k}|| + C.\end{aligned}$$

Thus

$$\begin{aligned} & \sigma^2_{.k} = \mathbf{1}\\ \Longleftrightarrow\quad & -\frac{1}{2}\sum_i\sum_j \tau_{ij}\left[(x_{jk}-\mu_{ik})'(x_{jk}-\mu_{ik})\right] \geqslant\\ & -\frac{1}{2}\sum_i\sum_j \tau_{ij}\left[\log(1+\Delta\sigma^2_{ik}) + (x_{jk}-\mu_{ik})'(1+\Delta\sigma^2_{ik})^{-1}(x_{jk}-\mu_{ik})\right] - \lambda_2\sqrt{g}||\Delta\sigma^2_{.k}||\\ \Longleftrightarrow\quad & \lambda_2\sqrt{g}||\Delta\sigma^2_{.k}|| \geqslant\\ & -\frac{1}{2}\sum_i\sum_j \tau_{ij}\left[\log(1+\Delta\sigma^2_{ik}) - (x_{jk}-\mu_{ik})'(\Delta\sigma^2_{ik}/(1+\Delta\sigma^2_{ik}))(x_{jk}-\mu_{ik})\right]\end{aligned}$$

From Taylor's expansions $\log(1+x) = x - x^2/2 + O(x^3)$ and $x/(1+x) = x - x^2 + O(x^3)$, we obtain

$$\begin{aligned}\Longleftrightarrow\qquad & \lambda_2\sqrt{g}||\Delta\sigma^2_{.k}|| \geqslant -\tfrac{1}{2}\textstyle\sum_i\sum_j \tau_{ij}[(1-(x_{jk}-\mu_{ik})^2)\Delta\sigma^2_{ik})\\ & \qquad -(1/2-(x_{jk}-\mu_{ik})^2)(\Delta\sigma^2_{ik}))^2 + O(\Delta\sigma^2_{ik})^3)]\\ \Longleftrightarrow\qquad & \lambda_2\sqrt{g} \geqslant -\tfrac{1}{2}\textstyle\sum_i\sum_j \tau_{ij}[(1-(x_{jk}-\mu_{ik})^2)\Delta\sigma^2_{ik}/||\Delta\sigma^2_{.k}||\\ & -(1/2-(x_{jk}-\mu_{ik})^2)(\Delta\sigma^2_{ik}))^2/||\Delta\sigma^2_{.k}|| + O(\Delta\sigma^2_{ik})^3)/||\Delta\sigma^2_{.k}||]\end{aligned}$$

Notice that, as $\Delta\sigma_{.k} \to \mathbf{0}$, $\frac{1}{2}\sum_i\sum_j \tau_{ij}(1/2-(x_{jk}-\mu_{ik})^2)(\Delta\sigma^2_{ik}))^2/||\Delta\sigma^2_{.k}|| \to 0^+$ if $\sum_j \tau_{.j}(1/2-(x_{jk}-\mu_{.k})^2) > \mathbf{0}$; $\frac{1}{2}\sum_i\sum_j \tau_{ij}(1/2-(x_{jk}-\mu_{ik})^2)(\Delta\sigma^2_{ik}))^2/||\Delta\sigma^2_{.k}|| \to 0$ otherwise. By the Cauchy-Schwarz inequality, we have $||\sum_j \tau_{.j}(\mathbf{1}-(x_{jk}-\mu_{.k})^2)|| \geqslant \sum_i\sum_j \tau_{ij}(1-(x_{jk}-\mu_{ik})^2)\Delta\sigma^2_{ik}/||\Delta\sigma^2_{.k}||$, and the equality can be attained.

Therefore, we obtain (25) as the sufficient and necessary condition for $\sigma^2_{.k} = \mathbf{1}$.

REFERENCES

Bickel P.J., Levina E. (2004). Some theory for Fisher's linear discriminant function, "naive

Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989-1010.

Dempster AP, Laird NM, Rubin DB. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *JRSS-B* **39**, 1-38.

Dudoit S, Fridlyand J, Speed T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77-87.

Efron B, Hastie T, Johnstone I, Tibshirani R. (2004). Least angle regression. *Annals of Statistics* **32**, 407-499.

Eisen M, Spellman P, Brown P and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**, 14863-14868.

Fan J, Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties. *JASA* **96**, 1348-1360.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611-631.

Ghosh D, Chinnaiyan, AM. (2002). Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275-286.

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

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27-30.

Li H. and Hong F. (2001). Cluster-Rasch models for microarray gene expression data. *Genome Biology* **2**, research0031.1-0031.13.

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.

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

McLachlan, G.J., Peel, D. and Bean, R.W. (2003). Modeling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* **41**, 379-388.

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

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

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

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

Tibshirani R, Hastie T, Narasimhan B, Chu G. (2003). Class prediction by nearest shrunken centroids, with application to DNA microarrays. *Statistical Science* **18**, 104-117.

Wang, S. and Zhu, J. (2006). Improved centroids estimation for the nearest shrunken centroid classifier. To appear in *Bioinformatics.*

Xie, B, Pan, W. and Shen, X. (2007). Penalized model-based clustering with cluster-specific diagonal covariances and grouped variables. Available at http://www.biostat.umn.edu./rrs.php as Research Report 2007-017, Division of Biostatistics, University of Minnesota.

Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977-987.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *JRSS-B*, **68**, 49-67.

Zou H, Hastie T, Tibshirani R. (2004). On the "Degrees of Freedom" of the Lasso. Technical report, Statistics Dept, Stanford University.

Available at http://stat.stanford.edu/~hastie/pub.htm.

First Draft May 24, 2007

Table 1

Simulation results: 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, $g = 2$, $z_1 = 0$, $z_2 = 290$ and $\mu_1 = 1.25$.

		$\lambda = 0$	L_1			VMG			HMG			VHMG		
Set-up	g	N	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2	N	z_1	z_2
1	1	100	100	10.0	290.0	100	10.0	290.0	100	10.0	290.0	100	10.0	290.0
	2	0	0	-	-	0	-	-	0	-	-	0	-	-
	3	0	0	-	-	0	-	-	0	-	-	0	-	-
2	1	100	100	5.0	295.0	78	5.0	295.0	76	5.0	295.0	59	5.0	295.0
	2	0	0	-	-	22	0.2	292.8	19	0.0	295.0	39	0.0	295.0
	3	0	0	-	-	0	-	-	5	0.0	295.0	2	0.0	295.0
3	1	100	18	10.0	290.0	1	10.0	290.0	2	10.0	290.0	0	-	-
	2	0	60	0.0	286.0	99	0.1	287.7	89	0.0	289.9	93	0.0	290.0
	3	0	22	0.0	286.2	0	-	-	9	0.0	290.0	7	0.0	290.0
4	1	100	92	10.0	290.0	45	10.0	290.0	53	10.0	290.0	63	10.0	290.0
	2	0	6	0.0	284.7	49	0.9	291.9	36	0.0	289.9	30	0.0	290.0
	3	0	2	0.0	287.0	6	0.2	293.8	11	0.0	290.0	7	0.0	290.0

Table 2

Clustering results for Golub's data.

Methods	$\lambda = 0$		L_1						VMG		
#genes	2000		1281						426		
Clusters	1	2	1	2	3	4	5	6	1	2	3
Samples (#)											
ALL-T (8)	0	8	0	0	0	0	0	8	8	0	0
ALL-B (19)	3	16	1	0	1	0	17	0	0	1	18
AML (11)	11	0	0	6	0	1	4	0	0	11	0

Table 3

Clustering results for Golub's data (continued).

Methods	HMG									VHMG			
#genes	504									54			
Clusters	1	2	3	4	5	6	7	8	9	1	2	3	4
Samples (#)													
ALL-T (8)	0	2	0	0	6	0	0	0	0	8	0	0	0
ALL-B (19)	0	5	1	8	0	4	1	0	0	0	1	18	0
AML (11)	7	1	0	0	0	0	0	3	0	0	11	0	0