Journal de la Société Française de Statistique
Vol. 153 No. 2 (2012)

SelvarClustMV: Variable selection approach in model-based clustering allowing for missing values

Titre : SelvarClustMV : sélection de variables pour la classification non supervisée avec données manquantes

Cathy Maugis-Rabusseau[1], Marie-Laure Martin-Magniette[2,3,4,5,6] and Sandra Pelletier[7]

Abstract: Overabundance of clustering methods exists but none was devised with a variable selection procedure and a missing data management. However in microarray datasets, genes are described by a growing number of experiments and missing data always exist. It is also important to detect the relevant experiments for improving the gene clustering and the data interpretation. A common practice is to remove genes with missing values or to replace missing values with estimation. However it is known to have an important impact on the clustering result. We tackle variable selection and missing data in a unique statistical framework: A versatile variable selection model based on multidimensional Gaussian mixtures is proposed, taking variable roles for clustering into account. Moreover this statistical framework manages missing values without imposing any data pre-processing. Numerical experiments highlight the gain of our method compared to imputation methods which do not allow to find the true variable roles and sometimes lose biological information.

Résumé : De nombreuses méthodes de classification non supervisée existent mais sont souvent conçues sans procédure de sélection de variables et ne permettent pas toujours de gérer les données manquantes. Dans les données issues de puces à ADN, les gènes sont décrits par un grand nombre d'expériences où il existe toujours des données manquantes. Il est donc important de détecter les expériences biologiques significatives afin d'améliorer la classification des gènes et son interprétation. Concernant les valeurs manquantes, il est courant d'écarter de l'étude les gènes non totalement observés ou d'estimer les valeurs manquantes avant classification. Dans cet article, nous traitons la sélection de variables et le problème des données manquantes grâce à une unique procédure. Nous proposons un modèle de sélection de variables pour prendre en compte le rôle des variables pour la classification non supervisée par mélanges gaussiens, où les données manquantes ne sont pas prétraitées. Des expériences numériques illustrent le gain de notre méthode par rapport aux méthodes avec imputation des données manquantes qui ne permettent pas toujours de retrouver le vrai rôle des variables et parfois perdent des informations biologiques.

Keywords: Variable selection, Missing values, Model-based clustering
Mots-clés : Sélection de variables, Données manquantes, Classification par mélanges gaussiens
AMS 2000 subject classifications: 62H30, 91C20

[1] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4 France. E-mail: cathy.maugis@insa-toulouse.fr
[2] INRA, UMR 518 Mathématiques et Informatiques Appliquées, F-75231 Paris, France. E-mail: marie_laure.martin@agroparistech.fr
[3] AgroParisTech, UMR Mathématiques et Informatiques Appliquées, F-75231 Paris, France
[4] INRA, UMR1165 Unité de Recherche en Génomique Végétale URGV, F-91057 Evry, France
[5] UEVE, UMR Unité de Recherche en Génomique Végétale URGV, F-91057 Evry, France
[6] CNRS, ERL8196 UMR Unité de Recherche en Génomique Végétale URGV, F-91057 Evry, France
[7] INRA, Institut de Recherche en Horticulture et Semences, UMR1345, PRES UNAM, 16 Bd Lavoisier, 49045 Angers Cedex 01, France. E-mail: sandra.pelleter@angers.inra.fr

 ISSN: 2102-6238

1. Introduction

Overabundance of clustering methods exist but none was devised with a variable selection procedure and a missing data management. However with microarray data, we are always confronted to a large number of available experiments describing the genes and missing data due to technical problems occurring during the production process. Intuitively, the more information we have about each gene, the better a clustering method is expected to perform. However when the number of experiments is large, the structure of interest may often be contained in a subset and some experiments could be useless or harmful. Thus it is important to detect the relevant experiments for the gene clustering. It leads to an improvement of gene clustering and its interpretation. This variable selection problem for clustering is a recent topic coming from the increasingly frequent study of high-dimensional datasets, such as gene expression datasets. Indeed expression data provide the main source of information about genes and are used to improve functional annotation by determining coexpressed gene clusters, usually assumed to be good candidats of coregulated genes [5].

Recently, several authors have recast the variable selection for clustering in the setting of Gaussian mixtures. After the work of Law et al. [11] and Raftery and Dean [18], successive improvements of the variable role modeling have been presented in [13, 14]: The so-called SR modeling and its generalization called SRUW modeling, which completely takes the variable role into account by subsetting the relevant variables for the clustering, the redundant variables and the independent variables. Theoretical properties of these methods are established: The model collection is identifiable and despite the model complexity, the variable selection is consistent. Moreover the interest of these two variable selection methods for studying gene expression data has been highlighted in [13, 15].

Nevertheless these procedures remain unusable for transcriptome dataset analysis since they do not take missing data into account. For this reason the transcriptome datasets studied in [13, 15] were preliminary restricted to the subset of totally observed genes, removing thus potential interesting genes. It is already known that missing values is a major factor of gene cluster instability [4], so missing value management is all the more crucial when gene clustering procedure includes a variable selection step. The objective of this paper consists of extending procedures proposed by [13, 14], briefly presented in Section 2. This extension, named *SelvarClustMV* [1], performs the variable selection and the clustering of data with missing values, without imposing any imputation of these latter. The required assumptions to extend the two models proposed by [13, 14] are discussed in Section 3. Imputation methods of missing values are presented in Section 4. *SelvarClustMV* and the same variable selection model-based clustering combined with an imputation method are compared in Section 5. *SelvarClustMV* seems to be more reliable than imputation methods: It is able to find the true model up to 20% of missing values, the error rates are among the smallest and it seems to keep easier biological information in the clusters.

[1] *SelvarClustMV* is available at `http://www.math.univ-toulouse.fr/~maugis/SelvarClustMVHomepage.html`

 ISSN: 2102-6238

2. Background

A sample of n objects $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ described by Q quantitative variables is considered. This sample is decomposed into $\mathbf{y} = (\mathbf{y}^{\mathbf{o}}, \mathbf{y}^{\mathbf{m}})$ where $\mathbf{y}^{\mathbf{o}}$ is the observed value subset and $\mathbf{y}^{\mathbf{m}}$ are the missing entries. In the model-based clustering context, the data $\mathbf{y}$ are assumed to come from several subpopulations (clusters) modeled with a multivariate Gaussian density. The observations are assumed to arise from a finite Gaussian mixture with K components

$$f(\mathbf{y}_i|K,m,\alpha) = \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i|\mu_k,\Sigma_k)$$

where $\mathbf{p} = (p_1, \ldots, p_K)$ is the mixing proportion vector ($p_k \in (0,1)$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$) and the function $\Phi(.|\mu_k,\Sigma_k)$ denotes the Q-dimensional Gaussian density with mean vector μ_k and variance matrix Σ_k. The mixture form, denoted by m, is related to the form of the possible variance matrices derived from assumptions on their eigenvalue decomposition which allow to control the volume, the orientation and the shape of each cluster. A collection of 28 Gaussian mixture models is available in the MIXMOD software, which allows one to estimate Gaussian mixture parameters (see for instance [2] for details.). Next, the *Maximum A Posteriori* (MAP) rule is considering to cluster objects. This clustering rule consists of assigning each object to the cluster with the highest conditional probability.

The variable selection problem for the model-based clustering with Gaussian mixtures is recast into a model selection problem. In the SR modeling [13], the model family consists of $\mathcal{N} = \{(K,m,S,R); (K,m) \in \mathcal{T}, (S,R) \in \mathcal{V}\}$. The set of variable partitions $\mathcal{V}$ contains couples (S,R) where S is the nonempty set of relevant clustering variables and R is a subset of S containing the relevant variables required to explain irrelevant variables according to a linear regression. For the model (K,m,S,R), the data distribution is modeled by

$$f(\mathbf{y}|K,m,S,R,\theta) = f_{\text{clust}}(\mathbf{y}^S|K,m,\alpha) f_{\text{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta,\Omega).$$

where θ denotes the parameter vector (α,a,β,Ω), the function

$$f_{\text{clust}}(\mathbf{y}^S|K,m,\alpha) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^S|\mu_k,\Sigma_k) \right\}$$

corresponds to the Gaussian mixture density on variables S with the parameter vector α. The function $f_{\text{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta,\Omega)$ corresponds to the multidimensional multivariate linear regression density of $\mathbf{y}^{S^c}$ on $\mathbf{y}^R$ with the intercept vector a, the regression coefficient matrix β and the variance matrix Ω.

In the SRUW modeling [14], a generalization of the variable roles is proposed. The irrelevant clustering variables are divided into two variable subsets U and W. The variables belonging to U are explained by a variable subset R of S according to a linear regression while the variables in W are assumed to be independent of all the relevant variables. The marginal distribution of the data on W is assumed to be a Gaussian distribution.

For these two models, a BIC-type criterion is proposed to solve the model selection problem. It corresponds to the maximized loglikelihood minus a penalty term defined by the number of free

 ISSN: 2102-6238

parameters times the logarithm of the number of objects. In practice, since the number of models is huge in the two modelings, an exhaustive search is impossible. The so-called *SelvarClust* [2] and *SelvarClustIndep* [3] algorithms, embedding two backward stepwise algorithms for variable selection for clustering and linear regression are proposed for the SR and the SRUW modelings respectively. At each step of these two algorithms, the parameter estimation and the difference of criterion values are required and calculated using MIXMOD (`http://www.mixmod.org/`). After selecting the best model and estimating the associated parameter vector, the MAP rule is considered to cluster objects.

3. Extension of the variable selection procedures for missing at random data

In this section, we are interested in the extension of our variable selection procedures for the study of datasets with missing values. The aim is to take the existence of these missing values into account by avoiding a preliminary estimation of the missing data. Thus we have to specify under which assumptions such an extension is possible. An adaptation of the model selection criterion, the parameter estimation and our variable selection algorithms have to be provided. In the sequel, we focus on the SR modeling to explain the procedure adaptation, but a similar extension for the SRUW modeling can be obtained and is explained in Section 3.3.

3.1. Nature of missing data

It is possible to distinguish three types of missing data according to the missing-data mechanism [19]: *missing completely at random*, *missing at random* and, *not missing at random*. In this paper, the missing data are assumed to be missing at random (MAR), namely the probability that a value is missing is related to the observed data but not to the missing data. In the MAR assumption, the missing-data mechanism is called ignorable [12, 20] and likelihood-based inference can be obtained by ignoring the missing-data mechanism. Considering the missing-data indicator matrix M defined by

$$M_{ij} = \begin{cases} 1 & \text{if} \quad y_i^j \text{ is observed} \\ 0 & \text{if} \quad y_i^j \text{ is missing} \end{cases} \tag{1}$$

as a random variable, the MAR assumption can be reformulated as "the distribution of the missing-data mechanism M is independent of missing values $\mathbf{y^m}$" [19].

3.2. Model selection criterion

The model selection in the SR procedure is based on the maximization of $f(\mathbf{y^o}, M|K, m, S, R)$, the integrated observed likelihood. Using the MAR assumption, it is equivalent to select the model maximizing the integrated observed likelihood ignoring the missing-data mechanism

[2] *SelvarClust* is available at `http://www.math.univ-toulouse.fr/~maugis/SelvarClustHomepage.html`

[3] *SelvarClustIndep* is available at `http://www.math.univ-toulouse.fr/~maugis/SelvarClustIndepHomepage.html`

 ISSN: 2102-6238

$f(\mathbf{y^o}|K,m,S,R)$. But this last quantity is difficult to evaluate and a BIC approximation is used. The chosen model maximizes the criterion

$$2\ln\left\{f(\mathbf{y^o}|K,m,S,R,\hat{\theta}\,)\right\} - \Xi_{(K,m,S,R)}\ln(n), \tag{2}$$

where $\hat{\theta}$ is the parameter vector maximizing the observed likelihood $f(\mathbf{y^o}|K,m,S,R,\theta)$ and $\Xi_{(K,m,S,R)}$ is the total number of free parameters for model (K,m,S,R).

In order to be able to use this model selection criterion, the observed likelihood has to be made explicit and the parameter vector $\hat{\theta}$ maximizing the observed likelihood $f(\mathbf{y^o}|K,m,S,R,\theta)$ has to be evaluated. These two points are addressed in Sections 3.2.1 and 3.2.2, respectively.

3.2.1. Explicit observed likelihood expression

The observed likelihood ignoring the missing-data mechanism is defined by

$$f(\mathbf{y^o}|K,m,S,R,\theta) = \int f(\mathbf{y^o},\mathbf{y^m}|K,m,S,R,\theta)d\mathbf{y^m}. \tag{3}$$

In order to evaluate explicitly this quantity, the distribution of the sample is recast into a global Gaussian mixture: The likelihood can be written as

$$f(\mathbf{y^o},\mathbf{y^m}|K,m,S,R,\theta) = \prod_{i=1}^{n}\left\{\sum_{k=1}^{K} p_k\Phi(\mathbf{y}_i^{\mathbf{o}},\mathbf{y}_i^{\mathbf{m}}|\nu_k,\Delta_k)\right\}$$

where for all variables $j \in \{1,\ldots,Q\}$,

$$\nu_{kj} = \begin{cases} \mu_{kj} & \text{if } j \in S \\ (a+\mu_k\tilde{\beta})_j & \text{if } j \in S^c \end{cases} \tag{4}$$

$\forall j \in S,\ \forall l \in S^c$,

$$\tilde{\beta}_{jl} = \begin{cases} \beta_{jl} & \text{if } j \in R \\ 0 & \text{if } j \in S\backslash R \end{cases} \tag{5}$$

and, for all variables l and j,

$$\Delta_{k,jl} = \begin{cases} \Sigma_{k,jl} & \text{if } j \in S, l \in S \\ (\Sigma_k\tilde{\beta})_{jl} & \text{if } j \in S, l \in S^c \\ (\tilde{\beta}'\Sigma_k)_{jl} & \text{if } j \in S^c, l \in S \\ (\Omega+\tilde{\beta}'\Sigma_k\tilde{\beta})_{jl} & \text{if } j \in S^c, l \in S^c. \end{cases} \tag{6}$$

In order to set apart the conditional distribution of the missing values according to the observed values, the mean vectors and the variance matrices are decomposed into

$$\nu_k = (\nu_{k,\mathbf{o}}^{(i)}, \nu_{k,\mathbf{m}}^{(i)}) \text{ and } \Delta_k = \begin{pmatrix} \Delta_{k,\mathbf{oo}}^{(i)} & \Delta_{k,\mathbf{om}}^{(i)} \\ \Delta_{k,\mathbf{mo}}^{(i)} & \Delta_{k,\mathbf{mm}}^{(i)} \end{pmatrix}$$

 ISSN: 2102-6238

according to the position of missing values for $\mathbf{y}_i$. According to Theorem 2.5.1 in [1], the Gaussian mixture density can then be decomposed as follows

$$\sum_{k=1}^{K} p_k \Phi\left(\mathbf{y}_i^{\mathbf{o}} | \nu_{k,\mathbf{o}}^{(i)}, \Delta_{k,\mathbf{oo}}^{(i)}\right) \Phi\left(\mathbf{y}_i^{\mathbf{m}} | \nu_{k,\mathbf{m}|\mathbf{o}}^{(i)} + \mathbf{y}_i^{\mathbf{o}} \Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)}, \Delta_{k,\mathbf{mm}|\mathbf{o}}^{(i)}\right),$$

where

$$\begin{cases} \nu_{k,\mathbf{m}|\mathbf{o}}^{(i)} = \nu_{k,\mathbf{m}}^{(i)} - \nu_{k,\mathbf{o}}^{(i)} \Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)}, \\ \Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)} = (\Delta_{k,\mathbf{oo}}^{(i)})^{-1} \Delta_{k,\mathbf{om}}^{(i)}, \\ \Delta_{k,\mathbf{mm}|\mathbf{o}}^{(i)} = \Delta_{k,\mathbf{mm}}^{(i)} - \Delta_{k,\mathbf{mo}}^{(i)} (\Delta_{k,\mathbf{oo}}^{(i)})^{-1} \Delta_{k,\mathbf{om}}^{(i)}. \end{cases}$$

We deduce thus the expression of $f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta)$ given by:

$$f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \Phi\left(\mathbf{y}_i^{\mathbf{o}} | \nu_{k,\mathbf{o}}^{(i)}, \Delta_{k,\mathbf{oo}}^{(i)}\right). \tag{7}$$

Consequently, it is possible to calculate explicitly the observed likelihood using the global Gaussian mixture with parameters $(p_k, \nu_k, \Delta_k)_{1 \leq k \leq K}$ and using the expression (7). The second task is then to derive the parameter vector $\hat{\theta}$ maximizing it.

3.2.2. Maximum observed likelihood estimator

Since the sample density can be formulated as a global Gaussian mixture, an EM algorithm could be used in order to estimate the parameters $(\hat{p}_k, \hat{\nu}_k, \hat{\Delta}_k)_{1 \leq k \leq K}$ and to deduce then $\hat{\theta} = (\hat{\alpha}, \hat{a}, \hat{\beta}, \hat{\Omega})$. But the constraints (4), (5) and (6) do not lead to a close form update of $\hat{\theta}$ in the M step. We propose thus to estimate the parameter vector α of the Gaussian mixture on S and the parameter vector (a, β, Ω) of the regression of S^c on R separately.

For estimating the Gaussian mixture parameter vector $\hat{\alpha}$ maximizing $f_{\text{clust}}(\mathbf{y}^{S,\mathbf{o}}|K,m,\alpha)$, an EM algorithm is used on the objects having at least one observed value on S. In this EM algorithm, the latent variables are composed of the unknown label vector z and the missing values $\mathbf{y}^{S,\mathbf{m}}$ (see for instance [12] for more details). The tricky point concerns the estimation of the maximum likelihood parameters for a multidimensional multivariate linear regression:

$$\forall i \in \{1,\ldots,n\},\ \mathbf{y}_i^{S^c} = a + \mathbf{y}_i^R \beta + \varepsilon_i$$

where the ε_i's are i.i.d $\mathcal{N}(0,\Omega)$, because both matrices $\mathbf{y}^{S^c}$ and $\mathbf{y}^R$ may contain missing values. A possible estimation method would consist of assuming that the vectors $(\mathbf{y}_i^{S^c}, \mathbf{y}_i^R)$ have a normal distribution (see Section 8.4 in [12]). The parameter vector for the global Gaussian density maximizing the observed likelihood would be then estimated using a classical EM algorithm and the parameters $(\hat{a}, \hat{\beta}, \hat{\Omega})$ would be deduced. This strategy would be easily carried out but the assumption is unrealistic because vectors $(\mathbf{y}_i^{S^c}, \mathbf{y}_i^R)$ cannot fulfill this normal law assumption in our context since the vectors $\mathbf{y}_i^R$ are modeled by a Gaussian mixture. Thereby, we suggest an other strategy requiring the predictor matrix $\mathbf{y}^R$ to be totally observed: The matrices $\mathbf{y}^{S^c}$ and $\mathbf{y}^R$ are restricted to the objects totally observed on R and then an EM algorithm is used to estimate the regression parameters $(\hat{a}, \hat{\beta}, \hat{\Omega})$.

 ISSN: 2102-6238

3.3. *Adaptation of* SelvarClust *and* SelvarClustIndep

An analogous extension is available for the SRUW modeling where the parameters of the additional Gaussian density on the independent variable subset are estimated by an usual EM algorithm and this Gaussian density can be included into a global Gaussian mixture for the observed likelihood calculation. Both softwares *SelvarClust* and *SelvarClustIndep* can be thus extended to take missing values into account. Nevertheless the software MIXMOD does not consider missing values so it is necessary to implement the EM algorithm for each mixture form. For this study, we focus on Gaussian mixture having the same variance matrix for each component ($\Sigma_k = \Sigma$, $\forall 1 \leq k \leq K$). Indeed when *SelvarClust* is applied on transcriptome datasets, this mixture form, denoted $[p_k LC]$, is most often selected with the form $[p_k L_k C]$ (variance matrices fulfill $\Sigma_k = \lambda_k \Sigma$, $\forall 1 \leq k \leq K$, λ_k a scalar).

4. Imputation methods

In the gene expression framework, imputation methods are usually used to complete the gene expression data. During the last decade, several imputation methods have been proposed (see for instance [21, 26] for an overview) and evaluated [3, 4, 25, 22]. According to their conclusions, we evaluate six imputations methods: The first three methods, called ZERO-imputation, ROW-imputation and COL-imputation methods, consist of replacing missing values with zero, the row (gene) average or the column (array) average respectively. They do not take the data correlation structure into account. Troyanskaya et al. [23] propose two correlation-based imputation methods including the KNN-imputation method. It consists of finding the k genes which are the closest to the gene of interest with missing values according to a distance metric, most frequently the Euclidian distance or Pearson correlation. The missing value is then estimated by the weighted average of these k-nearest neighbour genes for the same array where the weights are calculated from their similarity measurement. The R package *impute.knn* proposed by Hastie et al. [7] with the choice per default $k = 10$ number of neighbours has been used. Oba et al. [16] show that their BPCA-imputation method outperforms the previous method. The BPCA-imputation uses Bayesian estimation to fit a probabilistic PCA model, which is based on the assumption that the factor scores and the residuals obey normal distributions. The Matlab software *BPCAfill.m* proposed by Oba et al. [16] has been used. The last tested imputation method is the Local Least Squares (LLS-) imputation method of Kim et al. [10]. It estimates all missing values of an object simultaneously. First, it selects the k-nearest neighbours of an object with missing values based on the Pearson correlation. Second, this method performs a multiple regression using all k neighbours. Then the missing values are imputed, based on the least square estimates. Kim et al.[10] propose an heuristic method to select the number of neighbours k. The Matlab software *impute_llsq_l2.m* has been used in this paper.

For comparing the six imputation methods, the accuracy measure, usually based on the root mean squared error (RMSE) between the original matrix $\mathbf{y}$ and the imputed matrix $\hat{\mathbf{y}}$, is calculated. As in [17, 8, 24, 9], the RMSE is normalized by the root mean squared true values of the missing

 ISSN: 2102-6238

entries:

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{Q}\left(y_i^j - \hat{y}_i^j\right)^2 \mathbb{1}_{M_{ij}=0}}{\sum_{i=1}^{n}\sum_{j=1}^{Q}(y_i^j)^2 \mathbb{1}_{M_{ij}=0}}}$$

allowing us to consider the ZERO-imputation method as reference (NRMSE is equal to 1 for the ZERO-imputation). When the estimated values are accurate, the NRMSE reaches its minimum value 0 and when the missing value estimation is poor, the NRMSE becomes large.

5. *SelvarClustMV* behavior

The objective in this section is to discuss the best way to manage missing values in a model-based clustering with a variable role modeling. We compare *SelvarClustMV* which manages the missing values and *SelvarClust* combined with a preprocessing imputation method (among the six studied imputation methods).

5.1. Simulated dataset

A simulated dataset consisting of 2000 data points from a mixture of four Gaussian distributions $\mathcal{N}(\mu_k, \Sigma)$ is considered. The mean vectors are $\mu_1 = (0,0,0)$, $\mu_2 = (-6,6,0)$, $\mu_3 = (0,0,6)$, $\mu_4 = (-6,6,6)$ and the variance matrix is $\Sigma = A' \times \text{diag}(6\sqrt{2},1,2) \times A$ where the matrix A is the product of two 3×3 rotation matrices around the Oz and Ox axis with angle $\pi/6$ and $\pi/3$ respectively. The proportion vector of this mixture is $\mathbf{p} = (0.25, 0.25, 0.2, 0.3)$. The fourth and fifth variables are defined for all $i \in \{1,\ldots,n\}$ by

$$(y_i^4, y_i^5) = (-1,2) + (y_i^1, y_i^2)((0.5,2)', (1,0)') + \varepsilon_i,$$

ε_i being sampled from a $\mathcal{N}(0, \text{rot}(\pi/6)'\text{diag}(1,3)\text{rot}(\pi/6))$ density where $\text{rot}(\pi/6)$ is the 2×2 plane rotation matrix with angle $\pi/6$. Two noisy independent standard centered Gaussian variables are also appended. The true model under the SR modeling is thus

$$(K_0 = 4, m_0 = [p_k LC], S_0 = \{1,2,3\}, R_0 = \{1,2\}).$$

We applied *SelvarClustMV* and the six imputation methods combined with *SelvarClust*. The mixture form is fixed to $[p_k LC]$ and the component number varies from 1 to 8. Table 1 summarizes the results according to the percentage of missing values.

First remark concerns the NRMSE. As expected, the first three methods not taking the data correlation structure into account are inaccurate and the three others perform better. As [22], we point out that the performance of imputation methods improves initially as the missing value rate increase until an optimum point and then the performance are worse with increasing missing value rate. The BPCA and LLS methods are the more accurate, the KNN-imputation gives also reasonable NRMSE results. The second remark concerns the variable partition and the component number: In all scenarios, *SelvarClustMV* selects the true variable partition and the true number of clusters. In contrast, the variable selection procedure with a ZERO-, ROW- or COL-imputation in

 ISSN: 2102-6238

TABLE 1. *Results given by SelvarClustMV(indicated by "-" in the second column) or SelvarClust with a preprocessing imputation method for several percentages of missing values. The clustering error rate is calculated on objects with at least one observed value among the declared relevant variables and on objects totally observed on the first three variables (brackets). When the true model is selected, the solution is in bold.*

percentage of missing values	imputation method	NRMSE	$\hat{K}$	$\hat{S}$	$\hat{R}$	clustering error rate	
$c = 0\%$	-	-	**4**	**1,2,3**	**1,2**	1.15%	
$c = 1\%$	-	-	**4**	**1,2,3**	**1,2**	1.55%	[1.13%]
	ZERO	1	6	1,2,3,4,5,6,7	$\emptyset$	2.00%	
	ROW	1.1008	5	1,2,3,4,5	$\emptyset$	2.10%	
	COL	0.8013	6	1,2,3,4,5	$\emptyset$	1.70%	
	KNN	0.3617	**4**	**1,2,3**	**1,2**	**1.45%**	[1.13%]
	LLS	0.3337	**4**	**1,2,3**	**1,2**	**1.45%**	[1.13%]
	BPCA	0.3344	**4**	**1,2,3**	**1,2**	1.55%	[1.1%]
$c = 5\%$	-	-	**4**	**1,2,3**	**1,2**	3.85%	[1.10%]
	ZERO	1	6	1,2,3,4,5,6,7	$\emptyset$	6.55%	
	ROW	1.835	6	1,2,3,4,5,6,7	$\emptyset$	44.45%	
	COL	0.8255	6	1,2,3,4,5	$\emptyset$	6.35%	
	KNN	0.4883	6	1,2,3,4,5	$\emptyset$	6.65%	
	LLS	0.4175	**4**	**1,2,3**	**1,2**	**3.35%**	[1.04%]
	BPCA	0.3784	**4**	**1,2,3**	**1,2**	3.60%	[1.10%]
$c = 10\%$	-	-	**4**	**1,2,3**	**1,2**	7.17%	[1.17%]
	ZERO	1	6	1,2,3,4,5,6,7	$\emptyset$	44.90%	
	ROW	1.0737	6	1,2,4,5	4,5	45.25%	
	COL	0.8035	6	1,2,3,4,5,7	$\emptyset$	12.45%	
	KNN	0.5137	5	1,2,3,4,5,6	$\emptyset$	3.35%	
	LLS	0.4143	**4**	**1,2,3**	**1,2**	**6.45%**	[1.10%]
	BPCA	0.4020	**4**	**1,2,3**	**1,2**	6.85%	[1.17%]
$c = 15\%$	-	-	**4**	**1,2,3**	**1,2**	10.17%	[0.56%]
	ZERO	1	6	1,2,4,6	1,2,4	45.80%	
	ROW	1.0671	5	1,2,3,4,5	1,3,4	45.65%	
	COL	0.8284	6	1,2,3,4,5,6	$\emptyset$	11.40%	
	KNN	0.5180	6	1,2,3,4,5	$\emptyset$	9.35%	
	LLS	0.4467	**4**	**1,2,3**	**1,2**	**8.90%**	[0.96%]
	BPCA	0.4082	**4**	**1,2,3**	**1,2**	9.15%	[0.96%]
$c = 20\%$	-	-	**4**	**1,2,3**	**1,2**	13.27%	[1.17%]
	ZERO	1	6	1,2,4,5	1,4	45.60%	
	ROW	1.0925	6	1,2,3,4,5,6,7	$\emptyset$	45.85%	
	COL	0.8291	6	1,2,3,4,5,6,7	$\emptyset$	45.05%	
	KNN	0.5833	6	1,2,3,4,5	$\emptyset$	12.40%	
	LLS	0.5180	5	1,2,3,4,5	4,5	12.45%	
	BPCA	0.4450	**4**	**1,2,3**	**1,2**	**11.90%**	[1.37%]

 ISSN: 2102-6238

preprocessing does not select the true model in all the cases. With the KNN-imputation method, the procedure finds the true model only for $c = 1\%$ of missing values. With LLS- or BPCA-imputation method, the variable selection procedure has a better behavior since it selects the true model, except for $c = 20\%$ when the LLS-imputation is applied. It shows that it is better to use the extensions of our variable selection procedures because in all cases, the variable partition is well estimated.

Finally we estimate the clustering error rate. To do this, objects are assigned to a cluster according to the *Maximum A Posteriori* rule which assigns an object to the cluster with the highest conditional probability. Since the conditional probabilities of *SelvarClustMV* are based on the selected relevant variables, objects which have not at least one value for the relevant variables are not clustered. In our simulated dataset respectively, 5, 3 and 10 objects are not clustered for c equals 10, 15 and 20 %. For *SelvarClust* combined with an imputation method, the clustering error rate is always calculated from the 2000 objects. The difference of sample size being small, we think that the error rates can be compared. For all methods except KNN-imputation method, when the missing value percentage increases, the clustering error rate also increases. *SelvarClust* combined with ZERO-, ROW- or COL-imputation gives large error rates, it confirms that such naive imputation methods are too simple and should not be used. The other methods combined with *SelvarClust* have similar error rate although KNN-imputation seems to be less stable than LLS- ou BPCA-imputations. The error rate of *SelvarClustMV* is close to the error rate of the best method for each missing value percentage. In fact, this error rate puts *SelvarClustMV* at a disadvantage by comparison with the others because the conditional probabilities are evaluated on the available values of the relevant variables. To be specific, Table 2 shows the 264 misclassified objects for the scenario $c = 20\%$ according to their initial group and the position of their missing values on the three relevant clustering variables. Only four objects without missing values on the three relevant variables are misclassified. Moreover, the objects having at least one missing value for the third variable most often belong to Groups 2 and 3. According to the simulation model, the third variable allows to distinguish Group 1 from Group 3, and Group 2 from Group 4. For this reason the 52 objects of Group 2 which are not observed on the third variable are consequently all clustered in Cluster 4. It explains the increase of the clustering error rate with the missing value percentage and why our extended variable selection procedure has not the smallest error rate.

To pursue the comparison a second error rate based on the objects completely observed on the first three variables is calculated. The object number for each missing value percentage is respectively 1949, 1722, 1454, 1249 and 1024. When the missing value percentage increases, the clustering error rate for the objects completely observed on the first three variables is about 1% except for *SelvarClustMV* when $c = 15\%$. These percentages correspond to about ten misclassified objects. Note that in all cases, the error rate of *SelvarClustMV* is identical to *SelvarClust* combined with an imputation method or better when the missing value percentage is large (c greater than 15%). In conclusion, *SelvarClustMV* gives the best results since the true model is always found whatever the missing value percentage. Moreover its error rates are among the smallest. The difficult comes from objects with missing values on the relevant variables. For these objects, caution should be taken.

 ISSN: 2102-6238

TABLE 2. *Distribution of the 264 misclassified objects when $c = 20\%$ according to their initial group and the position of their missing values on the three relevant clustering variables.*

Positions of the missing values	Group 1	Group 2	Group 3	Group 4
Variable 1	9	7	5	10
Variable 2	0	1	8	0
Variable 3	1	52	53	10
Variables 1 and 2	0	13	14	0
Variables 1 and 3	1	15	12	3
Variables 2 and 3	0	16	21	1
No missing values	4	3	2	3
Total	15	107	115	27

5.2. Transcriptome dataset

The dataset already analyzed in [13] is considered. It concerns $n = 1267$ genes declared differentially expressed at least once in a time course of the hypocotyl growth switch and their expression profiles are studied on $T = 7$ projects with $Q = 27$ experiments. The dataset was extracted from the database CATdb [6]. For their study, [13] restricted the dataset to a subset of 1020 totally observed genes. The 247 removed genes are 118 genes manually flagged by the experimenter in some experiments for technical reasons and 129 genes removed because they did not satisfy the homoscedastic assumption in the differentially analysis (see [6] for a detailed description of the differential analysis). The first missing values are missing values at random but not the second ones. It is thus important to distinguish them. For the 129 genes, we propose to recalculate a test statistic as the expression difference normalized by the estimated standard deviation obtained with all the genes satisfying the homoscedastic assumption. Hence the dataset is composed of 1149 totally observed genes and 118 genes with missing values: 107 genes with one missing value, 10 genes with two missing values and 1 gene with three missing values. Consequently, 9.3% of genes have at least one missing value and the missing value percentage equals $c = 0.38\%$.

We apply *SelvarClustMV* and also *SelvarClust* combined with LLS- or BPCA-imputation method. In all cases, the mixture form is fixed to $m = [p_k LC]$, because this form was selected in the analysis of [13]. The number of mixture components varies between 2 and 20. *SelvarClustMV* selects a clustering with 17 clusters and Projects 1, 2, 3, 4, 6 and 7 are relevant, the last four ones being required to explain Project 5. The result differs from the one of [13] which found also 17 clusters but declared Projects 1, 3, 4, 6 and 7 relevant with the last four ones required in the regression model to explain Projects 2 and 5. The comparison of both results is not direct since the additional genes could reveal new structures. Nevertheless we expect that the results are coherent. Contingency table between the two clusterings on the 1020 genes studied in [13] is given Table 3. Fourteen clusters are close to the ones for the dataset restricted to the 1020 totally observed genes. Clusters 5 and 9 appear to be new clusters. They contain respectively 12 and 2 new genes and their expression profiles have characteristic expressions in Project 2 contrary to the genes of the other new clusters (see Figure 1). This can explain that Project 2 is now relevant. For Project 5, the same projects are selected to explain it. The explanations based on the redifferentiation of cells and the formation of giant cells given in [13] are still valid and the estimations of the regression parameters are similar.

SelvarClust combined with an imputation method is also tested. To begin with, we generate

 ISSN: 2102-6238

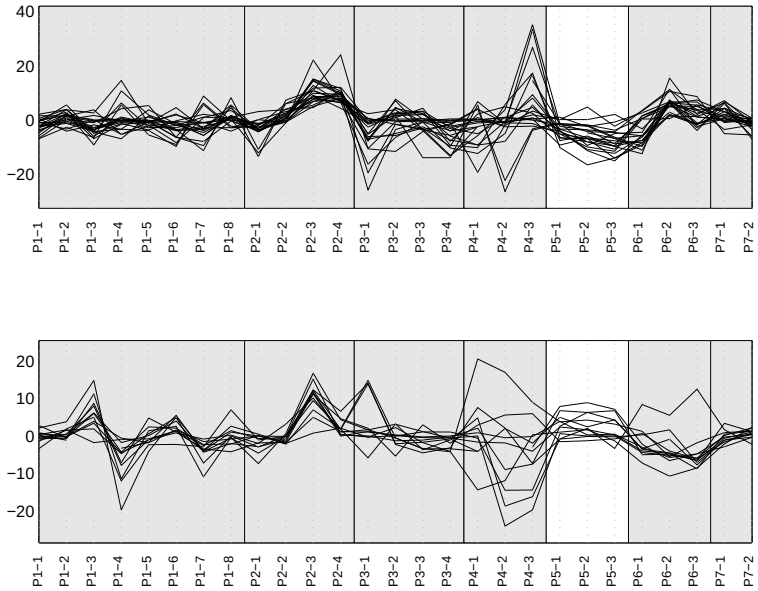

FIGURE 1. *Expression profiles of Clusters 5 and 9. The background color is white when the project is irrelevant.*

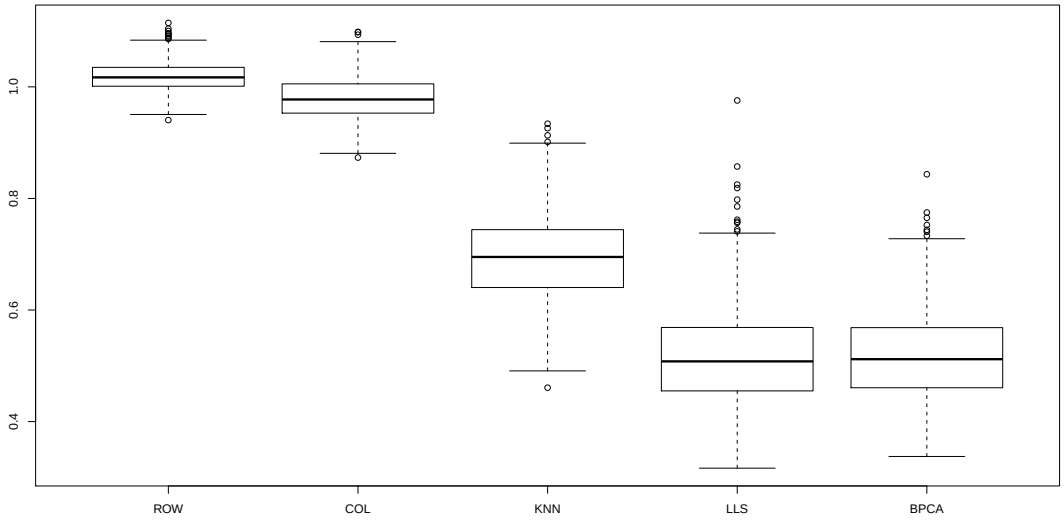

FIGURE 2. *Variability of the NRMSE for the different imputation methods (from left to right: ROW-,COL-,KNN-, LLS-, BPCA-impute methods).*

 ISSN: 2102-6238

$c = 0.38\%$ of missing values among the 1020 totally observed genes and we repeat this 1000 times to evaluate their NRMSE. Results are summarized in Figure 2. ZERO-imputation results are not given since by construction, NRMSE of this method always equals 1.

As previously observed on the simulated dataset, KNN-, LLS- and BPCA- imputations are better than imputations not taking data correlation structure into account. The KNN-imputation method has an intermediate behavior between the simple ZERO-, ROW- and COL-imputation methods and the better methods LLS- and BPCA-imputation methods. The ROW-imputation method behaves often worse. This method is clearly inappropriate for transcriptome data since a gene is often differentially expressed from an experiment to an other. The better results are given by the LLS- and the BPCA-imputation methods despite a large variability. Consequently, we only applied *SelvarClust* combined with LLS- or BPCA-imputation methods.

 ISSN: 2102-6238

TABLE 3. *Contingency table between SelvarClustMV (row) and SelvarClust without missing values (column). The number of genes per cluster for the dataset with missing values given in the second column, is followed by: [the number of genes totally observed and studied in [13], number of genes with at least one missing value, the number of genes totally observed but not studied in [13]].*

		Cl 6	Cl 12	Cl 2	Cl 10	Cl 9	Cl 3	Cl 5	Cl 13	Cl 17	Cl 8	Cl 1	Cl 4	Cl 7	Cl 15	Cl 11	Cl 14	Cl 16
Cl 8	510 [420, 56, 34]	**383**	0	1	6	3	1	6	0	0	10	0	4	2	0	4	0	0
Cl 1	143 [132, 10, 1]	0	**130**	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
Cl 14	167 [133, 20, 14]	1	5	**106**	11	0	0	0	0	0	6	0	0	4	0	0	0	0
Cl 15	143 [95, 8, 40]	11	7	1	**76**	0	0	0	0	0	0	0	0	0	0	0	0	0
Cl 3	67 [50, 8, 9]	1	3	5	0	**38**	0	0	0	1	0	0	1	1	0	0	0	0
Cl 11	40 [32, 0, 4]	5	0	0	0	0	**24**	0	0	0	0	0	0	1	0	0	0	2
Cl 12	36 [32, 7, 1]	0	2	1	0	0	0	**22**	0	0	0	0	1	5	0	1	0	0
Cl 10	33 [30, 1, 2]	20	0	0	0	0	0	7	0	0	1	0	0	0	0	2	0	0
Cl 7	28 [21, 2, 5]	0	0	0	0	0	0	0	**19**	0	0	0	0	1	0	1	0	0
Cl 16	24 [19, 0, 5]	0	0	0	0	0	0	0	0	**19**	0	0	0	0	0	0	0	0
Cl 13	16 [13, 0, 4]	0	0	0	0	0	0	0	0	0	**11**	0	0	0	1	1	0	0
Cl 5	22 [10, 3, 9]	3	0	0	0	1	0	6	0	0	0	0	0	0	0	0	0	0
Cl 6	8 [7, 0, 1]	0	0	0	0	0	0	1	0	0	0	**6**	0	0	0	0	0	0
Cl 9	12 [10, 2, 0]	0	2	6	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Cl 2	6 [6, 0, 0]	0	0	0	0	0	0	0	0	0	0	1	**5**	0	0	0	0	0
Cl 4	8 [6, 1, 1]	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**	1	0	0
Cl 17	4 [4, 0, 0]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0

TABLE 4. *Contingency table between SelvarClustMV (row) and SelvarClust combined with LLS-imputation (column)*

	Cl 1	Cl 8	Cl 13	Cl 3	Cl 6	Cl 7	Cl 9	Cl 14	Cl 11	Cl 4	Cl 2	Cl 12	Cl 15	Cl 10	Cl 5
Cl 8	**477**	**15**	0	1	0	0	0	5	1	3	0	8	0	0	0
Cl 2	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
Cl 1	0	1	**137**	0	0	0	0	0	3	0	0	2	0	0	0
Cl 3	3	3	4	**52**	0	0	0	0	4	0	0	0	1	0	0
Cl 4	0	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0
Cl 5	0	0	0	0	0	**21**	0	0	0	0	0	1	0	0	0
Cl 7	0	0	0	0	0	0	**27**	0	1	0	0	0	0	0	0
Cl 10	1	1	1	0	0	0	0	**26**	1	3	0	0	0	0	0
Cl 14	1	0	2	2	0	0	0	1	**156**	1	0	4	0	0	0
Cl 12	1	2	2	0	0	1	0	0	1	**29**	0	0	0	0	0
Cl 15	0	0	0	0	0	0	0	0	0	0	**31**	**112**	0	0	0
Cl 16	0	0	0	0	0	0	0	0	0	0	0	0	**24**	0	0
Cl 11	10	0	0	2	0	0	0	0	0	1	0	1	1	**25**	0
Cl 17	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1
Cl 6	0	1	0	3	0	0	0	0	1	2	0	0	0	0	1
Cl 9	0	1	5	3	0	0	0	0	2	0	0	1	0	0	0
Cl 13	2	0	1	0	3	0	0	1	9	0	0	0	0	0	0

Concerning the variable partition, the results are analogous to *SelvarClustMV* . Projects 1, 2, 3, 4, 6 and 7 are relevant and the four last ones are required to explain Project 5. The difference comes from the cluster number, which is smaller since *SelvarClust* with LLS selects 15 clusters. The contingency table between *SelvarClustMV* and *SelvarClust* with LLS (see Table 4) shows that the same gene clusters are defined, except Clusters 6, 9, 13 and 17 of *SelvarClustMV* which are scattered in the other clusters. Moreover a closer look shows that Clusters 6, 13 and 17 are mainly composed of genes totally observed (see Table 3), clustered together in the study of [13] and kept by *SelvarClustMV*. When *SelvarClust* is combined with BPCA-imputation, the same phenomenon is observed: 16 clusters are selected and genes of Clusters 6, 9, 13 and 17 are also scattered (data not shown). Consequently we think that the results of *SelvarClust* combined with an imputation method are less relevant since some biological interpretations are lost.

6. Discussion

In this paper, we were interested in the adaptability of the variable selection in model-based clustering proposed by [13, 14] to study datasets with missing values. It requires a new strategy to calculate the model selection criterion via the explicit expression of the observed loglikelihood and a new parameter estimation method. These different changes alter the backward stepwise algorithms and the code for constant variance matrix is available for the SR-model. The generalization for the whole mixture forms is straightforward and requires only a coding work.

Objects are assigned to a cluster according to the *Maximum A Posteriori* rule. Since the conditional probabilities are based on the relevant variables, objects which have not at least one value for the relevant variables are not clustered. It concerns a limited number of objects and solutions exist to overcome this problem. For example an intuitive solution would consist of attributing such object to the cluster for which the average profile on all the variables is the closest. This strategy is suitable if irrelevant variables are numerous and linked to the whole relevant variables.

Imputation methods are easy to use but could have a major impact on results. Missing value management inside the statistical model avoids this pre-processing step, often tricky. Applications on simulated data and real data shown that the variable selection procedure combined with an imputation method could have difficulties to find the variable partition. Whereas *SelvarClustMV* has a good behavior up to 20% of missing values. Moreover *SelvarClustMV* seems easier to keep the structure between genes totally observed. We emphasize that our method can be useful for other post-genomic data as well as proteins or metabolites. Our method is neither organism- nor data specific.

References

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition, 2003.

[2] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, 51(2):587–600, 2006.

[3] G. Brock, J. Shaffer, R. Blakesley, M. Lotz, and G. Tseng. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*, 9(1):12, 2008.

[4] M. Celton, A. Malpertuy, G. Lelandais, and A. de Brevern. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, 11(1):15, 2010.

 ISSN: 2102-6238

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.
[6] S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J.-P. Renou, A. Lecharny, and V. Brunaud. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36(Database Issues):986–990, 2008.
[7] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein. Imputing Missing Data for Gene Expression Arrays. Technical report, Stanford University Statistics Department, 1999.
[8] J. Hu, H. Li, M.S. Waterman, and X.J. Zhou. Integrative missing value estimation for microarray data. *BMC Bioinformatics*, 7:449, 2006.
[9] R. Jornsten, H.Y. Wang, W. J. Welsh, and M. Ouyang. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22):4155–4161, 2005.
[10] H. Kim, G. H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.
[11] M.H. Law, M.A.T Figueiredo, and A.K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.
[12] R.J.A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, USA, 1986.
[13] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65:701–709, 2009.
[14] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, 53:3872–3882, 2009.
[15] C. Maugis, M.-L. Martin-Magniette, J.-P. Tamby, J.-P. Renou, A. Lecharny, S. Aubourg, and G. Celeux. Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins. *Modulad*, 40, 2009.
[16] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, November 2003.
[17] M. Ouyang, W. J. Welsh, and P. Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923, 2004.
[18] A. E. Raftery and N. Dean. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.
[19] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, December 1976.
[20] J. L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, London, 1997.
[21] M. S. Sehgal, I. Gondal, L.S. Dooley, and R. Coppel. Ameliorative lmissing value imputation for robust biological knowledge inference. *Journal of Biomedical Informatics*, 41:499–514, 2008.
[22] Y. Sun, U. Braga-Neto, and E. R. Dougherty. Impact of Missing Value Imputation on Classification for DNA Microarray Gene Expression Data-A Model-Based Study. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009, 2009.
[23] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.
[24] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio. Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, 22(5):566–572, 2006.
[25] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics*, 9:202, 2008.
[26] X. Wang, A. Li, Z. Jiang, and H. Feng. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, 7(32), 2006.

 ISSN: 2102-6238