Computational Statistics manuscript No.
(will be inserted by the editor)

High-dimensional Variable Selection with the Plaid Mixture Model for Clustering

Thierry Chekouo · Alejandro Murua*

Received: date / Accepted: date

Abstract With high-dimensional data, the number of covariates is considerably larger than the sample size. We propose a sound method for analyzing these data. It performs simultaneously clustering and variable selection. The method is inspired by the plaid model. It may be seen as a multiplicative mixture model that allows for overlapping clustering. Unlike conventional clustering, within this model an observation may be explained by several clusters. This characteristic makes it specially suitable for gene expression data. Parameter estimation is performed with the Monte Carlo expectation maximization algorithm and importance sampling. Using extensive simulations and comparisons with competing methods, we show the advantages of our methodology, in terms of both variable selection and clustering. An application of our

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant number 327689-06.

Thierry Chekouo

Department of Mathematics and Statistics,

University of Minnesota Duluth,

1117 University Drive,

Duluth, MN,55812, USA

E-mail: tchekouo@d.umn.edu

Alejandro Murua

Département de mathématiques et de statistique,

Université de Montréal

CP 6128, succ. centre-ville,

Montréal, Québec H3C 3J7 Canada

*Corresponding author: E-mail: murua@dms.umontreal.ca

approach to the gene expression data of kidney renal cell carcinoma taken from The Cancer Genome Atlas validates some previously identified cancer biomarkers.

Keywords classification · model selection · multiplicative mixture model · Monte Carlo EM · kidney cancer genomic data

1 Introduction

Microarray data consist of many thousands of gene expression profiles but only tens or hundreds of samples. These data are typical examples of high-dimensional data for which the number of covariates (genes) is considerably larger than the sample size. Having so much information poses problems for model selection. It implies making decisions as to which data should be investigated or even retained. For this reason, a classical way to start the analysis of high-dimensional data is with exploratory techniques such as clustering or biclustering (Madeira and Oliveira, 2004; Tanay et al, 2005). Both techniques may be used for data compression and/or dimensionality reduction. However, in many situations clustering is the goal, such as when trying to detect subtypes of a disease. In this case, having a sound and efficient methodology to perform variable selection is key to advancing the study of the disease. For example, in cancer research, only a few genes in the genome are known to contribute to most of the characterization of cancer subtypes. Several authors have treated the problem of variable selection in the context of clustering. Tadesse et al (2005) formulated the clustering problem in a Bayesian context. In their model, the non-discriminating variables follow a multivariate normal distribution, while the discriminating ones follow a multivariate normal mixture model with an unknown number of components. In their model, a binary exclusion/inclusion latent vector is introduced to indicate whether a variable is selected (i.e., is discriminating) or not. Other authors (Kim et al, 2006; Hoff, 2006) have also introduced Bayesian variable selection methods through binary latent vectors to select the discriminating variables. Another approach to variable selection within a mixture model for clustering, described by Raftery and Dean (2006), uses Bayes factors. Raftery and Dean (2006) proposed a greedy search algorithm to find a local optimum in the model space, and used the Bayesian information criterion (BIC) to approximate the Bayes factor. A generalization of the Raftery and Dean (2006) model proposed by Maugis et al (2009a) does not need any prior assumptions about the linear link between the discriminating and the discarded variables.

Another class of models for variable selection uses penalization methods for model-based clustering (Pan and Shen, 2007; Xie et al, 2008; Wang and Zhu, 2008). A popular approach among these methods is that of Pan and Shen (2007), which is based on a penalized likelihood approach with an L_1 penalty term. Specifically, following Hoff (2006), Pan and Shen (2007) parameterized the cluster means, say μ_k, for each variable $j = 1, \ldots, p$, as $\mu_{jk} = \upsilon_j + \beta_{jk}$, where υ_j is the overall cluster-independent mean for variable j. They inferred that if $\beta_{jk} = 0$ for all clusters k, then the variable j is uninformative for clustering (at least in terms of the mean). The model is fitted with an expectation maximization (EM) algorithm. Witten and Tibshirani (2010) also apply a Lasso-type penalty to select the variables. Their method is based on sparse K-means and sparse hierarchical clustering. The method of Witten and Tibshirani (2010) differs from that of Pan and Shen (2007), for the Lasso penalty is applied on the weights of each variable. These weights are defined as the contribution of the variables to the resulting sparse clustering. The gap statistic (Tibshirani et al (2000)) is used to determine the optimal value of the tuning parameter for their sparse clustering algorithms.

Here, we propose a novel method to select the variables in the context of clustering. This method is inspired by the plaid model of Lazzeroni and Owen (2002). Let $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\} \subset \mathbf{R}^p$ be a random sample of n observations. Assume that the data have a structure consisting of K clusters. We introduce the latent variables for cluster labeling as $\rho = \{\rho_{ik}\}_{i=1,k=1}^{n,K}$ and the latent variables for variable selection as $\kappa = \{\kappa_j\}_{j=1}^{p}$. They are indicator variables, so that $\rho_{ik} = 1$ if the i-th observation is in cluster k; otherwise it is set to zero. Similarly, $\kappa_j = 1$ if the j-th variable is a discriminating variable; otherwise it is set to zero. We also use the notation $\rho_i = \{\rho_{ik}\}_{k=0}^{K}$, where $\rho_{i0} = \prod_{k=1}^{K}(1 - \rho_{ik})$, $i = 1, \ldots, n$. Note that the i-th observation has $\rho_{i0} = 1$ if it does not belong to any of the K clusters, that is, if it belongs to the *background* or *zero cluster* (see point (A) below). Our variable selection model for clustering is defined as follows. For any given pair (ρ, κ), the expectation of y_{ij} is a sum of layers or plaids $E(y_{ij}|\rho_i, \kappa_j) = \kappa_j \sum_{k=0}^{K} \rho_{ik}(\mu_k + \alpha_{ik} + \beta_{jk}) + (1 - \kappa_j)\upsilon_j$, where μ_k is the overall mean of the objects in cluster k, β_{jk} is the effect of the j-th variable in cluster k, α_{ik} is the random effect in cluster k associated with the i-th observation, and υ_j is the background mean of variable j (see below for further explanation). For identifiability purposes, we impose the constraints $\sum_{i=1}^{n} \rho_{ik}\alpha_{ik} = \sum_{j=1}^{p} \kappa_j\beta_{jk} = 0$, $k = 1, .., K$. Each plaid corresponds to a cluster. Note that the expectation of y_{ij} in the usual mixture model may be written as

$\mu_{jk} = E(y_{ij}|k) = \mu_k + \beta_{jk}$. Our model differs from other variable selection models based on mixtures in the three following ways.

(A) We consider the possibility that some observations are not well explained by the main clusters, but rather lie in what we call the zero cluster ($k = 0$) (note that the background mean υ_j is the zero cluster mean of variable j). These observations satisfy the constraints $\alpha_{i0} = \beta_{j0} = 0$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The presence of this cluster may be justified by some observations in real data sets. In clustering, there is often a "ragbag" cluster for data that do not belong to any well-defined cluster and which are thus considered to be noise. Hence, it is desirable to consider a model that can leave a few points un-clustered if necessary.

(B) We incorporate random effects in the observations. Therefore, the observations and the variables play a symmetrical role in each cluster. The random effects take into account the potential influence of single observations in the model. In addition, they introduce compound symmetry in the variance-covariance matrix associated with observations given the clusters. When this is not appropriate for the data at hand, then we can either simply eliminate the random effects from the model, or consider them as fixed effects (i.e., as parameters to be estimated). For example, in the case of gene expression data, the effect of each gene (the observations) is of interest, so it makes sense to incorporate fixed gene effects in the model (as opposed to random gene effects) and to avoid imposing compound symmetry. In particular, if $\mu_k + \beta_{jk} > 0$ for one gene j, then this gene is *upregulated* within cluster k; otherwise, it is *downregulated.* In the present study, we work with the assumption of fixed effects in the observations

(C) The observations may be explained by more than a single cluster ($\sum_{k=1}^{K} \rho_{ik} \geq 1$). This produces an aggregate overlapping (superimposition) of clusters that is different from the distributional overlapping of clusters (that is, when the mixture component densities overlap) implicit in the usual mixture model. For example, in clinical applications (Bhattacharya, 2005), a patient may belong to more than one clinical group, i.e., a patient who complains of headache may have migraine symptoms and other causes of headache (such as nasal or sinus problems/disease). Methods to address overlapping clustering are available in the literature (Fu and Banerjee, 2008, 2009; Heller and Ghahramani, 2007). These models, which are motivated by the product-of-experts model (Hinton, 2002), are often called *multiplicative mixture models*. We show that our approach is related to these approaches.

Similar to many models for clustering available in the literature, our model involves latent labels ρ, κ; thus, to estimate the parameters, we use a stochastic version of the EM algorithm that is based on the so-called Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990). This is a modified EM algorithm in which the expectation in the E-step is computed numerically through Monte Carlo simulations. We use a Gibbs sampler to perform Monte Carlo sampling in each iteration of the MCEM algorithm. However, as suggested by Levine and Casella (2001), we also use importance sampling to overcome the computational cost of the Monte Carlo sampling at each step of the EM algorithm.

We apply our method to the analysis of gene expression data associated with kidney renal cell carcinoma (KIRC), the most prevalent form of kidney cancer. Most treatments target the clear cell carcinoma type, which accounts for 80% of all KIRC cases. The data were obtained from The Cancer Genome Atlas (TCGA) (`https://tcga-data.nci.nih.gov/tcga/`), a large public repository for cancer-related genomic data. We aim to cluster the kidney cancer samples and identify important genes related to cancer development and progression that are capable of discriminating among the samples/patients.

The paper is organized as follows: Section 2 describes the proposed plaid mixture model for variable selection. The Monte Carlo EM procedure devised to estimate the parameters of the model is explained in Section 3. In Section 4 , we propose information criteria suitable to select the number of clusters. A simulation to compare the performance of our model with that of other popular methods is presented in Section 5. In section 6, we show an application of our approach to the analysis of KIRC gene expression data. Our conclusions are stated in Section 7.

2 The plaid mixture model

Throughout the paper, we follow the notation provided in the introduction. Inspired by the plaid model of Lazzeroni and Owen (2002), we propose a general model for variable selection in the context of clustering. Our model comprises the clustering label parameters ρ, the variable selection parameters κ, the variance parameters $\Sigma = (\{\varrho_j^2\}_{j=1}^p, \{\sigma_{jk}^2\}_{j=1,k=0}^{p,K})$, and the mean parameters $\Psi = (\mu, \{\mu_k\}_{k=0}^K, \beta, \alpha)$, with $\alpha = \{\alpha_{ik}\}_{i=1,k=0}^{n,K}$ and $\beta = \{\beta_{jk}\}_{j=1,k=0}^{p,K}$. The model is given by

$$y_{ij} = \kappa_j(\sum_{k=0}^{K}(\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik} + \eta_{ij}) + (1 - \kappa_j)(\upsilon_j + \epsilon_{ij}), \tag{1}$$

where the η_{ij}'s and ϵ_{ij}'s are assumed to follow independent zero-mean normal distributions. The variance of ϵ_{ij} is ϱ_j^2. The variance of η_{ij} is the harmonic mean of the variances σ_{jk}^2, which may depend on the cluster k and the variable j. It is given by $\tau_{ij}^2 = \sum_{k=0}^K \rho_{ik} / \sum_{k=0}^K (\rho_{ik}/\sigma_{jk}^2) = \left(\sum_{k=0}^K \rho_{ik}/r_i\sigma_{jk}^2\right)^{-1}$, where $r_i = \sum_{k=0}^K \rho_{ik} \geq 1$ is the number of plaids (that is, clusters) that jointly explain observation y_{ij}. This form of the variance allows us to cast our model as a multiplicative mixture model for which the variances for cluster k are $r_i\sigma_{jk}^2$ (see equation (3)).

Prior Distribution

We assume that the prior probabilities that any given variable j is selected are the same for all $j = 1, \ldots, p$. This is denoted by $\pi = P(\kappa_j = 1)$, any j. Thus, the prior distribution for the number of variables is binomial, with mean $p\pi$. This assumption is very common in the Bayesian variable selection setup (George and McCulloch, 1997; Li and Zhang, 2010).

The prior probability that the i-th observation is explained by cluster k is assumed to be the same for all observations $i = 1, \ldots, n$. It is denoted by $\pi_k = P(\rho_{ik} = 1)\ i = 1, \ldots, n$. We denote by $\mathit{\Pi} = (\pi, \{\pi_k\}_{k=0}^K)$ all the prior probability parameters. In addition, we assume that the Bernoulli latent variables $(\{\rho_i\}_{i=1}^n, \{\kappa_j\}_{j=1}^p)$ are mutually independent.

Likelihood

Let $\phi(.)$ denote the density function of the standard normal distribution. Hereafter, we write μ_{ijk} for $\mu_k + \alpha_{ik} + \beta_{jk}$. Let $\theta = (\mathit{\Sigma}, \mathit{\Psi}, \mathit{\Pi})$. The complete data likelihood is given by

$$\begin{aligned} L(\theta|\mathcal{Y}, \rho, \kappa) &= P(\mathcal{Y}|\rho, \kappa, \mathit{\Sigma}, \mathit{\Psi}) \prod_{i,k} {\pi_k}^{\rho_{ik}} (1-\pi_k)^{1-\rho_{ik}} \prod_j \pi^{\kappa_j} (1-\pi)^{1-\kappa_j} \\ &= \prod_{i,j} \left[\frac{1}{\tau_{ij}} \phi\left(\frac{y_{ij} - \sum_{k=0}^K \mu_{ijk}\rho_{ik}}{\tau_{ij}}\right)\right]^{\kappa_j} \left[\frac{1}{\varrho_j} \phi\left(\frac{y_{ij} - \upsilon_j}{\varrho_j}\right)\right]^{1-\kappa_j} \\ &\quad \times \prod_{i,k} {\pi_k}^{\rho_{ik}} (1-\pi_k)^{1-\rho_{ik}} \prod_j \pi^{\kappa_j} (1-\pi)^{1-\kappa_j} \end{aligned} \tag{2}$$

Let $\kappa^* = \{j : \kappa_j = 1, j = 1, \ldots, p\}$ be the set of the selected variables. One can show that the density of $\mathcal{Y}$ on the selected discriminating variables, that is $j \in \kappa^*$, is given by

$$P(\mathcal{Y}|\rho, \kappa^*, \theta) = \prod_{i,j} \frac{1}{c_{ij}(\rho, \kappa^*, \theta)} \prod_{k=0}^K \left[\frac{1}{\sqrt{r_i}\sigma_{jk}} \phi\left(\frac{y_{ij} - \mu_{ijk} r_i \sigma_{jk}^2/\tau_{ij}^2}{\sqrt{r_i}\sigma_{jk}}\right)\right]^{\rho_{ik}}, \tag{3}$$

where

$$c_{ij}(\rho, \kappa^*, \theta) = \frac{\tau_{ij}\sqrt{2\pi}}{\prod_k (\sqrt{r_i}\sigma_{jk}\sqrt{2\pi})^{\rho_{ik}}} \exp\left\{ \frac{1}{2\tau_{ij}^2} \left(\sum_{k=0}^{K} \mu_{ijk}\rho_{ik} \right)^2 - \frac{1}{2\tau_{ij}^4} \sum_{k=0}^{K} r_i \mu_{ijk}^2 \rho_{ik} \sigma_{jk}^2 \right\}.$$

Equation (3) shows that our model is similar to the multiplicative mixture model for overlapping clustering described by Fu and Banerjee (2008, 2009), and Heller and Ghahramani (2007) (see Section A in the Supplementary Material for more details). Our likelihood is proportional to

$$\prod_{i,k} \left[\frac{1}{\prod_j (\sqrt{2\pi r_i}\sigma_{jk})} \exp\{-\frac{1}{2}(\tilde{y}_i - \tilde{\mu}_{ik})^T D_{ik}^{-1} (\tilde{y}_i - \tilde{\mu}_{ik})\} \right]^{\rho_{ik}}, \tag{4}$$

where D_{ik} is the diagonal matrix with the main diagonal given by the vector $\{r_i\sigma_{jk}^2\}_{j=1}^p$, $\tilde{y}_i = \{y_{ij}\}_{j=1}^p$, and $\tilde{\mu}_{ik} = \{\mu_{ijk} r_i \sigma_{jk}^2 / \tau_{ij}^2\}_{j=1}^p$. Within this model, the mean and variance parameters associated with cluster k are $\tilde{\mu}_{ik}$ and D_{ik}. Note that when there is no aggregate overlapping of clusters (i.e., $r_i = 1$ for all $i = 1, \ldots, n$), the mean and the variance of cluster k are simply given by the parameters $\tilde{\mu}_{ijk} = \mu_{ijk}$ and σ_{jk}^2. Equation (3) is also related to the product of experts (PoE) of Hinton (2002). Indeed, the PoE model with $K+1$ components can be expressed as $P(\mathcal{Y}|\theta) \propto \prod_{k=0}^K p_k(\mathcal{Y}|\theta_k)$, where θ_k and $p_k(\mathcal{Y}|\theta_k)$ are respectively the set of parameters and density associated with component k. So when all components of ρ are set to 1, our multiplicative model (4) becomes a PoE model.

3 Estimation

The EM algorithm is particularly suitable for learning the parameters of our model (2) because the likelihood of the complete data $(\mathcal{Y}, \rho, \kappa)$ is much easier to calculate than the likelihood of the observed data $\mathcal{Y}$. More specifically, the EM algorithm starts with an initial guess of the unknown parameters, $\theta^{(0)} = (\Sigma^{(0)}, \Psi^{(0)}, \Pi^{(0)})$, and iteratively aims to estimate the maximum likelihood estimator (MLE) $\theta^\star$. Each iteration consists of the expectation (E) step and the maximization (M) step. Next, we show some of the details of the algorithm, which is summarized in Section 3.4

3.1 The E-step

Given an estimate of θ at the current iteration t, say $\theta^{(t)}$, the conditional expectation of the complete data log-likelihood with respect to the density $P(\rho, \kappa|\mathcal{Y}, \theta)$ is computed in the E-step:

$$Q(\theta|\theta^{(t)}) = E\big(\log(P(\mathcal{Y}, \rho, \kappa|\theta)) \,|\, \mathcal{Y}, \theta^{(t)}\big), \qquad t \geq 0. \tag{5}$$

We cannot compute the exact expectation (5) because we do not have a tractable closed-form expression for the joint conditional density $P(\rho, \kappa|\mathcal{Y}, \theta)$. However, since the full conditionals of ρ and κ are easily obtained, we propose to estimate $Q(\theta|\theta^{(t)})$ via an MCEM algorithm (Wei and Tanner, 1990). The proposed estimator is given by

$$Q_m(\theta|\theta^{(t)}) = \frac{1}{m}\sum_{l=1}^{m} \log(P(\mathcal{Y}, \rho(l), \kappa(l)|\theta)), \tag{6}$$

where $\rho(l), \kappa(l)$, $l = 1, .., m$ are samples from the conditional joint distribution of the latent variables ρ, κ given the observed data $\mathcal{Y}$ and the current value of the parameters $\theta^{(t)}$. The estimator in (6) converges to the theoretical expectation in (5) by the law of large numbers. Below, we explain how to obtain the label samples via a Gibbs sampler.

3.2 The M-step

The M-step maximizes the sum (6) with respect to θ subject to the identifiability constraints $\sum_i \rho_{ik}\alpha_{ik} = \sum_j \kappa_j\beta_{jk} = 0$, for all k. To overcome the computational cost of performing MCMC sampling within the MCEM algorithm when m is large, Levine and Casella (2001) proposed using importance sampling (IS) (Robert and Casella, 2004). The algorithm is initialized by m samples, $\rho(l), \kappa(l)$, $l = 1, .., m$ from the joint distribution $P(\rho, \kappa|\mathcal{Y}, \theta^{(0)})$. At iteration t, we estimate $Q(\theta|\theta^{(t)})$ by IS:

$$Q_{IS,m}(\theta|\theta^{(t)}) = \frac{1}{\sum_{l=1}^{m} w_l^{(t)}} \sum_{l=1}^{m} w_l^{(t)} \log(P(\mathcal{Y}, \rho(l), \kappa(l)|\theta)) \tag{7}$$

where $w_l^{(t)} = P(\mathcal{Y}|\rho(l), \kappa(l), \theta^{(t)})/P(\mathcal{Y}|\rho(l), \kappa(l), \theta^{(0)})$. Thus, we do not need to obtain a new sample of m labels from $P(\rho, \kappa|\mathcal{Y}, \theta^{(t)})$ at each iteration t in order to estimate $Q(\theta|\theta^{(t)})$. The cost of obtaining a new sample of m labels at each iteration is higher than that of obtaining the IS weights. Note that the weights may be written as

$$w_l^{(t)} = \prod_{i,j} w_l^{(t)}(i,j), \text{ with } w_l^{(t)}(i,j) = \frac{P(y_{ij}|\rho_i(l), \kappa_j(l), \theta^{(t)})}{P(y_{ij}|\rho_i(l), \kappa_j(l), \theta^{(0)})}. \tag{8}$$

The EM updating equations are given in Section B of the Supplementary Material.

3.2.1 Increasing the IS size $\mathbf{m}$

As pointed out by Robert and Casella (2004), the IS estimator in (7) would be inaccurate if the initial parameter values $\theta^{(0)}$ were poor. In addition, the estimator would take a long time to converge. Hence, as

suggested by Levine and Casella (2001), we obtain MCMC samples from $P(\rho, \kappa|\mathcal{Y}, \theta^{(t)})$ for the first few iterations. The choice of the MCMC sample size m is an issue within the MCEM algorithm because we do not want to use a large m when $\theta^{(t)}$ is far from the true MLE $\hat{\theta}$. The trade-off between the computational cost and the accuracy of the estimator of $Q(\theta|\theta^{(t)})$ can be resolved by increasing the sample size m as $\theta^{(t)}$ approaches the true MLE during the progression of the EM algorithm. This is what Booth and Hobert (1999) proposed within the context of generalized linear mixed models. In their procedure, the increase in m obeys a schedule induced by a simple confidence region test: at the $(t+1)th$ iteration of the MCEM, an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\hat{\theta}^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$ is constructed using the central limit theorem. If the previous estimate of the parameter $\theta^{(t)}$ lies in this region, then the procedure declares that "the EM-Step was *swamped* by Monte Carlo error" and the number of simulations, m, is increased. We note that this schedule is based on true Monte Carlo samples, whereas we use MCMC samples. The dependency between the MCMC samples does not allow us to directly use the central limit theorem to construct a confidence interval. However, we overcome this limitation by borrowing some ideas from Robert et al (1999) and Levine and Casella (2001) to limit the effect of the correlation between successive samples. We choose a sequence u_r, $r = 1, ..., N$ such that $u_r - 1 \sim \text{Poisson}(\nu_r)$ where $\nu_r = \nu r^d$ for some $\nu \geq 0$ and $d > 0$. The sums $l_r = \sum_{j=1}^{r} u_r$ are used as the sub-sampling points, and N, the number of such sub-samples taken from the m samples, is set to $\sup\{r : l_r \leq m\}$. For a more detailed description, see Section B1 of the Supplementary Material.

3.3 Sampling the Labels

As mentioned, the joint density of the labels $P(\rho, \kappa|\mathcal{Y}, \theta^{(t)})$ is not known in closed form; therefore, we cannot perform the Monte Carlo sampling of the labels (ρ, κ) required to compute $Q_{IS,m}(\theta|\theta^{(t)})$. However, we can obtain an MCMC estimate of this quantity. This is carried out with a Gibbs sampler because the marginal conditional distributions of the labels are known. For $i \in \{1, \ldots, n\}$ and $k \in \{0, \ldots, K\}$, let $\rho_{i0}^{(k)} = \prod_{k' \neq k}(1 - \rho_{ik'})$, and $\rho_{-ik} = \rho_k \setminus \{\rho_{ik}\}$. The labels ρ_i for each $i = 1, ..., n$ and κ_j for each $j = 1, .., p$ are generated independently according to the odds $P(\rho_{ik} = 1|\mathcal{Y}, \rho_{-ik}, \kappa, \theta)/P(\rho_{ik} = 0|\mathcal{Y}, \rho_{-ik}, \kappa, \theta)$ and $P(\kappa_j = 1|\mathcal{Y}, \rho, \theta)/P(\kappa_j = 0|\mathcal{Y}, \rho, \theta)$ which are respectively given by equations (9) and

(10).

$$\exp\left\{\sum_{j=1}^{p}\frac{\kappa_j}{2\sigma_j^2}(\mu_{ijk}-\mu_0\rho_{i0}^{(k)})(2y_{ij}-2\sum_{k'\neq k}\mu_{ijk'}\rho_{ik'}-\mu_0\rho_{i0}^{(k)}-\mu_{ijk})\right\}\frac{\pi_k}{1-\pi_k} \quad (9)$$

$$\frac{\varrho_j^n}{\sigma_j^n}\exp\left\{\frac{-1}{2\sigma_j^2}\sum_{i=1}^{n}(y_{ij}-\sum_{k}\mu_{ijk}\rho_{ik})^2+\frac{1}{2\varrho_j^2}\sum_{i=1}^{n}(y_{ij}-\upsilon_j)^2\right\}\frac{\pi}{1-\pi} \quad (10)$$

In the case of non-aggregate overlapping clusters, that is, $r_i = 1$ for all i, the Gibbs sampler alternatively uses $P(\rho_{ik} = 1|\kappa,\theta) = A_{ik}/\sum_{k=0}^{K} A_{ik}$, and $P(\kappa_j = 1|\rho,\theta) = B_{j1}/B_{j0} + B_{j1}$, where $A_{ik} = \prod_j \left[\frac{1}{\sigma_{kj}}\phi\left(\frac{y_{ij}-\mu_k-\alpha_{ik}-\beta_{jk}}{\sigma_{kj}}\right)\right]^{\kappa_j}\pi_k$, $B_{j1} = \prod_{i,k}\left[\frac{1}{\sigma_{kj}}\phi\left(\frac{y_{ij}-\mu_k-\alpha_{ik}-\beta_{jk}}{\sigma_{kj}}\right)\right]^{\rho_{ik}}\pi$, and $B_{j0} = \prod_i\left[\frac{1}{\varrho_j}\phi\left(\frac{y_{ij}-\mu}{\varrho_j}\right)\right](1-\pi)$.

3.4 The Algorithm

The sampling algorithm to estimate the model parameters is summarized below. In addition to the E-step and M-step, it includes a Monte Carlo error checking step to decide whether to increase the sampling size m of the IS scheme.

1. Initialize m, and $\theta^{(0)} = (\Sigma^{(0)}, \Psi^{(0)}, \Pi^{(0)})$ (See Section 5.2 for further details). Set $t = 0$.

2. Generate m label samples $\rho(l)$, $\kappa(l)$, $l = 1, .., m$ using the Gibbs sampler according to equations (9) and (10).

3. Compute the importance weights $w_l(i,j)$ for all i, j using the equation (8).

4. E-step: Estimate $Q(\theta|\theta^{(t)})$ by: $E_{IS}(\kappa_j|\mathcal{Y},\theta^{(t)}) = \sum_{l=1}^{m} w_l\kappa_j(l)/\sum_{l=1}^{m} w_l$, $E_{IS}(\rho_{ik}|\mathcal{Y},\theta^{(t)}) = \sum_{l=1}^{m} w_l\rho_{ik}(l)/\sum_{l=1}^{m} w_l$, and $E_{IS}(\rho_{ik}\kappa_j|\mathcal{Y},\theta^{(t)}) = \sum_{l=1}^{m} w_l\rho_{ik}(l)\kappa_j(l)/\sum_{l=1}^{m} w_l$.

5. M-step: Maximize $Q_m(\theta|\theta^{(t)})$ over θ to obtain $\theta^{(t+1)}$ through the EM updating equations given in the Supplementary Material.

6. MC error: Perform the tests described in Section 3.2.1. If any one of the tests is negative, that is, if any one of the components of the vector $Q_{IS,m}^{(1)}(\theta^{(t)}|\theta^{(t-1)})$ lies in the corresponding confidence interval, then (a) Set $m_0 = m$; (b) Set $m = m_0 + \lfloor m_0/c \rfloor$, where $c = 3$ in our simulations; and (c) Generate new labels $\rho(l), \kappa(l), l = m_0 + 1, ..., m$ via the Gibbs sampler.

7. Set $t = t + 1$. Repeat steps 3 through 6 until convergence is achieved.

As stated previously, if the initial value $\theta^{(0)}$ is poor, that is, if $P(\rho, \kappa|\mathcal{Y}, \theta^{(0)})$ is far from $P(\rho, \kappa|\mathcal{Y}, \theta^{*})$, then the algorithm will take a long time to converge. Thus, in our simulations, we include a burn-in period in step 1 so that at each burn-in iteration, we estimate $Q_m(\theta|\theta^{(t)})$ via MCMC instead of IS. Thus, our computations during the burn-in period behave like the MCEM algorithm described by McCulloch (1997).

4 Model selection

We propose a modified BIC (Schwarz, 1978) to use in model selection within our multiplicative plaid mixture model: $\mathrm{BIC}_{plaid} = -2\log L(\hat{\theta}|\mathcal{Y}) + d_e \log(n)$, where $L(\hat{\theta}|\mathcal{Y})$ is the likelihood of the incomplete data, $\hat{\theta}$ is the MLE, and $d_e = d - s$ is the effective number of parameters given by the difference between d, the total number of parameters, and s, the number of uninformative parameters. The latter number is given by the number of null parameters, $\alpha_{ik} = 0$, $\beta_{jk} = 0$, and the number of parameters associated with σ_j^2 for variables excluded from the model ($\kappa_j = 0$), or with $\upsilon_j = 0$ and ϱ_j^2 for variables included in the model ($\kappa_j = 1$). More formally, $d_e = \sum_{k=1}^{K} n_k + p_0' \times (K+1) + 2$, where n_k is the number of samples in cluster k and p_0' is the estimated number of selected variables. Table 2 of the supplementary material (see Section 5.1) shows that in scenario 1, d_e is much smaller that d. This definition of BIC is inspired by that of Pan and Shen (2007) for penalized model-based clustering with variable selection. We use it as a goodness-of-fit criterion to select an appropriate number of clusters K. The optimal K is the one that minimizes BIC_{plaid}. Note that our BIC_{plaid} is the analog of the usual BIC used in model-based clustering, since only those parameters actually used in the model are considered in the penalty term. The term $L(\hat{\theta}|\mathcal{Y})$ is intractable because it involves the sum of all possible combinations of label values. So, in order to compute BIC_{plaid}, we use an estimate of $L(\hat{\theta}|\mathcal{Y})$ derived by IS. This is given by $L_{IS}(\hat{\theta}|\mathcal{Y}) = \sum_{l=1}^{m} w_l P(\mathcal{Y}, \rho(l), \kappa(l)|\hat{\theta}) / \sum w_l$. In our experiments, we use $\mathrm{BIC}_{IS,plaid} = -2\log L_{IS}(\hat{\theta}|\mathcal{Y}) + d_e \log(n)$. We also looked at the model selection results yielded by a modified Akaike information criterion (AIC) (Akaike, 1974). Similar to the construction of the $\mathrm{BIC}_{IS,plaid}$, we consider a modified AIC, computed using IS, and given by $\mathrm{AIC}_{IS,plaid} = -2\log L_{IS}(\hat{\theta}|\mathcal{Y}) + 2d_e$. In our experiments and simulations, the criteria $\mathrm{AIC}_{IS,plaid}$ and $\mathrm{BIC}_{IS,plaid}$ performed similarly. We also applied the DIC (Deviance information criterion, Spiegelhalter et al (2002)) and ICL (Integrated Completed Likelihood, Biernacki et al (2000)) criteria to our data. The results from ICL, which are based on the complete likelihood, were very similar

to ones from BIC. On the other hand, DIC tended to select a much larger number of clusters than the true number of clusters.

5 Comparison of methods by simulation

In this section, we illustrate the effectiveness of our method by conducting a simulation study with two different data scenarios. The first one mimics the synthetic data described by Pan and Shen (2007), with $K = 1$ (i.e., two clusters) and no aggregate overlapping clusters. The second scenario concerns synthetic data sets built with aggregate sample overlap between clusters. By definition, for $K = 1$, there is no aggregate overlapping among the clusters. We applied two versions of the plaid model to the simulated data. The first one assumes that there are aggregate overlapping clusters. The second version assumes there is no aggregate overlapping at all. We respectively refer to these two versions of the model as Plaid-Full and Plaid-Restricted. We compare the performance of these two models with that of the Lasso-type L_1-penalization method of Pan and Shen (2007), the sparse K-means Lasso-type of Witten and Tibshirani (2010), and the Gaussian model-based clustering for variable selection of Maugis et al (2009a,b) , which generalized the procedure of Raftery and Dean (2006). We refer to these three methods as L_1-Penalty, SK-Means and SVM, respectively.

The L_1-Penalty of Pan and Shen (2007) penalizes the L_1-norm of the cluster means so as to obtain sparseness in the mean vectors. In this approach, a zero component across all cluster means corresponds to a variable not being selected. The L_1-Penalty algorithm was run with a maximum of 100 iterations, and 10 clusters. The penalty parameter λ, whose values were restricted to the interval $[1,\ 21]$, was estimated using the BIC criterion. We also try the method of Zhou et al (2009). This generalizes the approach of Pan and Shen (2007) by allowing unconstrained covariance matrices in a Normal mixture model. In our simulations, this method presented computational difficulties when run with data with a large number of variables. The SK-Means of Witten and Tibshirani (2010) uses an iterative algorithm to maximize a weighted between-cluster sum of squares subject to constraints on the weights. A weight of zero means that the corresponding feature is not involved in the clustering. When the weights are equal for all variables, the problem simply reduces to the standard K-means clustering criterion. We chose the gap-statistics to estimate both the tuning parameter and the number of clusters. To select the tuning parameter (an L_1-bound on variable weights),

we run their permutation approach algorithm with 10 permutations of the data and a possible choice of 20 tuning parameters ranging from 1.5 to 4. The "best" tuning was used to run their algorithm with a maximum of 30 iterations and 7 clusters. The SVM variable selection of Maugis et al (2009b) generalizes the Raftery and Dean (2006) method by accounting for three possible roles for variables: the relevant clustering variables (discriminative variables), the irrelevant clustering variables (non-discriminative) that depend on some relevant variables, and the irrelevant clustering variables (non-discriminative) totally independent of all relevant variables. We run their algorithm with the "general" family of the mixture model. We assumed three possible forms of the covariance matrix for the linear regression between some relevant and irrelevant variables: spherical, diagonal and general forms. For the other irrelevant variables, we considered the two possible covariance matrix forms: spherical and diagonal forms. We selected the best model using the BIC criterion. We used the *R* code published by Zhou (2009), the *R* packages *sparcl* and *SelvarMix* to run the L_1-Penalty, the SK-Means and the SVM methods, respectively. It is important to remember that our clustering model contains $K + 1$ clusters, which includes the zero cluster. If another clustering method selects two clusters, for example, then the corresponding K for comparison with our model is $K = 1$.

5.1 Simulated Data

Scenario 1. In the first scenario, we closely followed the simulation carried out by Pan and Shen (2007) so as to be able to compare fairly our results with those given by the L_1-Penalty method of Pan and Shen (2007). We generated a two-cluster 1000-dimensional data set with a hundred observations. To have unbalanced cluster sample sizes, eighty-five observations are located in the first cluster; the remaining fifteen are located in the second cluster. Only the first 80 variables are discriminating variables for clustering. Specifically, the first 80 variables were independent and identically distributed (iid) and generated as $y_{ij} \sim I_{\{1\leq i\leq 85\}}N(0,\sigma^2) + I_{\{86\leq i\leq 100\}}N(1.5,\sigma^2)$, $j = 1,\ldots,80$, whereas the remaining 920 variables were all iid $N(0,\sigma^2)$. As these data do not present fixed effects ($\{\alpha_{ik}\}$ and $\{\beta_{jk}\}$) in the response, any of the two clusters may be considered the zero cluster of the multiplicative plaid mixture model.

In order to study the effect of the level of noise in the analysis, we have varied the overall variance σ^2 for different datasets. We consider the values $\sigma^2 = 0.64, 1, 1.21, 1.44, 1.69$, which give respectively signal-to-noise ratios of 1.87, 1.5, 1.36, 1.25 and 1.15.

Furthermore, in order to study the robustness of our model against the presence of dependent variables, we have extended scenario 1 by considering data with different variable correlation structures. The first 40 discriminative variables are not correlated with any other variable. The second 40 discriminative variables are correlated within each other with correlation $\tau_w \in \{0, 0.2, 0.3\}$, and also with 40 other non-discriminative variables with correlation $\tau_b \in \{0, 0.1\}$. The overall variance was kept at $\sigma^2 = 1.69$. In the terminology of Maugis et al (2009b), this setup corresponds to 80 relevant variables for clustering (40 of them are relevant dependent variables), and 920 irrelevant variables (880 of them are irrelevant independent variables).

Scenario 2. In this scenario, we simulate datasets with more complicated clustering structures according to our model for $K = 2$ (that is, 3 clusters). Besides finding out how the plaid methods compared with those of Pan and Shen (2007), Maugis et al (2009b) and Witten and Tibshirani (2010), we want to study the behavior of the five methods with respect to the sample size and the number of (discriminating) variables. With these goals in mind, we generate ten replications of each of the six p-dimensional datasets with n observations, with $n \in \{50, 100\}$, and $p \in \{100, 500, 1000\}$. We set the number of discriminating variables as $p_0 = p/20$. Moreover, we create data with aggregate overlapping clusters. More specifically, the number of overlap samples between cluster $k = 1$ and $k = 2$ is $n/5$. The first p_0 variables are independently distributed $N(\sum_{k=0}^{2}(\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik}, 1)$, whereas the other $p - p_0$ variables are all iid $N(\upsilon_j, \varrho_j^2)$, $i = 1, ..., n$. More details on the simulation setup are provided in the supplementary material, Section C. We would expect a better performance of the Plaid-Full model since some clusters overlap. In addition, we would also expect a better clustering performance when $n = 100$ and p_0 is large.

Further simulation results based on this scenario but with larger number of clusters or with varying level of noise in the data are presented in the supplementary material, Section C.2 and Section D.2.

5.2 Results

The algorithms to fit the plaid models were run with $m = 60$. We included a burn-in period of 20 samples. We set a maximum of 100 iterations for finding the optimal parameters. In practice, our algorithm converged in much fewer iterations (about 50 iterations), and the time to achieve the convergence was approximately

2 minutes for $p = 1000$. We used a computer cluster of 24 cores at 2.6 GHz and 20 gigabytes of RAM in a 64-bit Red Hat Linux platform. The program was written in Java and only uses one core. To obtain good starting values for any given K, we ran the MCEM algorithm several times with random starting values. In order to initialize the labels, we randomly started several K-means algorithms. To find initial values for the cluster labels ρ, we ran K-means with $K + 1$ clusters, and found a "good" zero cluster among them. To do so, we repeated the algorithm $K + 1$ times by initializing the zero cluster as the K-means cluster k for every $k = 0, ..., K$. To initialize the variable labels κ, we also ran K-means algorithms, but on the variables. We set $K = 2$ and separately considered each of the two clusters as possible initial selected variables. For any given K, we performed multiple runs of this procedure. Initial values $\theta^{(0)}$ of θ were computed as follows. For each run and cluster k, we initialized μ_k as the sample mean of the data that were assigned to this cluster, and σ_j as the sample standard deviation of all the data for which $\kappa_j = 1$. The effects α_{ik} and β_{jk} were initialized to zero. Finally, the variances ϱ_j^2 for the non-discriminative variables were initialized as the sample variance of the data y_{ij} for which $\kappa_j = 0$. The final results were the ones associated with the optimal runs, that is, with the ones that yielded the highest log-likelihood for any given K.

In order to measure the quality of the clustering estimated by the methods, we compared the estimated clustering with the true clustering of the data through the so-called F_1-measure. This is defined as the harmonic average between recall and precision, which are two measures of retrieval quality introduced in the text mining literature (Allan et al, 1998). Let A, B be two clusters, and $|A|$ and $|B|$ be the number of elements in A and B, respectively. Recall and precision are given by recall $= |A \cap B|/|B|$, precision $= |A \cap B|/|A|$. So, recall is the proportion of elements in B that are in A, and precision is the proportion of elements in A that are also found in B. The F1-measure between A and B is given by $F_1(A, B) = 2|A \cap B|/(|A| + |B|)$. When an estimated clustering $M_1 = \{A_1, \ldots, A_K\}$ is to be compared with the true clustering $M_2 = \{B_1, \ldots, B_L\}$, we use the F1-measure average: $F_1(M_1, M_2) = \frac{1}{K} \sum_{k=1}^{K} \max_j F_1(A_k, B_j)$. We would like to stress that the more common measure of clustering quality, the adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985), is not properly defined for overlapping clusters. Instead, in this case, the F_1-measure is preferred in the literature. We computed the F_1-measure associated with the clustering of observations (F_1), and the F_1-measure associated with the selected variables (F_1^v). We also report their corresponding standard deviations (in brackets). F_1^v may be interpreted as a measure of the power of the method to detect

all discriminative variables. It can be written as $F_1^v = 2(p_0 - Z_1)/(p_0 - Z_1 + p - Z_2)$, where p_0 is the true number of informative variables, Z_1 is the number of discriminating variables excluded from the model, and Z_2 is the number of non-informative variables excluded from the model.

Table 1 shows the results for the first scenario based on 10 replications of each simulated dataset. As we can see, when the level of noise is small ($\sigma^2 \in \{0.64, 1\}$) the four methods, Plaid-Full, Plaid-Restricted, SK-Means and L_1-Penalty, detected the true structure of the clustering. But when the noise σ^2 is large, the plaid methods (Plaid-Full and Plaid-Restricted) performed much better than the other three methods. The clustering results of the SVM method of Maugis et al (2009b) are not as good as those obtained by the other methods (the F_1 is smaller). The plaid methods also performed better than the other three methods in terms of discriminative variable detection (larger F_1^v), but tended to keep slightly fewer (about 1.7% excluded) informative variables than the L_1-Penalty method. On average, SVM selected only about four variables among the 80 informative variables and any of the 920 noise variables. On average, the SK-Means method selected only 40 variables among the 80 informative variables and selected some noise variables for large σ^2.

Table 2 shows the results for scenario 1 with correlated data both within discriminative variables, and between discriminative and non-discriminative variables. In terms of discriminative variable detection, the plaid methods still perform much better than all the competitive methods considered here. In most cases, they also perform better in terms of clustering. However, when $\tau_w > 0$ and $\tau_b = 0.1$, the L1-Penalty performs as well as the Plaid-Restricted. As observed previously (see Table 1), SVM has a consistently lowest discriminative variable detection across all the cases.

Table 3 shows the results associated with each method for the second scenario. From this table, we observe in general that all the methods performed better when $n = 100$. The Plaid-Full method performed better in terms of clustering than the Plaid-Restricted method, which is expected since Plaid-Full accounts for the overlapping between clusters. It is clear from this table that the plaid methods performed much better than the three other methods in terms of both quality of clustering and discriminative variable detection (F_1^v is one for all cases). We stress that detecting all discriminative variables is of particular important in certain applications, such as those involving gene expression data. The table also shows that: (1) the L_1-Penalty method picks the right variables a good proportion of the time, but it does not obtain the data clustering

Table 1 Scenario 1: Results with variables generated independently with $\mu_0 = 0$, $\mu_1 = 1.5$, and different values of σ^2. F_1 is the F_1 measure evaluated between the true clustering and the clustering estimated by the corresponding method. F_1^v is the F_1 measure evaluated between the discriminative variables and the variables selected by the corresponding method. Z_1 is the number of variables excluded from the model out of the 80 informative variables. Z_2 is the number of variables excluded from the model out of the 920 noise variables. The numbers in the parentheses are the corresponding standard errors obtained from 10 replications of each dataset.

Method	σ^2	F_1	F_1^v	Z_1	Z_2
L1-Penalty	0.64	1.00 (0.00)	0.91 (0.01)	0.00 (0.00)	904.44 (1.91)
Plaid-Full	0.64	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	920.00 (0.00)
Plaid-Restricted	0.64	1.00 (0.00)	1.00 (0.00)	0.11 (0.11)	920.00 (0.00)
SK-Means	0.64	1.00 (0.00)	0.52 (0.01)	51.78 (0.83)	920.00 (0.00)
SVM	0.64	0.97 (0.01)	0.13 (0.01)	74.56 (0.53)	920.00 (0.00)
L1-Penalty	1	1.00 (0.00)	0.90 (0.01)	0.22 (0.15)	903.11 (2.06)
Plaid-Full	1	1.00 (0.00)	0.98 (0.00)	2.71 (0.42)	919.57 (0.26)
Plaid-Restricted	1	1.00 (0.00)	0.98 (0.00)	3.11 (0.61)	919.67 (0.24)
SK-Means	1	0.98 (0.02)	0.55 (0.02)	49.67 (1.34)	920.00 (0.00)
SVM	1	0.84 (0.03)	0.08 (0.01)	76.44 (0.38)	920.00 (0.00)
L1-Penalty	1.21	1.00 (0.00)	0.91 (0.01)	0.89 (0.31)	904.44 (1.63)
Plaid-Full	1.21	1.00 (0.00)	0.96 (0.01)	4.12 (0.63)	918.38 (0.53)
Plaid-Restricted	1.21	1.00 (0.00)	0.96 (0.01)	4.62 (0.47)	918.38 (0.53)
SK-Means	1.21	0.77 (0.04)	0.69 (0.02)	37.89 (1.63)	920.00 (0.00)
SVM	1.21	0.79 (0.03)	0.08 (0.01)	76.78 (0.36)	920.00 (0.00)
L1-Penalty	1.44	1.00 (0.00)	0.89 (0.01)	1.78 (0.55)	903.11 (2.06)
Plaid-Full	1.44	1.00 (0.00)	0.93 (0.01)	7.62 (0.73)	916.38 (0.64)
Plaid-Restricted	1.44	1.00 (0.00)	0.93 (0.01)	7.62 (0.73)	916.25 (0.66)
SK-Means	1.44	0.75 (0.04)	0.68 (0.02)	38.56 (2.30)	919.89 (0.11)
SVM	1.44	0.62 (0.04)	0.05 (0.01)	77.89 (0.26)	920.00 (0.00)
L1-Penalty	1.69	0.96 (0.01)	0.49 (0.16)	37.78 (13.36)	911.78 (2.74)
Plaid-Full	1.69	1.00 (0.00)	0.89 (0.01)	10.00 (0.99)	912.67 (1.91)
Plaid-Restricted	1.69	1.00 (0.00)	0.89 (0.01)	10.22 (0.95)	912.33 (2.10)
SK-Means	1.69	0.79 (0.05)	0.63 (0.05)	41.44 (4.22)	918.78 (0.88)
SVM	1.69	0.56 (0.04)	0.06 (0.00)	77.67 (0.17)	920.00 (0.00)

with the same accuracy; and that (2) SK-Means is not able to only select discriminative variables when p is small (very poor F_1^v).

We also looked at the model selection results yielded by the plaid models using the $\mathrm{AIC}_{IS,plaid}$ criterion. The results were very similar to those obtained with $\mathrm{BIC}_{IS,plaid}$. Overall, there was no statistically significant difference between the F_1 results from BIC and AIC. However, we note that AIC gives more pronounced peaks at the right number of clusters than BIC. It appears from our simulations that BIC tends to over-penalizes the number of clusters when the dimension is very large. These and further results are

Table 2 Scenario 1: Results based on 5 replications of each dataset with correlated variables with $\mu_0 = 0$, $\mu_1 = 1.5$. F_1, F_1^v, Z_1 and Z_2 are as in Table 1. τ_b is the correlation between discriminative and irrelevant variables, and τ_w is the correlation within discriminative variables. The numbers in the parentheses are the corresponding standard errors.

Method	τ_w	τ_b	F_1	F_1^v	Z_1	Z_2
L1-Penalty	0	0	0.98 (0.02)	0.67 (0.22)	22.25 (19.28)	907.75 (4.40)
Plaid-Full	0	0	1.00 (0.00)	0.89 (0.02)	10.75 (2.17)	912.75 (4.61)
Plaid-Restricted	0	0	1.00 (0.00)	0.88 (0.03)	11.00 (2.04)	912.25 (5.11)
SK-Means	0	0	0.89 (0.06)	0.62 (0.06)	43.25 (5.72)	919.75 (0.25)
SVM	0	0	0.57 (0.09)	0.05 (0.01)	77.75 (0.25)	920.00 (0.00)
L1-Penalty	0.2	0	0.99 (0.01)	0.90 (0.02)	0.75 (0.48)	902.75 (3.54)
Plaid-Full	0.2	0	0.99 (0.01)	0.93 (0.01)	4.00 (0.71)	913.25 (2.81)
Plaid-Restricted	0.2	0	0.98 (0.02)	0.94 (0.01)	4.75 (0.75)	915.50 (2.84)
SK-Means	0.2	0	0.82 (0.06)	0.61 (0.08)	43.50 (6.76)	919.00 (1.00)
SVM	0.2	0	0.63 (0.08)	0.05 (0.01)	77.75 (0.25)	920.00 (0.00)
L1-Penalty	0.3	0	0.84 (0.09)	0.60 (0.21)	20.25 (19.59)	870.00 (25.19)
Plaid-Full	0.3	0	1.00 (0.00)	0.96 (0.01)	3.00 (0.58)	916.00 (1.58)
Plaid-Restricted	0.3	0	1.00 (0.00)	0.96 (0.01)	3.00 (0.58)	916.00 (1.58)
SK-Means	0.3	0	0.82 (0.09)	0.48 (0.05)	52.75 (2.78)	913.50 (4.72)
SVM	0.3	0	0.65 (0.12)	0.06 (0.01)	77.50 (0.29)	920.00 (0.00)
L1-Penalty	0	0.1	0.90 (0.06)	0.60 (0.21)	21.00 (19.67)	889.75 (16.06)
Plaid-Full	0	0.1	0.91 (0.06)	0.89 (0.01)	11.00 (0.71)	914.25 (1.65)
Plaid-Restricted	0	0.1	0.91 (0.06)	0.88 (0.02)	11.75 (0.63)	913.00 (1.91)
SK-Means	0	0.1	0.80 (0.11)	0.60 (0.07)	44.25 (6.29)	919.75 (0.25)
SVM	0	0.1	0.63 (0.08)	0.05 (0.01)	77.75 (0.25)	920.00 (0.00)
L1-Penalty	0.2	0.1	0.92 (0.07)	0.82 (0.07)	0.50 (0.50)	882.00 (17.69)
Plaid-Full	0.2	0.1	0.81 (0.12)	0.84 (0.11)	13.00 (8.01)	906.25 (10.27)
Plaid-Restricted	0.2	0.1	0.88 (0.09)	0.91 (0.05)	7.25 (2.59)	912.25 (5.81)
SK-Means	0.2	0.1	0.81 (0.08)	0.43 (0.13)	55.50 (7.98)	912.00 (8.00)
SVM	0.2	0.1	0.69 (0.10)	0.06 (0.01)	77.50 (0.29)	920.00 (0.00)
L1-Penalty	0.3	0.1	0.75 (0.11)	0.58 (0.15)	17.25 (16.92)	856.00 (19.76)
Plaid-Full	0.3	0.1	0.69 (0.10)	0.73 (0.11)	18.50 (8.21)	893.50 (9.87)
Plaid-Restricted	0.3	0.1	0.77 (0.14)	0.75 (0.13)	19.75 (10.63)	899.25 (11.14)
SK-Means	0.3	0.1	0.69 (0.09)	0.30 (0.12)	63.00 (6.81)	906.50 (5.95)
SVM	0.3	0.1	0.66 (0.13)	0.07 (0.01)	77.25 (0.48)	920.00 (0.00)

shown with more details in the supplementary material, Section D.2. Based on these results, we decided to favor the results hinted by AIC in the applications with gene expression data described in the next section.

6 Application to TCGA Kidney Cancer Data

TCGA is a large public repository for cancer-related genomic data. In addition to detailed patient clinical information (age, overall survival time, tumor stage, etc.), TGCA has data on DNA methylation, mRNA

Table 3 Scenario 2: Results based on 10 replicates with $\mu_0 = 0$, $\mu_1 = 3$ and $\mu_2 = 6$. Cluster 1 and cluster 2 present an overlap. F_1 and F_1^v are as in Table 1. Z_1 is the number of variables excluded from the model out of the p_0 informative variables. Z_2 is the number of variables excluded from the model out of the $p - p_0$ noisy variables. The numbers in the parentheses are the corresponding standard errors.

Method	p	n	F_1	F_1^v	Z_1	Z_2
L1-Penalty	100	100	0.74 (0.00)	0.84 (0.25)	0.00 (0.00)	90.50 (10.34)
Plaid-Full	100	100	0.87 (0.08)	1.00 (0.00)	0.00 (0.00)	95.00 (0.00)
Plaid-Restricted	100	100	0.84 (0.06)	1.00 (0.00)	0.00 (0.00)	95.00 (0.00)
SK-Means	100	100	0.87 (0.00)	0.10 (0.00)	0.00 (0.00)	0.00 (0.00)
SVM	100	100	0.76 (0.01)	1.00 (0.00)	0.00 (0.00)	95.00 (0.00)
L1-Penalty	1000	100	0.75 (0.00)	0.92 (0.06)	0.00 (0.00)	940.50 (7.50)
Plaid-Full	1000	100	0.86 (0.14)	1.00 (0.00)	0.00 (0.00)	950.00 (0.00)
Plaid-Restricted	1000	100	0.82 (0.09)	1.00 (0.00)	0.00 (0.00)	950.00 (0.00)
SK-Means	1000	100	0.87 (0.00)	0.67 (0.04)	25.00 (2.20)	950.00 (0.00)
SVM	1000	100	0.83 (0.07)	0.14 (0.05)	46.29 (1.50)	950.00 (0.00)
L1-Penalty	500	100	0.75 (0.00)	0.93 (0.05)	0.00 (0.00)	471.00 (2.88)
Plaid-Full	500	100	0.92 (0.11)	1.00 (0.00)	0.00 (0.00)	475.00 (0.00)
Plaid-Restricted	500	100	0.85 (0.07)	1.00 (0.00)	0.00 (0.00)	475.00 (0.00)
SK-Means	500	100	0.87 (0.00)	0.92 (0.02)	3.75 (0.71)	475.00 (0.00)
SVM	500	100	0.78 (0.02)	0.38 (0.08)	19.00 (1.41)	475.00 (0.00)
L1-Penalty	100	50	0.75 (0.04)	0.85 (0.15)	0.00 (0.00)	92.88 (2.36)
Plaid-Full	100	50	0.93 (0.07)	1.00 (0.00)	0.00 (0.00)	95.00 (0.00)
Plaid-Restricted	100	50	0.78 (0.14)	1.00 (0.00)	0.00 (0.00)	95.00 (0.00)
SK-Means	100	50	0.87 (0.00)	0.10 (0.00)	0.00 (0.00)	0.00 (0.00)
SVM	100	50	0.76 (0.05)	0.70 (0.22)	2.12 (1.36)	95.00 (0.00)
L1-Penalty	1000	50	0.74 (0.00)	0.88 (0.10)	0.00 (0.00)	935.25 (14.57)
Plaid-Full	1000	50	0.86 (0.14)	1.00 (0.00)	0.00 (0.00)	950.00 (0.00)
Plaid-Restricted	1000	50	0.84 (0.07)	1.00 (0.00)	0.00 (0.00)	950.00 (0.00)
SK-Means	1000	50	0.87 (0.00)	0.68 (0.03)	24.25 (1.49)	950.00 (0.00)
SVM	1000	50	0.87 (0.01)	0.06 (0.02)	48.38 (0.52)	950.00 (0.00)
L1-Penalty	500	50	0.73 (0.03)	0.97 (0.03)	0.00 (0.00)	473.25 (1.67)
Plaid-Full	500	50	0.95 (0.09)	1.00 (0.00)	0.00 (0.00)	475.00 (0.00)
Plaid-Restricted	500	50	0.86 (0.06)	1.00 (0.00)	0.00 (0.00)	475.00 (0.00)
SK-Means	500	50	0.87 (0.00)	0.93 (0.01)	3.12 (0.64)	475.00 (0.00)
SVM	500	50	0.81 (0.10)	0.22 (0.08)	21.88 (1.36)	475.00 (0.00)

expression, microRNA expression, protein expression, single nucleotide polymorphism and copy number variations across 20 different cancers (`http://cancergenome.nih.gov`). We applied our methodology to 473 samples from TCGA KIRC data, using the mRNA log-expression information collected from the Illumina HiSeq2000 platform (which contains approximately 20,000 protein coding genes). This data set

Plaid-Restricted					
	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4
Sample size	11 (11)	260 (260)	47 (47)	24 (24)	131 (131)
μ_k	0.77	1.99	-1.13	2.47	1.13
Plaid-Full					
	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4
Sample size	20 (20)	248 (62)	227 (81)	130 (103)	48 (13)
μ_k	-0.12	1.58	0.61	1.33	0.98

Table 4 Overall means of each clustering. The numbers between parentheses are the number of samples that belong only to the corresponding cluster.

was assessed by The Cancer Genome Atlas Research Network (2013), who used unsupervised clustering to identify four molecular subsets in mRNA expression data that were associated with patient survival times. We first reduced the number of genes by taking the standard deviations (SDs) of all genes, then looking at the mean of the SDs; the distribution of the SD for all genes ranged between 0.1 and 4.3, with a mean SD of 0.7. The SD of the SD was 0.4. For our analysis, we selected genes with SD above the mean $+1$-SD, for a total number of 2835 genes. We then removed genes that contained more than 30% missing data, which removed 439 genes, leaving a total of 2396 genes. The remaining missing data were imputed using the k-nearest neighbor imputation method, with $k = 10$ of Troyanskaya et al (2001).

We fitted the plaid model for $K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The BIC (and AIC) selected $K = 4$ (5 clusters with the zero cluster included). Both plaid models (Plaid-Restricted and Plaid-Full) deemed that 156 genes (approximately 6% of the genes considered) were discriminative. However, they shared only 133 genes in common. The overall means μ_k's and the sizes of each cluster are summarized in Table 4. We see that the smallest size is the zero cluster for both clustering methods. In addition, there are 141 samples that belong to clusters 1 and 2. In general, only a small number of samples belong to more than one cluster. For example, only one sample belongs to clusters 1, 2 and 4. Only two samples belong to clusters 1, 2 and 3; and only two samples belong to clusters 2 and 3. Only three samples belong to clusters 1, 3 and 4; 2 and 4; and 3 and 4.

To determine whether our clustering schemas are associated with survival outcomes (65% observations are censored), we fitted two multivariate Cox regressions with covariates as clusters obtained from the Plaid-Full and the Plaid-Restricted clusterings. In addition to these covariates, we added some prognostic

factors such as age, sex and tumor stage. Clusters 1, 2 and 3 from the Plaid-Full clustering results provide negative associations with survival time (p-values$<$0.03) with respective hazard ratios of 0.51, 0.62 and 0.60. Similarly, clusters 1, 3 and 4 also have strong negative associations, with respective relative risks of 0.19, 0.1782 and 0.30 (p-values$<$0.002) compared to the 0 cluster. Both regressions have good predictive performance, with concordance indices of about 0.78. To compare the associated survival time between clusters, we performed two stratified Cox regressions with strata as clusters. Clusters from the Plaid-Full method were partitioned by combining samples that belong to both clusters 1 and 2, 1 and 3, and 1 and 4. Figure 1 from the stratified Cox regressions shows that clusters are associated with different survival outcomes for each clustering method. In particular, for both of our proposed clustering methods, samples in the zero cluster are associated with short survival times. Those belonging to both clusters 1 and 2 (cluster 12) are associated with long survival times. Clusters 1 and 3 from the Plaid-Restricted method are also associated with long survival times. Note that these clusters have the largest overall gene expression means, μ_k.

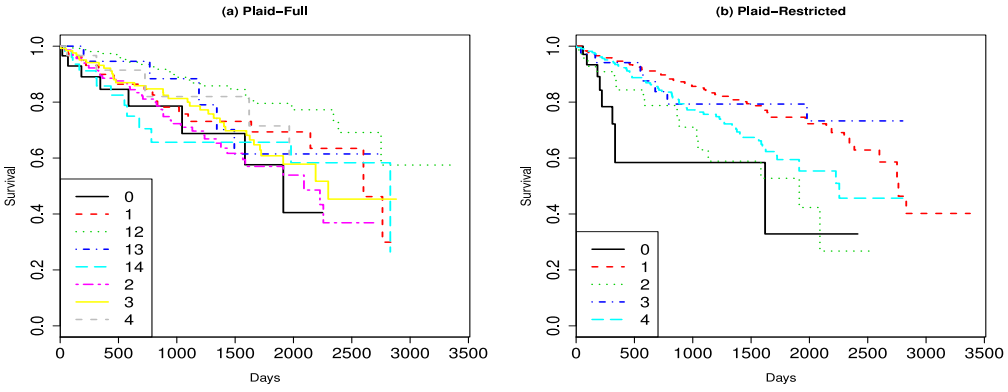

Fig. 1 The overall survival time associated with each non-overlapping cluster: (a) Plaid-Full method; (b) Plaid-Restricted method. Legend: 0 is zero cluster; 1 is cluster 1, which contains samples belonging only to cluster 1; 12 is the cluster with samples belonging only to clusters 1 and 2, etc. The other combinations of clusters are excluded as they contain a small number of samples (fewer than 4).

Using Ingenuity Pathway Analysis (IPA)[1], we determined which top-ranked biological function and disease categories would be statistically overrepresented with our 151 discriminating genes. An analysis of the over-represented diseases and disorders with our set of genes (p-value$<$0.001, Figure 2) shows that

[1] IPA (Ingenuity® Systems, `www.ingenuity.com`) is a software for interactive pathway analysis of complex 'omic data.

cancer is the most represented. Many other diseases or disorders related to kidney cancer such as renal and urological disease and metabolic disease are also over-represented. Linehan et al (2010) showed that more effective forms of therapy for kidney cancer can be achieved by targeting the fundamental metabolic abnormalities present in this disease.

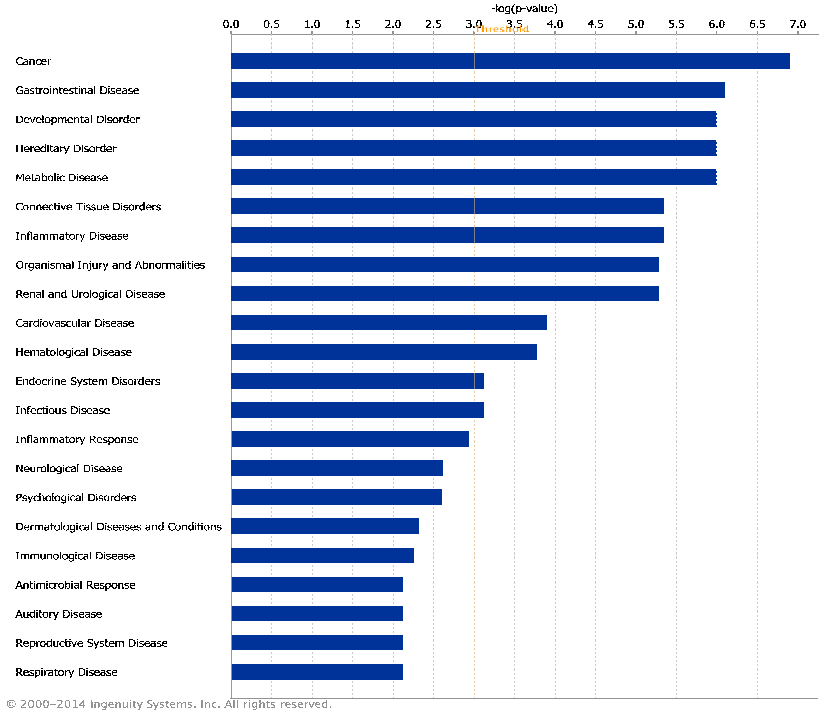

Fig. 2 Diseases and disorders over-represented.

7 Discussion and conclusions

In this work, we proposed two variable selection models that are inspired by the plaid model of Lazzeroni and Owen (2002). Our first model, the Plaid-Full model, assumes that each observation may be explained by more than one cluster, producing an aggregated clustering model. This model is related to the multiplicative mixture model for overlapping clustering that was developed by Fu and Banerjee (2008, 2009), and Heller and Ghahramani (2007). Our second model, the Plaid-Restricted model, forces each observation to belong to only one cluster. This model is more conventional in the sense that clusters are modeled as separate objects.

We can also link our models to an extreme type of *biclustering* (Madeira and Oliveira, 2004; Tanay et al, 2005; Chekouo and Murua, 2015). Biclustering is the simultaneous clustering of the observations (rows) and variables (columns) of a data matrix. The biclusters obtained are submatrices in which the rows exhibit a similar pattern across a subset of columns and vice versa. Note that when the same subset of columns is selected for each bicluster, then we have really obtained a clustering of the observations given by a selected subset of variables. This is a key observation that links our multiplicative plaid mixture model for simultaneous clustering and variable selection to a very particular case of biclustering. We stress that the methodology we proposed herein is not for biclustering. Rather, one can think of our model as an adaptation of the plaid model to the problem of variable selection within the framework of *clustering* (as opposed to *biclustering*).

We would like to stress that our model is cast into a Bayesian framework, and a full MCMC computational approach is possible. In particular, this would allow us to estimate the parameters associated with the compound symmetry (positive correlation) between the variables. However, in this work we have favored a faster estimation algorithm on a simpler model that only considers fixed effects. This is a Monte Carlo EM algorithm that efficiently estimate the parameters of our models. Despite the restriction of the simpler model, we have been able to show through our simulations that the simpler model performs very well and appears to be robust against the hypothesis of positively correlated variables. Furthermore, we also showed, through extensive simulations, that (a) the performance of the plaid models in terms of discriminative variable detection is much better than the performance of competing models such as the L_1-Penalty method of Pan and Shen (2007), the Gaussian model-based clustering for variable selection method of Maugis et al (2009b), and the SK-Means method of Witten and Tibshirani (2010); and (b) the performance of the plaid models in terms of quality of clustering is comparable to that of the aforementioned models. Our simulation study revealed that when the number of variables is large, the $AIC_{IS,plaid}$ criterion appears to select better models than the $BIC_{IS,plaid}$ criterion (See Table 3 of the supplementary material). This was a bit surprising given the popularity of BIC in the clustering literature. It appears that BIC over-penalizes larger models due to the large number of variables involved in the models.

The application of our methodology to kidney cancer data showed the usefulness of the plaid models. We found clusters that can be differentiated by the associated survival times. Moreover, the discriminating biomarkers (variables selected) found by the plaid models are related to kidney cancer.

In a Bayesian framework, prior distributions similar to those considered in the work of Chekouo and Murua (2015) within the context of biclustering may be incorporated within the context of clustering as well. Posterior inference may be achieved through an appropriate MCMC algorithm. In particular, the number of clusters $K+1$ could be made a model parameter. In the case of the standard clustering (non-aggregate clustering), many Bayesian approaches have been proposed. A popular choice, is the use of nonparametric Bayesian models, such as the Dirichlet process, to model the prior probabilities of variable inclusion. However, one would need to adapt such processes to the case of aggregate clusters (to our knowledge, this has not yet been done, and it does not seem easy to do). Another possibility would be to assume that K follows a uniform or truncated Poisson distribution. The use of reversible jump techniques may be useful in these latter cases.

Supplementary Materials

The accompanying supplementary document presents: a more detailed description of the similarity of our model with the multiplicative mixture model (Section A); further details on the EM updating equations and the Monte Carlo error (Section B), the simulation setup (Section C), and the effective number of parameters, including a comparison between AIC and BIC results (Section D).

Acknowledgements The authors are grateful to LeeAnn Chastain at MD Anderson Cancer Center for editing assistance.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723

Allan J, Carbonell J, Doddington G, Yamron J, Yang Y (1998) Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp 194–218

Bhattacharya AK (2005) Evaluation of headache. Journal, Indian Academy of Clinical Medicine 6(1):17–22

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):719–725

Booth JG, Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society Series B (Statistical Methodology) 61(1):265–285

Chekouo T, Murua A (2015) The penalized biclustering model and related algorithms. Journal of Applied Statistics 42(6):1255–1277

Fu Q, Banerjee A (2008) Multiplicative mixture models for overlapping clustering. In: Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on, pp 791 –796

Fu Q, Banerjee A (2009) Bayesian overlapping subspace clustering. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, pp 776–781

George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. Statistica Sinica (7):339–374

Heller KA, Ghahramani Z (2007) A nonparametric Bayesian approach to modeling overlapping clusters. Journal of Machine Learning Research - Proceedings Track 2:187–194

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800

Hoff PD (2006) Model-based subspace clustering. Bayesian Analysis 1(2):321–344

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2:193–218

Kim S, Tadesse MG, Vannucci M (2006) Variable selection in clustering via Dirichlet process mixture models. Biometrika 93(4):877–893

Lazzeroni L, Owen A (2002) Plaid models for gene expression data. Statistica Sinica 12:61–86

Levine R, Casella G (2001) Implementations of the Monte Carlo EM algorithm. Journal of Computational and Graphical Statistics 10(10):422–439

Li F, Zhang NR (2010) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association 105(491):1202–1214

Linehan W, Srinivasan R, Schmidt L (2010) The genetic basis of kidney cancer: a metabolic disease. Nature reviews Urology 7(5):277–285

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans Comput Biol Bioinformatics 1(1):24–45, DOI 10.1109/TCBB.2004.2, URL http://dx.doi.org/10.1109/TCBB.2004.2

Maugis C, Celeux G, Martin-Magniette ML (2009a) Variable selection for clustering with gaussian mixture models. Biometrics 65(3):701–709

Maugis C, Celeux G, Martin-Magniette ML (2009b) Variable selection in model-based clustering: A general variable role modeling. Computational Statistics and Data Analysis 53(11):3872 – 3882

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association 92(437):162–170

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

Raftery AE, Dean N (2006) Variable selection for model-based clustering. Journal of the American Statistical Association 101:168–178

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66:846–850

Robert C, Casella G (2004) Monte Carlo Statistical Methods. Springer Texts in Statistics, Springer

Robert CP, Rydn T, Titterington D (1999) Convergence controls for MCMC algorithms, with applications to hidden Markov chains. Journal of Statistical Computation and Simulation 64(4):327–355

Schwarz GE (1978) Estimating the dimension of a model. Annals of Statistics 6(2):461–464

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B 64:583–639(57)

Tadesse MG, Sha N, Vannucci M (2005) Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association 100:602–617, URL http://EconPapers.repec.org/RePEc:bes:jnlasa:v:100:y:2005:p:602-617

Tanay A, Sharan R, Shamir R (2005) Biclustering algorithms: A survey. In: Handbook of Computational Molecular Biology Edited by: Aluru S. Chapman and Hall/CRC Computer and Information Science Series

The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–49

Tibshirani R, Walther G, Hastie T (2000) Estimating the number of clusters in a dataset via the gap statistic 63:411–423

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(6):520–525

Wang S, Zhu J (2008) Variable selection for model-based high-dimensional clustering and its application to microarray data. Biometrics 64(2):440–448, DOI 10.1111/j.1541-0420.2007.00922.x, URL http://dx.doi.org/10.1111/j.1541-0420.2007.00922.x

Wei GCG, Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85(411):699–704

Witten DM, Tibshirani R (2010) A framework for feature selection in clustering. Journal of the American Statistical Association 105(490):713–726

Xie B, Pan W, Shen X (2008) Variable selection in penalized model-based clustering via regularization on grouped parameters. Biometrics 64(3):921–930, DOI 10.1111/j.1541-0420.2007.00955.x, URL http://dx.doi.org/10.1111/j.1541-0420.2007.00955.x

Zhou H (2009) Manual for program of the algorithm of Pan, W. and Shen, X. (2007). Available online at http://www.biostat.umn.edu/~weip/prog.html

Zhou H, Pan W, Shen X (2009) Penalized model-based clustering with unconstrained covariance matrices. Electron J Statist 3:1473–1496, DOI 10.1214/09-EJS487, URL http://dx.doi.org/10.1214/09-EJS487