Supervised and Unsupervised Model-Based Multi-Partitioning

Alessio Farcomeni and Maurizio Vichi
Department of Statistics, Probability and Applied Statistics
University of Rome "La Sapienza"
Email: alessio.farcomeni@uniroma1.it, maurizio.vichi@uniroma1.it

ABSTRACT:
Rocci and Vichi (2006) have recently introduced the two-mode multi-partitioning model with the aim to cluster both objects (rows) and variables (columns) of a two-way data matrix. The new methodology allows to partition the set of objects and to obtain a partition of the variables for each class of the partitions of the objects. In this paper a model-based approach in the field of the maximum likelihood clustering is proposed. The model is extended to the supervised classification framework. A specific algorithm is introduced and its performances are discussed by means of a simulation study. Finally, the new methodology is applied to data sets to show its features.

Keywords: Double partitioning, maximum likelihood clustering.

1. INTRODUCTION

In the last years increasing attention has been paid to new methodologies for clustering both objects and variables of a two-way two mode (objects, variables) data matrix. There are several situations where such methodologies can be applied. In general when a very large data set is observed it would be very useful to identify disjoint classes of objects which are perceived as similar to one another within each class. However, clusters of objects would be very likely similar only according to some subsets of observed variables that can vary passing from one cluster of objects to another. In other terms, different variable partitions for each set of the partition of the objects would be very likely expected. This clustering setting is called two-mode multi-partitioning (Rocci & Vichi, 2006).
For example, in marketing research customers are segmented according to the preference

on products, but also products are partitioned according to preferences of customers. Therefore, a partition of the customers and a simultaneous partition of the products is very useful. In particular a different partition of the products would be likely to be found for each class of the partition of the customers, because it is very easy that groups of people with different tastes would specify different clustering preferences of products.

In DNA Microarrays studies, researchers are interested in clustering tissue samples into homogeneous clusters according to the similarly expressed genes; however, researchers are also interested in finding partitions of genes that possibly can vary for each class of the partition of the tissue samples.

In this paper we introduce a new model-based multi-partitioning methodology for a two-way data matrix. The model is defined both in the supervised and unsupervised classification framework. It is known or to be assumed that the population of interest from which the data are observed consists of P different subpopulations. Conditionally on each class, variables are partitioned into Q_p independent possibly different blocks, i.e., there is a between block independence, which induces a possible different partition of the variables for each class of objects. Furthermore, assuming class conditional Gaussian distribution it is possible to define the complete log-likelihood of the double classification/clustering problems which can be maximized using a coordinate ascent algorithm of the type ECM (Expectation conditional maximization, (Meng and Rubin, 1993)).

To give an idea, Figure 1 shows the reordering of a DNA microarray experiment. The experiment will be described in Section 8. It can be seen nevertheless that genes (on the columns, here) are partitioned into 5 blocks; and that each block is divided into a different number of groups of slides (on the rows, here). For instance, the second group of genes is divided in two groups of slides, of which one is composed by only a single outlier.

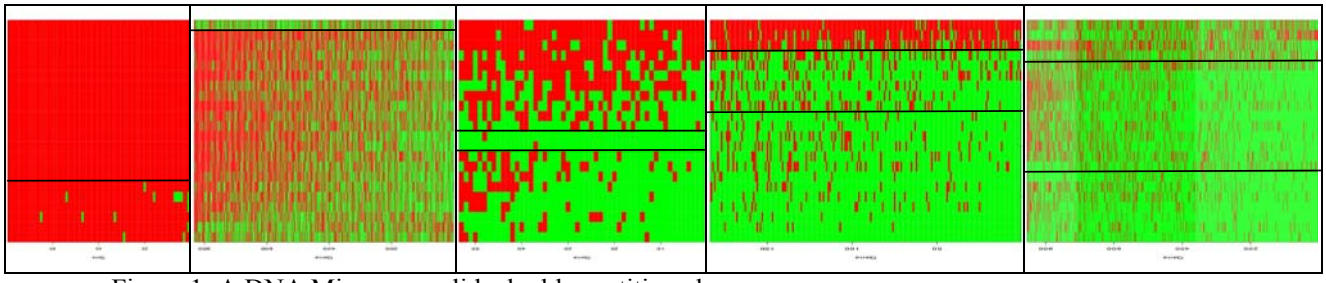

Figure 1: A DNA Microarray slide doubly partitioned

The rest of the paper is organized as follows. In section 2 notation and terminology used in this paper are listed for convenience of the reader. In section 3 the two mode multi-partitioning model under fixed and random effect formulations is introduced. Section 4 is devoted to the Maximum Likelihood (ML) estimation of the parameters via a coordinate ascent algorithm of ECM type. A general parameterization of the class-conditional covariance matrix is given in section 5. In section 6 a simulation study illustrates the performances of the new methodology; while other simulations in section 7 show the possible performance of information criteria for model selection. An application of the technique is given in section 8, and a final discussion follows in section 9.

2. NOTATION

For the convenience of the reader, the terminology used in this paper is listed here:

I, J	number of units, variables, respectively;
P, Q	number of classes of the partition for units and variables;
$C_1,C_2,\ldots,C_P$	partition in P clusters of units;
$V_{1p},V_{2p},\ldots,V_{Qp,p}$	partition of the variables for the p^{th} cluster of units;
$\mathbf{X}=[x_{ij}]$	$(I \times J)$ data matrix; where x_{ij} is the value of the j^{th} variable observed on the i^{th} object. Variables are supposed to be commensurable in order to avoid clustering of the units depending on the variables unit of measurements. If this is not the case variables are supposed appropriately standardized;
$\mathbf{U} = [u_{ip}]$	$(I \times P)$ membership function matrix defining a partition of units, into P classes, where $u_{ip} = 1$ if the i^{th} object belongs to class p, $u_{ip} = 0$ otherwise.

Matrix **U** is constrained to have only one nonzero element per row;

I_p cardinality of cluster C_p, i.e. $I_p = |C_p| = \sum_{i=1}^{I} u_{ip}$;

J_{pq} cardinality of cluster V_{qp};

$\mathbf{x}_i$, $\mathbf{u}_i$, $\boldsymbol{e}_i$, column vectors representing the i^{th} row of **X** and **U**, respectively;

$\mathbf{x}_{.j}$ column vector representing the j^{th} variable of **X**.

3. THE MODEL

Let the population, from which the data are observed, be structured into P homogeneous subpopulations with elements in proportions

$$\pi_1, \pi_2, ..., \pi_P, \sum_{i=1}^{P} \pi_i = 1, \tag{1}$$

and let $\mathbf{x} = (x_1,\ldots,x_J)'$ be a multivariate variable characterizing the populations. The complete data is given by the $(J+P)$-dimensional vector $(\mathbf{x}_i', \mathbf{u}_i')'$, where the P-dimensional vector $\mathbf{u}_i$ is binary and has a unique non null element, (denoting whether unit i belongs to subpopulation p, (i=1,…,I, p=1,…,P)); it identifies the value of categorical non observable variables specifying the membership of the unit to a subpopulation.
The first classification/clustering assumption in P subpopulations can be formalized modeling data conditionally on class p, (p=1,…,P), by requiring that

$$\mathbf{x}_i = \boldsymbol{\mu}_p + \mathbf{e}_i \qquad (i = 1,\ldots,I) \tag{2}$$

where $\mathbf{e}_i$ is the random error with

(i) $E(\mathbf{e}_i)=\mathbf{0}$, (3)
(ii) $Cov(\mathbf{e}_i)=\boldsymbol{\Sigma}_p$ (4)
(iii) $Cov(\mathbf{x}_i, \mathbf{x}_l) = \mathbf{0}$, for all i,l=1,…,I, ($i \neq l$), (5)

Thus: 1. the J-column vector $\boldsymbol{\mu}_p$ is the expected value of the random vector $\mathbf{x}_i$; 2. heteroskedastic subpopulations are assumed (from condition (ii)); 3. a random sample from **x** is drawn.
The second assumption we make is on within-cluster distribution of the objects $\mathbf{x}_i$. Conditionally on class p, (p=1,…,P), the density distribution of $\mathbf{x}_i$ is

$$\mathbf{x}_i \sim f_p(\mathbf{x};\boldsymbol{\theta}_p) = \mathrm{MVN}_J(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p). \tag{6}$$

The third assumption models the membership of the objects to subpopulations identified by

$\mathbf{u}_i$. Two cases can be distinguished whether a fixed or a random effect assumption on the clustering model is supposed.

In general, variable $\mathbf{u}_i$ can be:

1. a fixed effect in the clustering problem, where for any random sample of I units, respectively $I\pi_1, I\pi_2, ..., I\pi_P$ units are always supposed to belong to the P subpopulations; in other terms the number of objects belonging to each population is fixed from sample to sample and thus $\mathbf{u}_i$ cannot be considered as a random variable.
2. a random effect in the clustering problem, where for any random sample of I units a random number of objects belongs to the P classes of the partition, i.e.

$$\mathbf{u}_i \sim \prod_{P=1}^{P} \pi_p^{u_{ip}} \tag{7}$$

has a Multinomial distribution of one draw on P categories with probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_P)$. Therefore, in this case, $\mathbf{u}_i$ has to be considered as a random variable.

The fourth assumption models the dependence structure within row clusters. Thus, conditionally on class p, (p=1,…,P), variables are partitioned into Q_p independent blocks, i.e., there is a between blocks independence.
Recall that the row cluster specific mean is partitioned as: $\boldsymbol{\mu}_p = [\boldsymbol{\mu}'_{p1}, ..., \boldsymbol{\mu}'_{pq}, ..., \boldsymbol{\mu}'_{pQ_p}]'$.
Formally, for each unit i^{th} belonging to p^{th} class, $\mathbf{x}_i = [\mathbf{x}'_{i1_p}, ..., \mathbf{x}'_{iQ_p}]'$,

$$cov(\mathbf{x}_{iq_p}, \mathbf{x}_{ir_p}) = \mathbf{0}, \qquad (q,r = 1, ...,Q_p \text{ for } r \neq q), \tag{8}$$

therefore, the covariance structure of the matrix $\boldsymbol{\Sigma}_p$ is

$$\boldsymbol{\Sigma}_p = \text{diag}(\boldsymbol{\Sigma}_{p1}, \boldsymbol{\Sigma}_{p2}, ..., \boldsymbol{\Sigma}_{pq},...,\boldsymbol{\Sigma}_{pQp}). \tag{9}$$

where, $\boldsymbol{\Sigma}_{pq}$ is the covariance matrix of order J_{pq} of the random vector $\mathbf{x}_{iq_p}$ and diag(.) specifies a block diagonal matrix. The between blocks independence could be relaxed without problems. We keep it here because it is sensible in some applications. For instance, in DNA Microarrays genes are thought to be dependent in blocks (the so called *clumpy dependence*); and in local inference approximate independence of submatrices may be desirable. Moreover, between-block independence avoids some numerical instability problems and allows for a fast and efficient algorithm for maximizing the likelihood. For further discussion on covariance modeling see Section 5.

3.1 Fixed Effect Multi-Partitioning Model

For a fixed effect classification model it is assumed that the sample is drawn from the population and the P clusters have cardinality equal to $I_p = I\pi_p$, (p=1,⋯,P) from sample to

sample. This assumption is realistic for large data sets.
Let ($\mathbf{x}_1$, $\mathbf{u}_1$; $\mathbf{x}_2$, $\mathbf{u}_2$; …; $\mathbf{x}_I$, $\mathbf{u}_I$) be a random sample of I multivariate units drawn from the population under the above specified scheme.
The classification (supervised) problem here considered, (i.e., supposing each couple ($\mathbf{x}_i$, $\mathbf{u}_i$) is completely observed) is characterized by the complete-data likelihood

$$L_C(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_P)=\prod_{i=1}^{I}\prod_{p=1}^{P}\left[f_p(\mathbf{x}_i,\mathbf{u}_i;\boldsymbol{\theta}_p)\right]^{u_{ip}}. \tag{10}$$

With the assumption (6) and supposing a fixed model effect for the $\mathbf{u}_i$, the complete data likelihood can be written:

$$L_C(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;...;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)=\prod_{i=1}^{I}\prod_{p=1}^{P}\left[(2\pi)^{-J/2}\left|\Sigma_p\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_p)'\boldsymbol{\Sigma}_p^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_p)\right\}\right]^{u_{ip}} \tag{11}$$

Now, without loss of generality let $\boldsymbol{\mu}_p$ be partitioned accordingly to $\mathbf{x}_i$, that is $\boldsymbol{\mu}_p=[\boldsymbol{\mu}'_{p1}, ..., \boldsymbol{\mu}'_{pq}, ..., \boldsymbol{\mu}'_{pQ_p}]'$. Therefore, the likelihood (11) can be written

$$L_C(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;...;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)=\prod_{i=1}^{I}\prod_{p=1}^{P}\left[\prod_{q=1}^{Q_p}(2\pi)^{-J_{pq}/2}\left|\Sigma_{pq}\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})'\boldsymbol{\Sigma}_{pq}^{-1}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})\right\}\right]^{u_{ip}} \tag{12}$$

where J_{pq} is the number of variables in the q^{th} block for the p^{th} cluster of units.
When there is a clustering problem, i.e., values $\mathbf{u}_i$, (i=1,…,I) are not observed, it is assumed that they are "missing" in the observed sample and they have to be estimated. Passing to the complete log-likelihood we have

$$-2l_C(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;...;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P,u_{ip})=\sum_{i=1}^{I}\sum_{p=1}^{P}u_{ip}\left[J\log(2\pi)+\sum_{q=1}^{Q_p}\log\left|\Sigma_{pq}\right|+\sum_{q=1}^{Q_p}(\mathbf{x}_{iq}-\boldsymbol{\mu}_{pq})'\boldsymbol{\Sigma}_{pq}^{-1}(\mathbf{x}_{iq}-\boldsymbol{\mu}_{pq})\right] \tag{13}$$

Thus, our task is to partition the observed data matrix in p clusters of units (row groups); each class of objects is characterized by a possibly different partition of the variables in Q_p classes (column groups), and variables do not necessarily belong to the same column groups across row groups.

3.2 Random Effect Multi-Partitioning Model

Again let ($\mathbf{x}_1$, $\mathbf{u}_1$; $\mathbf{x}_2$, $\mathbf{u}_2$; …; $\mathbf{x}_I$, $\mathbf{u}_I$) be a random sample of I multivariate units drawn from the population under the mixture sampling scheme, corresponding to drawing for each unit first its class value $\mathbf{u}_i$, from the population with p.d.f. (7) and second drawing values of $\mathbf{x}_i$ from the population with c.p.d.f. (6).
The supervised classification problem here considered is characterized by the complete-data likelihood

$$L_M(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_P)=\prod_{i=1}^{I}\prod_{p=1}^{P}\left[f_p(\mathbf{x}_i;\boldsymbol{\theta}_p)\pi_p\right]^{u_{ip}}. \quad (14)$$

With the assumption (6) and (9) the complete data likelihood can be written:

$$L_M(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;.;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P,\pi_p)=\prod_{i=1}^{I}\prod_{p=1}^{P}\left[\pi_p\prod_{q=1}^{Q_p}(2\pi)^{-J_{pq}/2}\left|\Sigma_{pq}\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})'\boldsymbol{\Sigma}_{pq}^{-1}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})\right\}\right]^{u_{ip}} \quad (15)$$

Passing to the complete data log-likelihood we have

$$-2l_M(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;...;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P,\pi_p)=\sum_{i=1}^{I}\sum_{p=1}^{P}u_{ip}\left[J\log(2\pi)+\sum_{q=1}^{Q_p}\log\left|\Sigma_{pq}\right|+\sum_{q=1}^{Q_p}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})'\boldsymbol{\Sigma}_{pq}^{-1}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})\right]+$$

$$+\sum_{i=1}^{I}\sum_{p=1}^{P}u_{ip}\log\pi_p, \quad (16)$$

which corresponds to the complete data log-likelihood of a mixture model with observed log-likelihood

$$l_O(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;.;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P,\pi_p)=\sum_{i=1}^{I}\ln\left[\sum_{p=1}^{P}\pi_p\prod_{q=1}^{Q_p}(2\pi)^{-J_{pq}/2}\left|\Sigma_{pq}\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})'\boldsymbol{\Sigma}_{pq}^{-1}(\mathbf{x}_{iq_p}-\boldsymbol{\mu}_{pq})\right\}\right] \quad (17)$$

Also in this case if there is a clustering problem, i.e., values $\mathbf{u}_i$, (i=1,…,I) are not observed, it is assumed that they are "missing" in the observed sample and they have to be estimated; therefore, in this case (16) is also function of $\mathbf{u}_i$, (i=1,…,I).

4. MODEL FIT

For the fixed effect multi-partitioning model, the maximization of the likelihood corresponds to the minimization of $-2l_C(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;...;\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P,u_{ip})$ that can be conveniently performed via a coordinate descent algorithm. At each step of the algorithm, each parameter vector or matrix is to be updated in turn by maximizing (12) or equivalently minimizing (13) with respect to one of the parameter matrices conditionally upon the others. The loss function -2l_C decreases at each step, or at least never increases, and the algorithm stops when the loss decrement is less than a arbitrary small positive threshold. Since -2l_C is bounded below, the monotonicity property of the algorithm guarantees that the sequence of function values converges to a stationary point, which usually turns out to be, at least, a local minimum.

4.1. The Clustering Algorithm in the Fixed Effect Case.

In the case of the fixed effect model, the basic steps of the previously introduced algorithm can be described as follows.

Step 1) Update $\boldsymbol{\mu}_{p1}, \boldsymbol{\Sigma}_{p1}; \ldots; \boldsymbol{\mu}_{PQ}, \boldsymbol{\Sigma}_{PQ}$ for fixed $\hat{\mathbf{U}} = [\hat{u}_{ip}]$

The estimation of the parameters $\boldsymbol{\mu}_{p1}, \boldsymbol{\Sigma}_{p1}; \ldots; \boldsymbol{\mu}_{PQ}, \boldsymbol{\Sigma}_{PQ}$ are obtained simply by using the sample mean and covariance of the entries of the matrix belonging to each block,

$$\hat{\boldsymbol{\mu}}_{pq} = \frac{1}{\sum_{i=1}^{I} \hat{u}_{ip}} \sum_{i=1}^{I} \hat{u}_{ip} \mathbf{x}_{iq}, \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_{pq} = \frac{1}{\sum_{i=1}^{I} \hat{u}_{ip}} \sum_{i=1}^{I} \hat{u}_{ip} \left(\mathbf{x}_{iq} - \hat{\boldsymbol{\mu}}_{pq}\right)\left(\mathbf{x}_{iq} - \hat{\boldsymbol{\mu}}_{pq}\right)'. \tag{19}$$

Step 2) Update $\mathbf{U}=[u_{ip}]$ for fixed $\hat{\boldsymbol{\mu}}_{p1}, \hat{\boldsymbol{\Sigma}}_{p1}; \ldots; \hat{\boldsymbol{\mu}}_{PQ}, \hat{\boldsymbol{\Sigma}}_{PQ}$ (the partition of the rows)

Function (13) can be rewritten

$$-2l_C\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1; \ldots; \hat{\boldsymbol{\mu}}_P, \hat{\boldsymbol{\Sigma}}_P, u_{ip}\right) = \sum_{i=1}^{I} \sum_{p=1}^{P} u_{ip} \left[f_{ip}\right] \tag{20}$$

where $f_{ip} = J \log(2\pi) + \sum_{q=1}^{Q_p} \log\left|\hat{\boldsymbol{\Sigma}}_{pq}\right| + \sum_{q=1}^{Q_p} \left(\mathbf{x}_{iq_p} - \hat{\boldsymbol{\mu}}_{pq}\right)' \hat{\boldsymbol{\Sigma}}_{pq}^{-1} \left(\mathbf{x}_{iq_p} - \hat{\boldsymbol{\mu}}_{pq}\right)$ is constant, when function (13) is minimized with respect to u_{ip}. To minimize function (13), we can observe that I independent assignment problems in the binary variables u_{ip} are at stake; whose solution is given by setting, for each v=1,..,P,

$$u_{iv} = \begin{cases} 1 & \text{if } f_{iv} = \min\left\{f_{ip}; p = 1,2,\ldots,P\right\} \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Note that the minimization of f_{ip} is equivalent to minimizing the weighted norm

$$\sum_{q=1}^{Qp} \left\| \mathbf{x}_{iq_p} - \hat{\boldsymbol{\mu}}_{pq} \right\|_{\boldsymbol{\Sigma}^{-1}}^2, \tag{22}$$

that corresponds to the sum of the squared Mahalanobis distances between the row profile $\mathbf{x}_{iq_p}$ and the corresponding centroid $\hat{\boldsymbol{\mu}}_{pq}$. Therefore, the i^{th} unit is assigned to the closest cluster in terms of Mahalanobis distance, and each column to the group leading to maximization of the likelihood.

Step 3) Update $\mathbf{x}_i = [\,\mathbf{x}'_{i1_p},\ \ldots,\ \mathbf{x}'_{iQ_p}\,]'$, the partition of the columns for a fixed row-partition

To mimimize function (13) for the partition of the columns we have to minimize (20) for each x_{ij} ($\underline{i}$=1,⋯,I), that is, for each variable j^{th}

$$\mathbf{x}_j \in \mathbf{x}_{iq_p} \text{ if } \sum_{i=1}^{I}\sum_{p=1}^{P}\hat{u}_{ip}\sum_{q=1}^{Q_p}\left\|\mathbf{x}_{iq_p}-\hat{\boldsymbol{\mu}}_{pq}\right\|^2_{\boldsymbol{\Sigma}^{-1}} = \min\{\sum_{i=1}^{I}\sum_{p=1}^{P}\hat{u}_{ip}\sum_{v=1}^{Q_p}\left\|\mathbf{x}_{iv_p}-\hat{\boldsymbol{\mu}}_{pq}\right\|^2_{\boldsymbol{\Sigma}^{-1}} : \mathbf{x}_j \in \mathbf{x}_{iv_p}, v=1,\cdots,Q_p\ (v\neq q)\}.$$

Now it is well known that the estimation procedures may encounter problems when the number of parameters increases indefinitely with the sample size, and therefore it is necessary to check for consistency of this procedure in the fixed effects case.
The parameters of each normal distribution are in this case estimated on a single truncated distribution adding the tails of the other P-1 distributions (mainly, all those objects which are closer to the distribution of interest according to (22)). This biases the estimates inducing inconsistent estimators.
The classification parameter is based on the Mahalanobis distance, which is based on the not consistent covariance matrix. Hence, also the classification parameter is not consistent.
Consistency of the previous estimators is guaranteed only when all the normal distributions have equal proportions (see next section).
Therefore, it is necessary to properly estimate the proportions and consistently estimate the vector means and covariance matrices of the mixture of normal distributions.
In practice this corresponds to replacing step 1 with

Step 1a) Update $\boldsymbol{\mu}_{p1}, \boldsymbol{\Sigma}_{p1}; \ldots; \boldsymbol{\mu}_{PQ}, \boldsymbol{\Sigma}_{PQ}$ (p=1,⋯,P), $\hat{\mathbf{V}} = [\,\hat{v}_{ip}\,]$ and

$$\hat{v}_{ip} = \frac{f_p\left(\mathbf{x}_i;\boldsymbol{\theta}_p\right)}{\sum_{j=1}^{P} f_j\left(\mathbf{x}_i;\boldsymbol{\theta}_j\right)}, \tag{23}$$

$$\hat{\boldsymbol{\mu}}_{pq} = \frac{1}{\sum_{i=1}^{I}\hat{v}_{ip}}\sum_{i=1}^{I}\hat{v}_{ip}\mathbf{x}_{iq}, \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}_{pq} = \frac{1}{\sum_{i=1}^{I}\hat{v}_{ip}}\sum_{i=1}^{I}\hat{v}_{ip}\left(\mathbf{x}_{iq}-\hat{\boldsymbol{\mu}}_{pq}\right)\left(\mathbf{x}_{iq}-\hat{\boldsymbol{\mu}}_{pq}\right)'. \tag{25}$$

With (23), we take the expectation of u_{ip} conditionally on the observed data, thus having a formal E-step. In (24) and (25) we have a closed form solution for the formal M-step.
After convergence, the last estimates in step 1 give the optimal clustering.
Now, the mean vectors and the covariance matrices of the normal distributions are consistently estimated, so as the final assignment of the function (21) since the Mahalanobis distance is consistently estimated.
Updating the parameters of the multinormals distributions according to step 1a the loss (13) still decreases at each step, or at least never increases.

4.2. The Clustering Algorithm in the Random Effect Case.

In the case of a random effect model, the algorithm is based on the steps 1a and 3 only, which decrease the loss function (17) (replacing in it u_{ip} with v_{ip}) if one replaces (23) with:

$$\hat{v}_{ip} = \frac{\hat{\pi}_p f_p\left(\mathbf{x}_i;\boldsymbol{\theta}_p\right)}{\sum_{j=1}^{P}\hat{\pi}_p f_j\left(\mathbf{x}_i;\boldsymbol{\theta}_j\right)} \tag{26}$$

and estimate

$$\hat{\pi}_p = \frac{1}{I}\sum_{i=1}^{I}\hat{v}_{ip}\ . \tag{27}$$

Once again, we are using an expectation-maximization algorithm.
At the end if we need a row-partition (since a fuzzy partitioning is obtained) we then have to assigning objects to clusters with a maximum a-posteriori strategy. Conditional expectation (26) comes from the complete data log-likelihood for the random effect case. As noted, if we assume $\pi_{p=}1/P$, (16) comes back to (13) and (26) to (23).

4.3 The Classification Algorithm

In the classification case the partition of the rows is known a priori, thus the multi-partitioning problem simplifies to the problem of finding the best partition of the columns of the data matrix for each class of the rows. This corresponds to fixing $\mathbf{u}_i$ according to the observed data (step 2) and to iterate steps 1a and 3 until convergence.

A new object can be assigned to a row group by minimizing the distance from the estimated centroid, that is, by applying "step 2" after convergence.

Formattato: Allineato a sinistra

4.4 Starting Solutions and Model Choice

In both fixed and random cases the algorithm may stop at local minima of the -2log-likelihood, therefore we suggest to use a classical multistart procedure to increase the chance to identify the global optimal solution. However, there are various possibilities for choosing the starting values. We compared a random assignment of rows and columns (pure random multistart), together with classical non-hierarchical clustering methods like *k*-means and partitioning around medoids (PAM): the row-groups are assigned through any of these methods, then the matrix is transposed; and for each row cluster, column clusters are assigned via the same one-way clustering method. Simulations showed that PAM is more likely to initialize the algorithm with a solution close to the optimal, also giving a lower number of iterations.

Another important problem is model choice.

While in usual non-hierarchical clustering methods a single choice (for the number of row groups) must be done, here there are many possibilities: after choosing the number of row groups P, a vector of size P is to be specified, in which each entry gives the number of column groups for each row group. Automatic model choice may be desirable. In our case, usual likelihood based methods can be suggested, like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Such methods are based on comparing penalized log-likelihood for different models, and choosing the one performing better with the minimal amount of parameters.

AIC is defined as $-2\log(L)+2Ip$, where L is the maximized likelihood, and Ip is the number of parameters for the given model. BIC is very similar, and defined as $-2\log(L)+\log(I)Ip$, where I is the number of observations. In all cases, the chosen model is the one achieving lowest index. It is intuitive the BIC in general performs a stronger penalization than AIC and so it will tend to favor more parsimonious models. As we will see in Section 7, both criteria seem to be likely to choose the right model with high probability in our case.

5. PARAMETERIZATIONS OF THE COVARIANCE MATRICES

In our general approach we assumed a different covariance matrix for each block.
This is quite flexible, but in many cases a reduction in the number of parameters can be achieved by separately describing size, shape and orientation of each covariance matrix (Banfield & Raftery, 1993; Celeux and Govaert, 1995; Bensmail & Celeux, 1996).

Each covariance matrix can be decomposed, based on the spectral decomposition, as:

$$\hat{\mathbf{\Sigma}}_{pq} = a_{pq}\, \mathbf{B}_{pq}\, \mathbf{D}_{pq}\, \mathbf{B}'_{pq}, \tag{27}$$

where a_{pq} is the largest eigenvalue (size), $\mathbf{D}_{pq}$ contains the diagonal matrix of eigenvalues divided by the largest eigenvalue (shape), and $\mathbf{B}_{pq}$ is the matrix of eigenvectors (orientation).

By assuming combinations of the described characteristics to be equal across blocks, one can achieve a suitable trade-off between number of parameters and fit.
For instance one could assume equal size and shape across blocks, but different orientation. The special case in which the matrices have the same orientation $\mathbf{B}_{pq} = \mathbf{B}$ leads to a particularly consistent and efficient reduction in the number of parameters, and is known as the common principal component model (see Flury, 1988; also for details on maximum likelihood estimation).

There are now many possible models, and as usual information criteria like BIC and AIC can be used to choose one.

6. SIMULATION STUDY

The design of the simulation is as follows. The number of clusters for rows is fixed as a *small* or *large* number, corresponding to *P=4* and *P=10* row-groups; the number of column-groups as Q_p=3 and Q_p=5. Furthermore, two data matrices sizes have been considered: 200 × 20 and 600 × 60.

Data were generated according to the model described in section 3, from *P* multinormal distributions. Three different separations between clusters where allowed corresponding to three general situations: clusters of units well separated (difference in mean between closest groups equal to 6 + random standard normal value); clusters separated with few overlaps (difference in mean between closest groups equal to 3 + random standard normal value); clusters with moderate overlap (difference in mean between closest groups equal to 1,5 + random standard normal value).

Finally, three different covariance matrices where considered, in which each entry is given by $e^{-(1/\tau)\, d((i,j),(i',j'))}$, where τ is a parameter controlling the dependence (the higher, the stronger the dependence) and *d(., .)* is just the Euclidean distance, computed between the indices of the position in the (reduced) data matrix.

For each cell of the simulation study, 1000 replications were computed.

The algorithm for fixed effect was used in the simulation study including the consistent estimation of the parameters (step 2a).

The results of the simulations were evaluated according to the average modified rand index (*M-Rand*), the proportion of times the algorithm manages to exactly find the correct clustering (*Exact*), and the proportion of times a Kruskal-Wallis test on the final clustering is significant at 5% level (*Sig*). In order to compute a single modified rand index we identified and vectorized the blocks given by the combination of the row and column partitions, and compared with the true vectorized blocks.

Table 1 shows the results for a 200 × 20 data matrix, with difference in mean between closest groups given by *6* plus a random value generated from a standard normal.

Table 2 shows the results for a 600 × 60 data matrix, with difference in mean between closest groups given by *6* plus a random value generated from a standard normal.

τ	*P*	*M-Rand*	*Exact*	*Sig*
0	4	0.99	0.98	1
5	4	0.99	0.99	1
20	4	0.99	0.99	1
0	10	0.99	0.55	1
5	10	0.99	0.61	1

20	10	0.99	0.56	1

Table1: Modified Rand Index, proportion of correct clustering and of significant clustering for 200 × 20 data matrix with separation *6* plus a standard normal.

τ	*P*	*M-Rand*	*Exact*	*Sig*
0	4	1	1	1
5	4	1	1	1
20	4	1	1	1
0	10	1	1	1
5	10	1	1	1
20	10	1	1	1

Table2: Modified Rand Index, proportion of correct clustering and of significant clustering for 600 × 60 data matrix with separation *6* plus a standard normal.

The tables show that when the separation between the groups is clear, a perfect clustering is very likely to occur. A lower proportion of perfect clustering is given in the 200 × 20 data matrix, when the number of row groups is 10. This happens because there can be very few observations in some blocks.

It is interesting to note that in this case the modified algorithm (including step 2a) and the original algorithm found solutions that could often differ only on the third decimal of the results in the table 1 and 2. This is so because clusters are very well distinct and therefore the estimation of the parameters based on truncated multinormals are not significantly different from those consistently estimated. As major examples, by using the unmodified algorithm we could observe a Table 4 exactly as it is, only with 0.99 instead of 0.93 as proportion of exact classification for the case of strong dependence; and the same for Table 5, only a 0.67 instead of 0.71 as proportion of exact classification for the independence case with 4 row groups.

Table 3 shows the results for a 200 × 20 data matrix, with difference in mean between closest groups given by 3 plus a random value generated from a standard normal, and Table 4 the same for a difference in mean given by 1.5 plus a random value generated from a standard normal.

Tables 5 and 6 show the same for the case of a 600 × 60 data matrix.

τ	*P*	*M-Rand*	*Exact*	*Sig*
0	4	0.97	0.54	1

5	4	1	1	1
20	4	1	1	1

Table 3: Modified Rand Index, proportion of correct clustering and of significant clustering for 200 × 20 data matrix with separation 3 plus a standard normal.

τ	*P*	*M-Rand*	*Exact*	*Sig*
0	4	0.83	0.03	1
5	4	0.99	0.98	1
20	4	0.99	0.93	1

Table4: Modified Rand Index, proportion of correct clustering and of significant clustering for 200 × 20 data matrix with separation 1.5 plus a standard normal.

τ	*P*	*M-Rand*	*Exact*	*Sig*
0	4	0.97	0.71	1
5	4	1	1	1
20	4	1	1	1
0	10	0.96	0.07	1
5	10	1	1	1
20	10	1	1	1

Table 5: Modified Rand Index, proportion of correct clustering and of significant clustering for 600 × 60 data matrix with separation 3 plus a standard normal.

τ	*G*	*M-Rand*	*Exact*	*Sig*
0	4	0.87	0.09	1
5	4	1	1	1
20	4	1	1	1
0	10	0.82	0.01	1
5	10	1	1	1
20	10	1	1	1

Table 6: Modified Rand Index, proportion of correct clustering and of significant clustering

for 600 × 60 data matrix with separation 1.5 plus a standard normal.

It can be seen that the model performance is very good also when the groups are not well separated; and that as the dependence within blocks increases, both the modified rand-index and the proportion of perfect clustering increase.

7. MODEL CHOICE

In order to evaluate the performance of AIC and BIC in the multi-partitioning model choice data matrices with size 200 × 20 were simulated. Different choices for the number of row and column groups, with the same setting as the previous section were considered.

For each simulation, we iterated 100 times and gave the proportion of times the AIC and BIC manage to choose the correct model in the space of all possible models generated by 3,4,5 row groups and 3,4,5 column groups for each row group.

Table 7 shows the results when the difference in mean between closest groups is given by *6* plus a random value generated from a standard normal, under independence.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.89,0.92)	(0.86,0.86)	(1.00,1.00)
4	(0.98,1.00)	(0.97,0.98)	(1.00,1.00)
5	(0.81,0.81)	(1.00,1.00)	(1.00,1.00)

Table 7: Proportion of times AIC and BIC manage to choose the correct model, for a 200 × 20 data matrix with separation *6* plus a standard normal, under independence.

Table 8 shows the results when the difference in mean between closest groups is given by 6 plus a random value generated from a standard normal, with τ=5.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.95,1.00)	(0.97,1.00)	(1.00,1.00)
4	(0.95,1.00)	(0.93,1.00)	(1.00,1.00)
5	(0.89,0.82)	(0.95,1.00)	(1.00,1.00)

Table 8: Proportion of times AIC and BIC manage to choose the correct model, for a 200 ×

20 data matrix with separation *6* plus a standard normal, with τ=5.

Table 9 shows the results when the difference in mean between closest groups is given by 6 plus a random value generated from a standard normal, with τ=20.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.96,1.00)	(0.91,1.00)	(1.00,1.00)
4	(0.96,1.00)	(0.89,1.00)	(1.00,1.00)
5	(1.00,1.00)	(1.00,1.00)	(1.00,1.00)

Table 9: Proportion of times AIC and BIC manage to choose the correct model, for a 200 × 20 data matrix with separation *6* plus a standard normal, with τ=20.

Table 10 shows the results when the difference in mean between closest groups is given by 3 plus a random value generated from a standard normal, under independence.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.84,1.00)	(0.72,0.73)	(0.71,0.71)
4	(0.86,0.89)	(0.74,0.79)	(0.69,0.69)
5	(0.70,0.70)	(1.00,1.00)	(1.00,1.00)

Table 10: Proportion of times AIC and BIC manage to choose the correct model, for a 200 × 20 data matrix with separation *3* plus a standard normal, under independence.

Table 11 shows the results when the difference in mean between closest groups is given by *3* plus a random value generated from a standard normal, with τ=*5*.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.72,1.00)	(0.77,1.00)	(0.80,0.93)
4	(0.97,1.00)	(0.82,1.00)	(0.93,0.94)
5	(1.00,0.54)	(1.00,1.00)	(0.96,0.96)

Table 11: Proportion of times AIC and BIC manage to choose the correct model, for a 200

× 20 data matrix with separation *3* plus a standard normal, τ=5.

Table 12 shows the results when the difference in mean between closest groups is given by *3* plus a random value generated from a standard normal, with τ=*20*.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(1.00,1.00)	(1.00,1.00)	(0.93,0.94)
4	(0.92,1.00)	(0.75,1.00)	(0.81,0.81)
5	(1.00,1.00)	(1.00,1.00)	(0.95,0.95)

Table 12: Proportion of times AIC and BIC manage to choose the correct model, for a 200 × 20 data matrix with separation *3* plus a standard normal, τ=*20*.

Table 13 shows the results when the difference in mean between closest groups is given by *1.5* plus a random value generated from a standard normal, under independence.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.74,1.00)	(0.70,0.75)	(0.64,0.68)
4	(0.74,0.84)	(0.70,0.75)	(0.62,0.74)
5	(0.67,0.60)	(0.91,0.93)	(0.84,0.84)

Table 13: Proportion of times AIC and BIC manage to choose the correct model, for a 200 × 20 data matrix with separation *1.5* plus a standard normal, under independence.

Table 14 shows the results when the difference in mean between closest groups is given by *1.5* plus a random value generated from a standard normal, τ=*5*.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(0.86,1.00)	(0.66,1.00)	(0.73,0.77)
4	(0.98,1.00)	(0.89,1.00)	(0.60,0.63)
5	(0.61,0.68)	(1.00,1.00)	(0.68,0.66)

Table 14: Proportion of times AIC and BIC manage to choose the correct model, for a 200

$\times$ 20 data matrix with separation *1.5* plus a standard normal, $\tau = 5$.

Table 15 shows the results when the difference in mean between closest groups is given by *1.5* plus a random value generated from a standard normal, $\tau = 20$.

Column Groups	3	4	5
Row Groups	(AIC, BIC)	(AIC, BIC)	(AIC, BIC)
3	(1.00,1.00)	(0.76,1.00)	(0.75,0.84)
4	(0.93,1.00)	(0.57,0.59)	(0.53,0.57)
5	(1.00,1.00)	(1.00,1.00)	(0.59,0.65)

Table 15: Proportion of times AIC and BIC manage to choose the correct model, for a 200 $\times$ 20 data matrix with separation *1.5* plus a standard normal, $\tau = 20$.

Though no measure seems to dominate, BIC seems to choose the correct model more frequently than AIC.

8. REAL DATA

8.1 Clustering Genes and Tissues

Clustering methods are one of the most frequently used tools in gene expression profiling, expecially for cancer (Alon *et.al.* (1999), Golub *et.al.* (1999)). Patterns in gene expression may lead to early diagnosis, and to the identification of genes connected with the disease.

We show an application of our methodology to the data from Alon *et.al.* (1999), The dataset refers to 2000 genes recorded on 62 individuals, with 22 safe and 40 ill of colon cancer.

After filtering, pre-processing, global normalization and computation of log-fold changes a data matrix of I=2000 genes by J=22 will be used for clustering.

Both BIC and AIC choose a model with 5 row groups, and (2,3,3,3,2) column groups.

1	**2**	**3**	**4**	**5**
79	183	53	853	832

Table 16: Number of genes for each row group.

Table 16 shows the number of genes for each row group. The column groups are illustrated in Figures 2 through 6, where expression for each gene at each slide are plot.

In group 1 over expressed genes are detected. It can be seen that the sample (i.e., column) clustering sensibly divides genes that are seen to be always over expressed with genes that are only often over expressed. Candidate under expressed genes are seen in groups 2 and 3, where in group 2 we have a block of slides (number 5,14 and 10) that do not follow the general tendency. Note further that in group 5 there is a group made up of the single outlying slide 20, while all the other slides are put in the same column group.

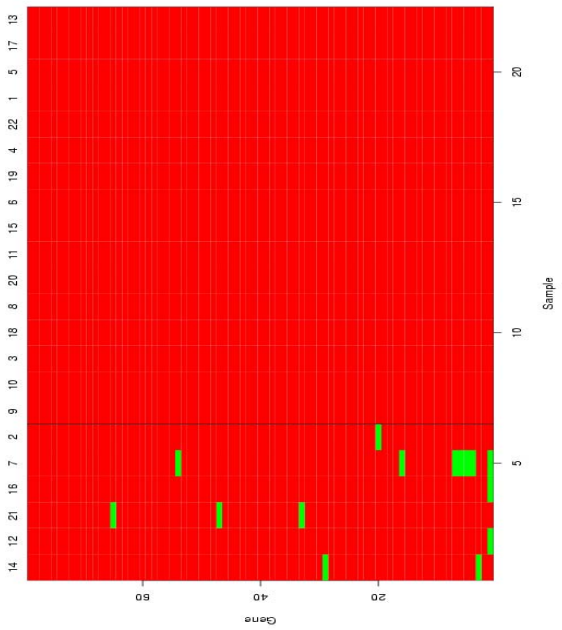

Figure 2. Gene expression data, first row cluster, split into 2 groups of samples.

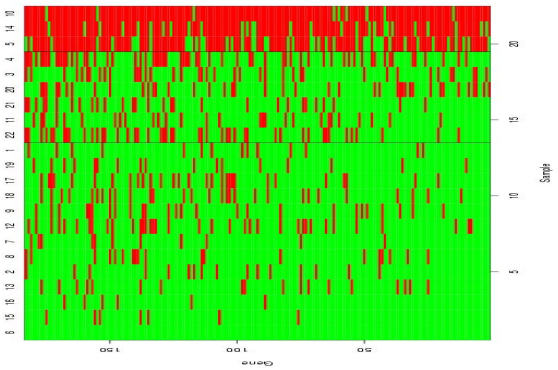

Figure 3. Gene expression data, second row cluster, split into 2 groups of samples.

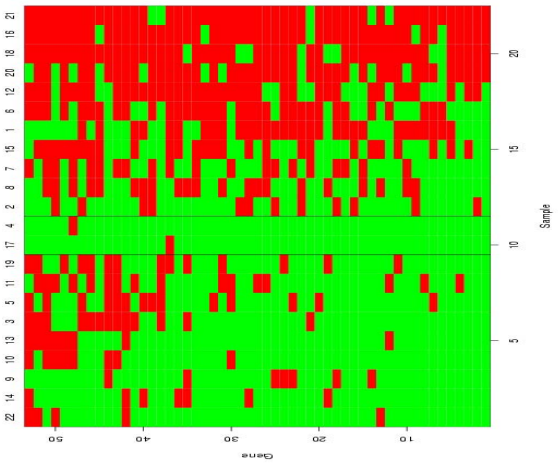

Figure 4. Gene expression data, third row cluster, split into 3 groups of samples.

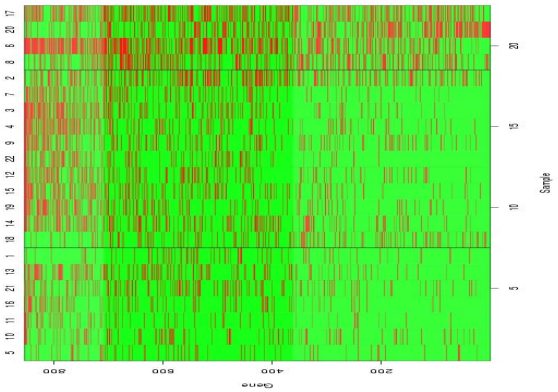

Figure 5. Gene expression data, fourth row cluster, split into 3 groups of samples.

.

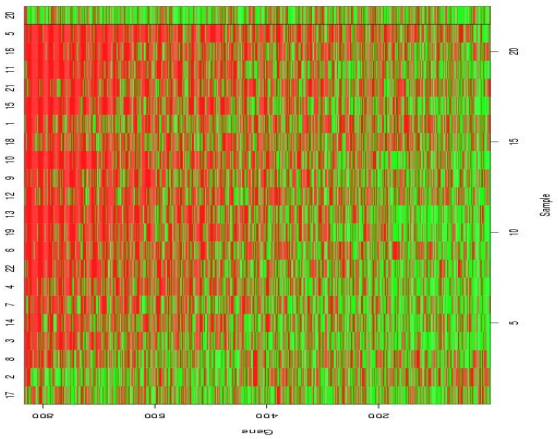

Figure 6. Gene expression data, fifth row cluster, split into 2 groups of samples.

8.2 Recognition of Glass and Ceramic Glass Fragments

Data come from an original study on discrimination between glass and ceramic glass fragments in recycling plants. Glass ceramic is hard to be manually sorted from glass, and melts at a much higher temperature. Thus, it is desirable to build automatic methods for distinguishing between glass ceramic fragments and glass, in-line during the recycling process, to avoid negative effects on the quality of recycled glass.

A detailed description of the study and the implications of the findings are illustrated in Farcomeni *et. al.* (2007).

The infrared spectra of 161 objects, of which 109 glass and 52 ceramic glass, was recorded at wavelengths intervals of 50 nm, from 1282 to 4482 nm.

Figure 7 shows the average absorbance for each wavelength sampled for glass (red line) and glass ceramic (green line). The dotted lines give 95% bands for the estimated mean.

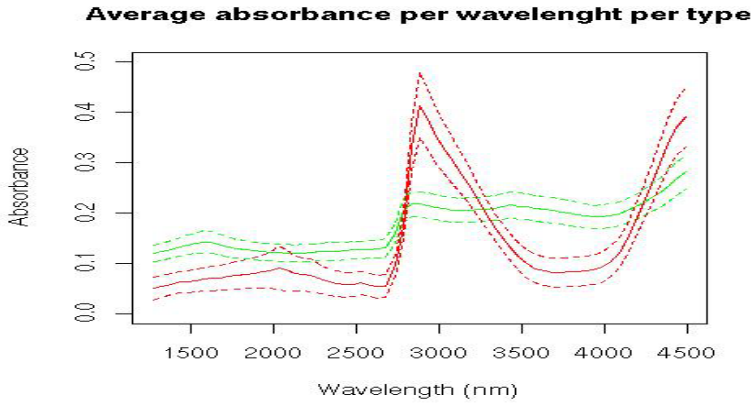

Figure 7: average absorbance of glass (green) and glass ceramic (red) fragments, with 95% bands.

Here we are interested both in clustering fragments, to check for separation between glass and ceramic glass (thus fixing P=2) and to identify which wavelengths better contribute to separation, thereby clustering also wavelengths. It is apparent in fact that certain ranges of wavelengths may be more useful than others in separating between the two groups, and a reduction on the number of sampled wavelengths is desirable in order to speed up the classification algorithms, which work in-line.

Both BIC and AIC lead us to choose Q_p=2 column groups for each row cluster p=1, 2. First, we check the classification of fragments in Table 17. By comparing it with Table 18, where we performed a PAM with 2 groups, we can see that even if the groups are not well separated there is some improvement by using a double clustering method.

	Glass	**Glass Ceramic**
Predicted Glass	79	40

Predicted Glass Ceramic	30	12

Table 17: Classification of fragments using double clustering

	Glass	**Glass Ceramic**
Predicted Glass	68	39
Predicted Glass Ceramic	41	13

Table 18: Classification of fragments using PAM

Interestingly enough, neighboring wavelengths in each row group are assigned to the same column group up to a point, which was not imposed by the algorithm but obviously sensible (and useful).

Further, Table 19 gives the difference between the estimated mean absorbance for each wavelength for each row cluster, allowing us to identify wavelengths between 3382 and 4132 as the most important for cluster separation.

These results are well in agreement with those obtained by Farcomeni *et. al.* (2007) with formal testing methods.

Wavelength	**Difference**
[1282-2732]	-0.14
[2782-3332]	-0.17
[3382-4132]	-0.27
[4182-4482]	-0.17

Table 19: Difference in estimated centroid for each wavelength

Table 19 illustrates well also the implications of the enhanced flexibility of our method: for each row group wavelengths are clustered differently. In fact, for the first row group we have wavelengths from 1282 to 2732 and from 3382 to 4132 in one column group and all the other wavelengths in the other column group. For the second row group we have wavelengths from 1282 to 2732 in one column group and all the following in the other column group.

In order to illustrate the classification version of the algorithm, we also provide results about the supervised partitioning. The observed data are randomly split into training and a test sets and the algorithm is run on the training set. After parameter estimation, the performance of the algorithm is measured on the remaining test set. We used a test set of 25 objects, leaving the remaining for training of the model.

The procedure is repeated 1000 times, and Table 20 gives the estimated probabilities of classification. These results are competitive and comparable to the best results obtained in Farcomeni *et al.* (2007) with formal signal processing and variable selection, and use of k-nearest neighbors classifier. Finally, while the use of the supervised algorithm leads to identification of similar ranges of wavelenghts as most discriminant, now the column groups are not mostly made of neighboring wavelenghts.

	Glass	**Glass Ceramic**
Predicted Glass	0.739	0.005
Predicted Glass Ceramic	0.076	0.180

Table 20: Estimated classification error with supervised algorithm

9. DISCUSSION

In this paper a model based multi-partitioning methodology has been proposed. It allows to partition units of a multivariate data set and simultaneously to partition variables for each class of the partition of units. This model has been also studied by Rocci & Vichi, 2006 using a semi-parametric approach. Here the multi-partitioning model is specified in a model based framework. The parametric assumption is that the population from which the data are observed structures into P homogeneous subpopulations in proportions $\pi_1, \pi_2, ..., \pi_P$, each having multivariate normal distribution. If subpopulations are not expected to be well distinct, a fuzzy (overlapping) classification may be more useful to classify units (and a multi-covering problem methodology would be defined). It is straightforward to use our algorithms for this case too.

The unknown membership of units to the clusters can be specified to have a fixed or a random effect, which correspond to consider the membership as a fixed or a random variable. Maximum likelihood estimation has been used in both cases and the corresponding coordinate ascent algorithms of the EM type are given.

The fixed effect model can be used when subpopulations are well represented in the sample and this is generally achieved for large samples. In this case the expected proportions can be considered fixed from sample to sample and therefore the membership are fixed

variables. For this case a quite fast algorithm is given which recovers with high probability the cluster membership of the units and variables in the generated data according to the simulations study given in section 6.

We have observed that for the fixed effect case the parameters of the multivariate normal distributions may not be consistently estimated, because their direct estimation truncates the tails of the densities involved. In the case of well separated subpopulations this becomes an irrelevant problem. However, when multinormals heavily overlap a modified EM algorithm has been introduced to estimate consistently their parameters. It uses a complete step of a usual EM algorithm. The modified EM still increases the complete data likelihood of the fixed effect model.

It is interesting to note that assignment of the units to the clusters specifies always an optimal partition of the units, as in the case of the CEM algorithm (Classification EM Celeux and Goveart, 1992); however, the new algorithm for the fixed effect model differs from CEM for two points: (i) it maximizes a criterion (13) of maximum likelihood clustering type (Scott and Symons, 1971); (ii) it estimates consistently the parameters of the multinormal distributions.

If a fuzzy partition of the units is required (the data show overlapping clusters) and/or a random (multinomial) effect is expected, i.e., a random effect from sample to sample is predictable, it is convenient to use the random effect model. In this case an EM algorithm is given; however, its convergence is more time-consuming, with respect to the algorithm for the fixed effect model.

We suggested to parameterize the covariance matrices as indicated by Banfield & Raftery, (1993); Celeux and Govaert, (1995) in order to reduce the number of parameters to estimate and to choose the model according to a classification criterion such as AIC or BIC. In our simulation study emerged that BIC may outperform AIC in detecting the correct number of row and column clusters; however, this is not so often the case and therefore we suggest to examine different criteria and to choose the more parsimonious and interpretable solution.

REFERENCES

Alon,U. ,Barkai, N. ,Notterman, D.A. , Gish, K. , Ybarra, S., Mack, D. , Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. {USA}, **96**: 6745-6750*

Banfield, J & Raftery, A. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803-821.

Bensmail, H. & Celeux, G. (1996) Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**, 1743-1748.

Celeux G. & Govaert G. (1992) A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14, 3, 315-332.
Celeux C. & Govaert G. (1995) Gaussian Parsimonious Clustering Models, *Pattern Recognition*, vol. 28, pp. 781-793.

Farcomeni, A., Serranti, S. & Bonifazi, G. (2007) Non-parametric analysis of infrared spectra for recognition of glass and glass ceramic fragments in recycling plants. *Waste Management*, ePub ahead of print, available online.

Flury, B. (1988) *Common Principal Components and Related Multivariate Models.* New York; Wiley.

Golub, T.R. , Slonim, D.K., Tamayo, P. , Huard, C. , Gaasenbeek, M. , Mesirov, J.P. ,Coller, H. , Loh, M.L. , Downing, J.R. , Caligiuri, M.A., Bloomfied, C. D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring *Science,* **286**: 531-537

Rocci & Vichi (2006) Two mode Multi-Partitioning, accepted by *Computational Statistics and Data Analysis*.

Scott A.J. & Symons M.J. (1971), Clustering Methods based on Likelihood Ratio Criteria, *Biometrics*, 27, 387-397.