Noname manuscript No.
(will be inserted by the editor)

Unifying Data Units and Models in (Co-)Clustering

Christophe Biernacki · Alexandre Lourme

Received: date / Accepted: date

Abstract Statisticians are already aware that any modelling process issue (exploration, prediction) is wholly data unit dependent, to the extend that it should be impossible to provide a statistical outcome without specifying the *couple* (unit,model). In this work, this general principle is formalized with a particular focus in model-based clustering and co-clustering in the case of possibly mixed data types (continuous and/or categorical and/or counting features), being also the opportunity to revisit what the related data units are. Such a formalization allows to raise three important spots: (*i*) the couple (unit,model) is not identifiable so that different interpretations unit/model of the same whole modelling process are always possible; (*ii*) combining different "classical" units with different "classical" models should be an interesting opportunity for a cheap, wide and meaningful enlarging of the whole modelling process family designed by the couple (unit,model); (*iii*) if necessary, this couple, up to the non identifiability property, could be selected by any traditional model selection criterion. Some experiments on real data sets illustrate in detail practical beneficits from the previous three spots.

Keywords Measurement units · Mixed data · Mixture models · Model selection · Non identifiability

1 Introduction

Usually, statistical analysis relies on two coupled and fundamental materials: data and models. This basic description can be further enriched by dividing

C. Biernacki
University of Lille & CNRS & Inria
E-mail: Christophe.Biernacki@math.univ-lille1.fr

A. Lourme
University of Bordeaux
E-mail: alexandre.lourme@u-bordeaux.fr

data into two complementary parts [25,54,30]: the object features (namely: the variables) and the associated measurement units (namely the units). Substantially, data (variables and units) are expected to be provided by the *practitioner* whereas models are the domain of the *statistician* but, in practice, both of them are not completely unrelated as we illustrate now both in the predictive and in the descriptive contexts.

In the predictive framework, data are composed by, for instance, one non-negative predictor variable x provided in order to predict a real outcome variable y, where variables x and y and their associated units are practitioner-defined. In such a situation, the statistician could propose the standard linear regression model [52] $y = \beta x + e$, with β a real but unknown parameter and e the realization of a standardized normal distribution. Alternatively, he could propose two other models which are $y = e$ and $y = \beta \ln(x) + e$. The former model equivalently corresponds either to a new regression model ($\beta = 0$), or to a variable selection situation (variable x has been canceled). The latter model equivalently corresponds either to a new regression model (a logarithm regression model) with x still expressed in the initial practitioner-defined unit, or to the initial linear regression model where the unit of variable x is now expressed on a logarithm scale. Thus, in this context, there exists a straightforward bridge between models and data (variables and units), leading in particular to non-identifiability situations from the interpretation point of view. As a matter of fact, the practitioner may beneficit from this link by using directly any model selection paradigm (as BIC [51], cross-validation. . .) for helping him in his task.

In the descriptive framework, where the clustering task is emblematic, [22,21] identified the following associated fundamental challenges, namely: (*i*) “What is a cluster?”, (*ii*) “What features should be used?”, (*iii*) “Should the data be normalized?”. Such challenges arise even before the so much discussed in literature “selecting the number of clusters”, only numbered (*vi*) in [21]. We recognized challenges (*ii*) and (*iii*) as being respectively the variable selection and the unit definition problems we previously discussed. Mixture of distributions is now a classical way for answering (*i*), a cluster being itself modelled by a homogeneous distribution. This approach met many successes from both the practical and the theoretical point of views (see for instance a survey in [17,40,38]). In addition, success of mixtures is also sensible on questionings (*ii*) and (*iii*) thanks to their comprehensive modelling property, allowing to reformulate them as particular mixture models. Thus, concerning (*ii*), several attempts essentially focused in the Gaussian mixture setting exist, as the SRUW modelling and related works [26,55,45,33,34,53,32] or also other ℓ_1 penalization procedures combined with variable unit transformation (centering) leading to so-called PS-Lasso [44,60] and Lasso-MLE [41,42] strategies. Concerning (*iii*) now, the unit definition has been also recasted as a particular mixture model, essentially in the Gaussian setting also [59,56,61,14].

However, such works aiming at embedding data units and models in clustering are only early, albeit attractive, attempts failing to fully reveal all its potential for clustering. In particular, we defend that a suitable formulation of data units as a couple (unit,model) is an opportunity to enlarge without any effort and in a meaningful way the whole traditional model collection, with benefits for both the practitioner and the statistician. From this key recast, it becomes possible to extend straightforwardly these model families in the case of non continuous data (categorical and counting data typically) and to address also similarly the model-based co-clustering framework (see a review of this topic in [18]).

The outline of this work is the following. In Section 2, notations, estimation and model selection are fixed for model-based clustering and model-based co-clustering by underlying also the case of various types of data (continuous, categorical, counting). Section 3 introduces formalization of data units, its fusional link with modelling and related gains that both the practitioner and the statistician may expect from this new light. In Section 4 and 5, various real data sets in clustering and co-clustering, respectively, and for different data types, are involved to illustrate the practical interest of the previous concept unifying units and models. Finally, Section 6 concludes this work by sketching also some future prospects.

In the following, sets, sums and products on i, j, k and l stands for ranges $\{1,\ldots,n\}$, $\{1,\ldots,d\}$, $\{1,\ldots,K\}$ and $\{1,\ldots,L\}$ respectively. Also, capital letters designate random variables/vectors.

2 Model-based (co-)clustering for multiple data kinds

2.1 Model-based clustering

Mixture hypothesis Cluster analysis is one of the main data analysis method. It aims at partitioning a data set $\mathbf{x}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)=(\mathbf{x}^O,\mathbf{x}^M)$, composed by n individuals $\mathbf{x}_i=(x_{i1},\ldots,x_{id})$ of dimension d, and lying in a space $\mathbb{X}$, into K classes $G_1,\ldots,G_K$. Here the observed part of $\mathbf{x}$ has been denoted by $\mathbf{x}^O$ whereas the missing is denoted by $\mathbf{x}^M$. Moreover, $\mathbb{X}$ designates possibly a mixed feature space, it means a space mixing features of different kinds like continuous, categorical or integer.

The target partition is denoted by $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)$, lying in a space $\mathbb{Z}$, where $\mathbf{z}_i=(z_{i1},\ldots,z_{iK})'$ is a vector of $\{0,1\}^K$ such that $z_{ik}=1$ if individual $\mathbf{x}_i$ belongs to the kth class G_k, and $z_{ik}=0$ otherwise. Model-based clustering allows to reformulate cluster analysis as a well-posed estimation problem both for the partition $\mathbf{z}$ and for the number K of classes. It considers data $\mathbf{x}_1,\ldots,\mathbf{x}_n$ as n i.i.d. realizations of a mixture distribution $\mathrm{p}(\cdot;\boldsymbol{\theta})=\sum_k \pi_k \mathrm{p}(\cdot;\boldsymbol{\alpha}_k)$, where $\mathrm{p}(\cdot;\boldsymbol{\alpha}_k)$ generically indicates the probability distribution function (pdf), parameterized by $\boldsymbol{\alpha}_k$, associated to the class k, where π_k indicates the mixture proportion of this component ($\sum_k \pi_k=1$, $\pi_k\geq 0$) and where $\boldsymbol{\theta}=\{(\pi_k,\boldsymbol{\alpha}_k)\}$ indicates the whole mixture parameters.

Mixture parameter estimation From the observed data set $\mathbf{x}^O$ it is then possible to obtain a mixture parameter estimate $\hat{\boldsymbol{\theta}}$ by maximizing the observed log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{x}^O) = \ln \mathrm{p}(\mathbf{x}_i^O; \boldsymbol{\theta})$ where

$$\mathrm{p}(\mathbf{x}_i^O; \boldsymbol{\theta}) = \sum_k \pi_k \mathrm{p}(\mathbf{x}_i^O; \boldsymbol{\alpha}_k) = \sum_k \pi_k \int_{\mathbf{x}_i^M} \mathrm{p}(\mathbf{x}_i^O, \mathbf{x}_i^M; \boldsymbol{\alpha}_k) d\mathbf{x}_i^M, \qquad (1)$$

provided that missing data $\mathbf{x}^M$ are obtained by a missing at random (MAR) process [28].

For optimizing $\ell(\boldsymbol{\theta}; \mathbf{x}^O)$, the EM (Expectation-Maximization) algorithm of [13] is often performed, or some of its variants (see also [49]) like the SEM (Stochastic EM) [11]. Indeed, a SEM algorithm can be used to maximize the observed-data log-likelihood, described as follows for iteration $q \geq 1$, when starting from a parameter $\boldsymbol{\theta}^{(0)}$ selected at random:

- **E-step**: compute conditional probabilities $\mathrm{p}(\mathbf{x}_i^M, \mathbf{z}_i | \mathbf{x}_i^O; \boldsymbol{\theta}^{(q-1)})$,
- **S-step**: draw $(\mathbf{x}_i^{M(q)}, \mathbf{z}_i^{(q)})$ from $\mathrm{p}(\mathbf{x}_i^M, \mathbf{z}_i | \mathbf{x}_i^O; \boldsymbol{\theta}^{(q-1)})$,
- **M-step**: maximize $\boldsymbol{\theta}^{(q)} = \arg\max_{\boldsymbol{\theta}} \ln \mathrm{p}(\mathbf{x}^O, \mathbf{x}^{M(q)}, \mathbf{z}^{(q)}; \boldsymbol{\theta})$.

Since the parameter sequence $(\boldsymbol{\theta}^{(q)})$ generated by SEM does not punctually converges, due to the S-step definition, the algorithm generally stops after a predefined number of iterations. This sequence converges in distribution towards the unique stationary distribution. Asymptotically on q, the mean of the sequence $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}$ and thus provides a sensible local estimate of the maximum likelihood. In addition, the variance of the sequence $(\boldsymbol{\theta}^{(q)})$ gives confidence intervals on $\boldsymbol{\theta}$. SEM has also advantage to be less dependent on the initial value $\boldsymbol{\theta}^{(0)}$ than EM does if a “sufficient” iteration number is performed and so avoids uninteresting local maxima. Finally, managing missing data is easier than with EM thanks to its so-called stochastic S-step, while preserving a classical M-step like EM.

Partition (and missing data) estimation Once $\hat{\boldsymbol{\theta}}$ is obtained, a so-called SE algorithm (a SEM without the M step) can be used to estimate partition $\mathbf{z}$, and simultaneously missing data $\mathbf{x}^M$. Its qth iteration is given by

- **E-step**: compute conditional probabilities $\mathrm{p}(\mathbf{x}^M, \mathbf{z} | \mathbf{x}^O; \hat{\boldsymbol{\theta}})$,
- **S-step**: draw $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$ from $\mathrm{p}(\mathbf{x}^M, \mathbf{z} | \mathbf{x}^O; \hat{\boldsymbol{\theta}})$.

After a given iteration number, the mean and/or mode of the sequence $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$ estimates $(\mathbf{x}^M, \mathbf{z})$, denoted by $(\hat{\mathbf{x}}^M, \hat{\mathbf{z}})$, with again the possibility to derive some confidence intervals on these unknown quantities.

Estimation of the class number From the Bayesian model selection principle, it is now possible to derive an estimate $\hat{K}$ from an estimate of the observed conditional probability $\hat{\mathrm{p}}(K | \mathbf{x}^O)$ or also from the completed-partition conditional probability $\hat{\mathrm{p}}(K | \mathbf{x}^O, \mathbf{z})$. The first one leads to retaining $\hat{K}$ which minimizes the so-called BIC (Bayesian Information Criterion) criterion [51],

$$\mathrm{BIC} = \frac{D}{2} \ln n - \ln \mathrm{p}(\mathbf{x}^O; \hat{\boldsymbol{\theta}}), \qquad (2)$$

whereas the second one corresponds to minimizing the so-called ICL (Integrated Completed Likelihood) criterion [5], defined by

$$\mathrm{ICL} = \frac{D}{2} \ln n - \ln \mathrm{p}(\mathbf{x}^O, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}), \tag{3}$$

D denoting the number of free (continuous) parameters in the model at hand.

Continuous data and Gaussian distribution The multivariate mixture model is certainly the most known and used model for continuous data. It has a long history of use in clustering (see for instance [58], [9]). In that case, $\mathbf{x}_i$ are continuous variables ($\mathbb{X} = \mathbb{R}^d$) and the conditional density of components is written $\mathrm{p}(\cdot; \boldsymbol{\alpha}_k) = \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ the component mean (or centre) and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ its covariance matrix. Since $\boldsymbol{\Sigma}_k$ requires to estimate a quadratic number of parameters with D, there exists also many parsimonious and meaningful constraints on it relying on the spectral decomposition [12], on factor analyzers [15,36,37] or also on the so-called statistical decomposition RTV [8]. For instance, the spectral family of [12] includes the diagonal case assuming that all variables x_{ij} of $\mathbf{x}_i$ are *conditionally independent* knowing the latent classes. Thus,

$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_j \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{kj}) \tag{4}$$

where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{kj})$, $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kj})$ denoting the univariate distribution associated to the variable j in the class k, with $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{N}(\mu_{kj}, \sigma^2_{kj})$. More information is provided on this family in Section 4.1.

Other data types and related distributions It is also possible to easily extend the previous diagonal case to all kinds of data types by assuming again conditional independence (4) knowing the latent classes. Thus, only the univariate distribution associated to the variable j in the class k has to be defined, depending on the data type:

- **Categorical**: given variable j is categorical, $x_{ij} = (x_{ijh}; h = 1, \ldots, m_j)$ has m_j response levels where $x_{ijh} = 1$ if i has response level h for variable j and $x_{ijh} = 0$ otherwise. The standard model for clustering observations described through categorical variables is the so-called latent class model (see for instance [16]) where $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{M}(\boldsymbol{\alpha}_{kj})$ is the multinomial distribution with $\boldsymbol{\alpha}_{kj} = (\alpha_{kjh}; h = 1, \ldots, m_j)$, α_{kjh} denoting the probability that variable j has level h for one individual in cluster k.
- **Integer**: given variable j is a count, each $x_{ij} \in \mathbb{N}$ and $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{P}(\lambda_{kj})$, the Poisson distribution of parameter λ_{kj}.
- **Other**: x_{ij} could be also an ordinal data or a ranking data, for instance (see respective univariate distributions in [7,6]).

Mixed data type and related distribution It is frequent in practice to mix different kinds of data types, for instance continuous, categorical and integer ones. Thus the ith individual is composed by three parts, $\mathbf{x}_i = (\mathbf{x}_i^{cont}, \mathbf{x}_i^{cat}, \mathbf{x}_i^{int})$, $\mathbf{x}_i^{cont}$, $\mathbf{x}_i^{cat}$ and $\mathbf{x}_i^{int}$ designing the continuous, the categorical and the integer ones respectively. In that case, the proposed solution for symmetry between data types is to mix all types by inter-type conditional independence [43]:

$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \mathrm{p}(\mathbf{x}_i^{cont}; \boldsymbol{\alpha}_k^{cont}) \times \mathrm{p}(\mathbf{x}_i^{cat}; \boldsymbol{\alpha}_k^{cat}) \times \mathrm{p}(\mathbf{x}_i^{int}; \boldsymbol{\alpha}_k^{int}) \tag{5}$$

with $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_k^{cont}, \boldsymbol{\alpha}_k^{cat}, \boldsymbol{\alpha}_k^{int})$ the obvious associated parameters by data type.

2.2 Model-based co-clustering

Mixture hypothesis Simultaneous clustering of rows and columns, usually designated by bi-clustering, co-clustering or block clustering, is an important technique in two way data analysis allowing very simple models even with many variables. They consider the two sets simultaneously and organize the data into homogeneous blocks. Two partition representations are thus now needed. First, as usual, a partition of n individuals (lines of the data matrix $\mathbf{x}$) into K clusters still noticed $\mathbf{z}$. Second, and symmetrically, a partition of d variables (columns of the data matrix $\mathbf{x}$) into L clusters is denoted by $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$ where $\mathbf{w}_j = (w_{j1}, \ldots, w_{jL})$ with $w_{jl} = 1$ if j belongs to cluster l and $w_{jl} = 0$ otherwise. Both space partitions are respectively denoted by $\mathbb{Z}$ and $\mathbb{W}$.

We refer to the book of [18] for providing more details on co-clustering techniques, probabilistic or not. Here, we focus on model-based co-clustering as being often a generalization of non-probabilistic methods and allowing coherent formulation from estimation to model selection. Block model-based clustering can be seen as an extension of the traditional mixture model-based clustering previously described in Section 2.1. The basic idea is to extend the latent class principle of local (or conditional) independence expressed in (4): each data point x_{ij} is assumed to be independent once $\mathbf{z}_i$ and $\mathbf{w}_j$ are fixed. We note $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$ and where $\boldsymbol{\pi} = (\pi_k)$ and $\boldsymbol{\rho} = (\rho_k)$ are the vectors of probabilities π_k and ρ_l that a row and a column belong to the kth row component and to the lth column component, respectively. Assuming also independence between all $\mathbf{z}_i$ and $\mathbf{w}_j$, the latent block mixture model has the global probability distribution

$$\mathrm{p}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,j,k,l} \left(\pi_k \rho_l \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{kl})\right)^{z_{ik} w_{jl}} . \tag{6}$$

Finally, the distribution $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kl})$ depends on the data type of x_{ij} (continuous, categorical, integer) and thus is similar to these ones defined in Section 2.1, except that mixed data are not allowed this time. Such models can be very parsimonious even when d is very large, provided that L is moderate. Indeed, by comparison to a classical intra-type conditional independence model with D parameters to be estimated (see Section 2.1), the corresponding co-clustering model requires only $D \times \frac{L}{d}$ parameters.

Mixture parameter estimation EM-based algorithms are the standard approach to estimate model parameters by maximizing the observed log-likelihood. Here, the complete data is represented as a vector $(\mathbf{x}, \mathbf{z}, \mathbf{w})$ where unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. Unfortunately, difficulties arise owing to the dependence structure in the model, and more precisely in the combinatorial difficulty for evaluating the terms $\mathrm{p}(Z_{ik} = 1, W_{jl} = 1|\mathbf{x}; \boldsymbol{\theta}^{(q)})$ involved in the E step (see details in [18]). Several solutions exist for skirting this difficulty, including the so-called *variational approach* which constraints the problematic joint probability to satisfy the relation $\mathrm{p}(\mathbf{z}, \mathbf{w}|\mathbf{x}; \boldsymbol{\theta}) \approx \mathrm{p}(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\mathrm{p}(\mathbf{w}|\mathbf{x}; \boldsymbol{\theta})$. Alternatively, the E-step can be replaced by a S-step by using a SEM algorithm instead of EM (see details on SEM in Section 2.1). In the S-step, random couples $(\mathbf{z}, \mathbf{w})$ (conditionnally to $\mathbf{x}$) are drawn sequentially by the following two-step Gibbs algorithm (see more details in [24]): $\mathbf{Z}|\mathbf{x}, \mathbf{w}; \boldsymbol{\theta}$ and $\mathbf{W}|\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}$. Estimating the block membership designed by the pair $(\mathbf{z}, \mathbf{w})$ can then be performed by a SE algorithm similar to this one described in Section 2.1.

Estimation of the block number In addition, a specific expression of the ICL criterion (3) can be invoked for selecting the pair (K, L) (see [29] and [24] which provide ICL expressions for the Gaussian situation and for the Bernoulli/multinomial case, respectively).

3 Data units as a piece of model and associate properties

3.1 Formalization of data unit transformation

Data unit as a bijective mapping Data $\mathbf{x}$, lying in a space $\mathbb{X}$, are implicitly provided by a practitioner measurement unit denoted by $\mathbf{id}$ which acts as a kind of canonical unit from the practitioner point of view. Consequently, both the data $\mathbf{x}$ and its measurement space $\mathbb{X}$ should be indexed by this unit $\mathbf{id}$, leading respectively to new notations $\mathbf{x}^{\mathbf{id}}$ and $\mathbb{X}^{\mathbf{id}}$, even if it should be more convenient in practice to simplify them into more traditional, but equivalent, notations $\mathbf{x} = \mathbf{x}^{\mathbf{id}}$ and $\mathbb{X} = \mathbb{X}^{\mathbf{id}}$.

We denote by $\mathbf{u}$ a new measurement unit for data $\mathbf{x}$, being now expressed as a data set $\mathbf{x}^{\mathbf{u}}$ lying in the corresponding new space $\mathbb{X}^{\mathbf{u}}$. Formally, such a data unit transformation $\mathbf{u}$ acts as the following general *bijective* mapping:

$$\begin{aligned} \mathbf{u} : \mathbb{X} = \mathbb{X}^{\mathbf{id}} &\longrightarrow \mathbb{X}^{\mathbf{u}} \\ \mathbf{x} = \mathbf{x}^{\mathbf{id}} = \mathbf{id}(\mathbf{x}) &\longmapsto \mathbf{x}^{\mathbf{u}} = \mathbf{u}(\mathbf{x}). \end{aligned} \tag{7}$$

The bijective assumption over $\mathbf{u}$ is important to preserve the whole data set information quantity when performing unit transformations. We denote by $\mathbf{u}^{-1}$ the reciprocal of $\mathbf{u}$, so $\mathbf{u}^{-1} \circ \mathbf{u} = \mathbf{id}$. Moreover, in this setting, $\mathbf{id}$ is thus only a particular unit $\mathbf{u}$.

However, within this too general framework we add the two realistic constraints on $\mathbf{u}$ that it proceeds only both individual and variable wise, it means lines by lines and rows by rows in the whole data set $\mathbf{x}$. Such constraints are expressed respectively by

$$\mathbf{u}(\mathbf{x}) = (\mathbf{u}(\mathbf{x}_1), \dots, \mathbf{u}(\mathbf{x}_n)), \tag{8}$$

and

$$\mathbf{u}(\mathbf{x}_i) = (\mathbf{u}_1(x_{i1}), \dots, \mathbf{u}_d(x_{id})). \tag{9}$$

Notation $\mathbf{u}(\mathbf{x}_i)$ straightforwardly means that transformation $\mathbf{u}$ is applied to the particular data set $\mathbf{x}_i$, restricted to the single individual i. Notation $\mathbf{u}_j$ corresponds to the specific (bijective) transformation unit associated to variable j.

It is obviously possible to weaken these assumptions by cancelling the variable wise hypothesis defined in (9). In particular, this relaxation would encompass important transformations such that the linear ones involved in principal component analysis (PCA), when all $x_{ij} \in \mathbb{R}$. Nevertheless, restriction (9) has advantage to respect the variable definition, transforming only its unit.

Categories of data units according to data types It is crucial to notice that allowed units $\mathbf{u}_j$ depend on the kind of variable j (continuous, binary, integer...) as we described now. In each case, all bijective mappings are possible while respecting the space where variable j is. Here are some typical situations (non exhaustive) of unit tranformation in each case:

- **Continuous variables**: It corresponds to $x_{ij} \in \mathbb{R}$, or a subset of $\mathbb{R}$ like $\mathbb{R}^+$. The simplest and probably the most straightforward unit transformation is the scaling and shifting one, namely

$$x_{ij}^{\mathbf{u}_j} = a_j x_{ij} + b_j, \tag{10}$$

  with $a_j \in \mathbb{R}\backslash\{0\}$ and $b_j \in \mathbb{R}$. It is the typical situation when transforming feet (F) unit into inches (I) unit ($F = 12I$, thus $b_j = 0$), when converting Celsius (C) unit into Fahrenheit (F) unit ($C = 5F/9 - 160/9$, thus $b_j \neq 0$), or when performing any standardization for being "unit-free" ($x_{ij}^{\text{stand}} = x_{ij}/\hat{\sigma}_j - \hat{\mu}_j/\hat{\sigma}_j$, with $\hat{\mu}_j$ and $\hat{\sigma}_j$ respectively the mean and the standard deviation of the whole marginal sample). Also, other transformations are classical, for instance the logarithm scale, namely $x_{ij}^{\mathbf{u}_j} = a_j \ln(x_{ij})$, when obtaining the decibel unit from the ratio of the two power levels unit (measured power and reference power).
- **Counting variables**: It corresponds to integer values for counting, thus $x_{ij} \in \mathbb{N}$. Two common unit transformations (for variable j) can be the *shifted* one $\mathbf{u}_j^{\text{shift}}(x_{ij}) = x_{ij} - b_j$ with $b_j \in \mathbb{N}$ or the *scaled* one $\mathbf{u}_j^{\text{scale}}(x_{ij}) = a_j x_{ij}$ with $a_j \in \mathbb{N}\backslash\{0\}$. As an illustration of both units, consider x_{ij} to be the *total* number of educational years (canonical unit $\mathbf{id}_j$) of student i. Alternatively, it is possible to propose either $\mathbf{u}_j^{\text{shift}}(x_{ij}) = x_{ij} - 8$ as being the *university* number of educational years[1], or $\mathbf{u}_j^{\text{scale}}(x_{ij}) = 2x_{ij}$ as being the total number of educational *semesters*.

[1] Eight is the number of years spent by English pupils in a secondary school.

- **Categorical variables**: It corresponds to the m_j dimensional vector of binary values, thus $x_{ij} \in \{0,1\}^{m_j}$ with $\sum_{h=1}^{m_j} x_{ijh} = 1$, m_j denoting the level number (see Section 2.1). In this situation, possible unit transformations $\mathbf{u}_j$ are exhaustively restricted to all *permutations* of level coding. It is denoted by $\mathbf{u}_j^{\text{perm}} \in \mathcal{P}_{m_j}$, where $\mathcal{P}_{m_j}$ is the standard symmetric permutation group on $\{1,\ldots,m_j\}$. As a straightforward example, consider the following binary case ($m_j = 2$) with the canonical unit $\mathbf{id}_j$: $x_{ij} = (0,1)$ corresponds to holidays in the mountains whereas $x_{ij} = (1,0)$ corresponds to holidays at the sea. The (only) alternative unit is here the permuted (or reverse) unit $\mathbf{u}_j^{\text{perm}}(x_{ij}) = (1 - x_{ij1}, 1 - x_{ij2})$. Concretely, the new unit $\mathbf{u}_j^{\text{perm}}$ is as follows: $(0,1)$ designates now holidays at the sea whereas $(1,0)$ corresponds now to holidays in the mountains. Application for categorical non-binary data (more than two levels) will be illustrated through a real data set of Section 5.2.
- **Other types of variables**: Other situations should be approached case by case. Some common types of variables are in fact particular categorical variables, as *ordinal* variables are. For instance, consider the school level (variable j) of pupil i among values $\{\text{high grade}, \text{middle grade}, \text{low grade}\}$. A canonical unit $\mathbf{id}_j$ can be expressed by: high grade $>$ middle grade $>$ low grade, where " $>''$ means "greater in *strength* than". Alternatively, the (only) other unit $\mathbf{u}_j^{\text{perm}}$ is expressed by: low grade $>$ middle grade $>$ high grade , where " $>''$ means now "greater in *weakness* than".

3.2 Revisiting units as a modelling component

Classically, the couple composed by the parametric pdf $\mathrm{p}(\cdot;\boldsymbol{\theta})$ and a space $\Theta_{\mathbf{m}}$ where evolves this parameter defines a so-called *model*, denoted now by $\mathrm{p}_{\mathbf{m}}$:

$$\mathrm{p}_{\mathbf{m}} = \{\cdot \in \mathbb{X} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}. \tag{11}$$

In practice it will be sometimes a convenient shortcut to confound the index $\mathbf{m} \in \mathbb{M}$ and the corresponding distribution parameterized by $\Theta_{\mathbf{m}}$. In model-based clustering (see Section 2.1), a model is defined simultaneously by a number of clusters K and each component distribution $\mathrm{p}(\cdot;\boldsymbol{\alpha}_k)$, including eventual constraints on $\boldsymbol{\pi}$ and $\boldsymbol{\alpha}$. In model-based co-clustering (see Section 2.2), a model is defined simultaneously by a number of clusters in lines K and in columns L and each block distribution $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{kl})$, including eventual constraints on $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$.

However, model definition (11) can be equivalently rewritten by explicitly expressing the canonical data units $\mathbf{id}$ (see Section 3.1)

$$\mathrm{p}_{\mathbf{m}}^{\mathbf{id}} = \{\cdot \in \mathbb{X}^{\mathbf{id}} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}, \tag{12}$$

enlightening that the space where evolve the variables and the probabilistic modelling are irrevocably embedded. This phenomenon is totally related to the standard probability theory where any probabilistic modelling explicitly gathers both a *sample space* and a *probability measure*. By pursuing this idea, changing the previous units from $\mathbf{id}$ to $\mathbf{u}$, while preserving the same probabilistic distribution indexed by $\mathbf{m}$, is expected to produce a new probabilistic modelling expressed by

$$\mathrm{p}_{\mathbf{m}}^{\mathbf{u}} = \{\cdot \in \mathbb{X}^{\mathbf{u}} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}. \tag{13}$$

As a matter of fact, a model should be now imperatively defined by a couple of *mesurement units* $\mathbf{u}$ and *a probabilistic distribution* $\mathbf{m}$. We will denote it by $\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}$.

3.3 Properties of models $\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}$

Interpretation and identifiability From the standard probability theory again, we know that there exists another probabilistic model index, designated here by the (very) shortcut notation $\mathbf{u}^{-1}(\mathbf{m})$, such that

$$\mathbf{u}^{-1}(\mathbf{m}) \in \{\mathbf{m}' \in \mathbb{M} : \mathrm{p}_{\mathbf{m}'}^{\mathbf{id}} = \mathrm{p}_{\mathbf{m}}^{\mathbf{u}}\}, \tag{14}$$

this set being usually restricted to a single element. In other words, it means that there exists *two alternative interpretations* of strictly the same model:

- $\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}$: data measured with unit $\mathbf{u}$ arise from the probabilistic distribution indexed by $\mathbf{m}$;
- $\mathrm{p}_{\mathbf{u}^{-1}(\mathbf{m})}^{\mathbf{id}}$: data measured with unit $\mathbf{id}$ arise from the probabilistic distribution indexed by $\mathbf{u}^{-1}(\mathbf{m})$.

From the statistician point of view, it means that the decomposition of a model into unit *vs.* distribution is *not identifiable*. From the practitioner point of view, it means that he has the freedom to choose the interpretation which is the most meaningful for him.

Opportunity for designing new models Expressing a model as a combination of a unit and a distribution can be in practice a great opportunity for the statistician to build easily numerous new but meaningful models. Indeed, The Cartesian product of, say, a set of $N_{\mathbf{u}_j}$ *standard* units $\{\mathbf{u}_j\}$ for each variable j and a set of $N_{\mathbf{m}}$ *standard* distribution families $\{\mathbf{m}\}$ straightforwardly leads to potentially $\prod_j N_{\mathbf{u}_j} \times N_{\mathbf{m}}$ different models $\{\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}\}$. The new family $\{\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}\}$ can even be huge, since it could involve some combinatorics (we will discuss about combinatorics in the discussion of Section 6). However, it should be underlined again that all these by-product models are positively meaningful since their interpretation is directly related to the interpretation of both the $\mathbf{u}_j$'s and $\mathbf{m}$'s, themselves meaningful since standard.

Besides the latter model building process, note that there may exist some situations where $\mathbf{m} = \mathbf{u}(\mathbf{m})$, meaning that some probabilistic distribution $\mathbf{m}$ is invariant to the unit transformation $\mathbf{u}$. It is for instance the case with the scaling and shifting unit transformation (10) associated to the multivariate

Gaussian mixture. Such a situation will be studied in detail in Section 4.1. Invariance is also verified for the (unconstrained) latent class model with categorical variables (see Section 2) for any permutation transformation unit.

Model selection Among the set of proposed models by combining units and distributions, the practitioner can let the statistician choose one of them by a model selection principle like the BIC or the ICL criteria, respectively defined in (2) and (3). Nevertheless, both criteria relying on the log-likelihood, it is required to compute all of them with the same unit reference, say **id**. It means that when using any model $\mathrm{p}^{\mathbf{u}}_{\mathbf{m}}$, the associated log-likelihood involved in BIC or ICL *has to* be systematically converted into this one associated to the equivalent model $\mathrm{p}^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})}$. For instance, for absolutely continuous variables $\mathbf{x}$ and a differentiable unit $\mathbf{u}$, the likelihood model conversion from $\mathrm{p}^{\mathbf{u}}_{\mathbf{m}}$ to $\mathrm{p}^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})}$ can be explicitly obtained from the relationship

$$\mathrm{p}^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})} = \{\cdot \in \mathbb{X}^{\mathbf{id}} \mapsto \mathrm{p}(\mathbf{u}(\cdot);\boldsymbol{\theta}) \;\times\; |\mathbf{J}^{\mathbf{u}}(\cdot)| : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}, \tag{15}$$

with $\mathbf{J}^{\mathbf{u}}(\cdot)$ the Jacobian associated to the transformation $\mathbf{u}$. We will note in the following $\mathrm{BIC}^{\mathbf{u}}_{\mathbf{m}}$ and $\mathrm{ICL}^{\mathbf{u}}_{\mathbf{m}}$ for criteria values BIC and ICL computed with the model $\mathrm{p}^{\mathbf{u}}_{\mathbf{m}}$.

About more complex units A much more complex transformation unit is proposed in [61] for the multivariate Gaussian mixture situation. Its fundamental purpose is to approach class normality, in order to match as possible with the Gaussian hypothesis by clusters. Consequently, this unit depends both on classes and on variables. In addition, it is also parameterized, its unit parameter ($\boldsymbol{\lambda}$) having to be estimated by an EM algorithm simultaneously with the mixture parameter ($\boldsymbol{\theta}$). More precisely, it corresponds to the Manly transformation unit [31] $\mathbf{u}_{\boldsymbol{\lambda}} = \{\mathbf{u}_{\lambda_{kj}}\}$ defined by

$$\mathbf{u}_{\lambda_{kj}}(x_{ij}) = \begin{cases} \dfrac{\exp(\lambda_{kj}x_{ij}) - 1}{\lambda_{kj}}, & \lambda_{kj} \neq 0 \\ x_{ij}, & \lambda_{kj} = 0 \end{cases} \tag{16}$$

where $\boldsymbol{\lambda} = \{\lambda_{kj}\}$ gathers the unit parameters ($\lambda_{kj} \in \mathbb{R}$). Technically, models where λ_{kj} stands, among $\mathbb{R}\backslash\{0\}$ and $\{0\}$ are selected by a BIC criterion, through a forward and backward algorithm for avoiding combinatorial difficulties when the involved dimension d grows.

It is clear that the work of [61] produces very high flexible mixtures. However, it acts as a good "technical" transformation unit rather than a meaningful practitioner unit for two reasons. Firstly, the Manly transformation was originally designed for turning skew unimodal distributions into nearly symmetric normal distributions, thus conveying no particular interpretation as a "human" or "physical" unit does. Secondly, its class dependent transformation seems to be inconsistent with the traditional definition that any practitioner should expect. Conversely, [61] defends invariance on the estimated partition resulting from any scaling or shifting transformation of Manly as a desirable property, whereas in the present work we argue that a non-invariant transformation unit could be instead an opportunity for enlarging the model family.

4 Real data sets experiments: the clustering case

4.1 Scaling units with parsimonious Gaussian models

Gaussian models are probably the most frequent mixture distributions and scaling transformations (see (10) with $b_j = 0$) are probably the most current unit transformations. We have already discussed in Section 3.2 that *general* Gaussian models are invariant to any scaling transformation. However, it is not always the case for some classical *constrained* Gaussian models. We propose to identify them and to illustrate that it could be an opportunity for enlarging easily the whole Gaussian mixture family on a real example.

Eigenvalue decomposition models (EIG) Initiated by [3], each covariance matrix $\boldsymbol{\Sigma}_k$ can be decomposed as $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{S}_k \boldsymbol{\Lambda}_k \mathbf{S}'_k$ where: (*i*) $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/d}$ is the volume of the class k, (*ii*) $\mathbf{S}_k$ is an orthogonal matrix the columns of which are the eigenvectors of $\boldsymbol{\Sigma}_k$ and corresponds to the orientation of the class k, and (*iii*) $\boldsymbol{\Lambda}_k$ is a diagonal positive definite matrix with determinant 1 and with diagonal coefficients in decreasing order, corresponding to the shape of the class k. A combination of parsimonious hypotheses on λ_k, $\mathbf{S}_k$ and $\boldsymbol{\Lambda}_k$ parameters provided by [12], allowing volume, shape or orientations to vary or not between components, and also including two parsimonious families which are the so-called spherical and diagonal models associated respectively to the identity matrix for $\boldsymbol{\Lambda}_k$ and to a permutation matrix for $\mathbf{S}_k$, leads to 14 different models and we will now refer to them by the name EIG as in [8]. For instance, the so-denoted $[\lambda \mathbf{S}_k \boldsymbol{\Lambda} \mathbf{S}'_k]$ EIG model assumes that the Gaussian components have identical shapes, same volumes and free orientations.

Statistical models (RTV) [8] proposes the statistical sensible decomposition $\boldsymbol{\Sigma}_k = \mathbf{T}_k \mathbf{R}_k \mathbf{T}_k$ and $\boldsymbol{\mu}_k = \mathbf{T}_k \mathbf{V}_k$, on $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ (centers) respectively, where $\mathbf{T}_k$ is the corresponding diagonal matrix of conditional standard deviations, where $\mathbf{R}_k$ is the associated matrix of conditional correlations and where $\mathbf{V}_k$ gathers standardized means. It is possible to combine meaningful constraints on $\mathbf{T}_k$ (free, isotropic ($\forall\ k : \mathbf{T}_k = a_k \mathbf{T}_1$ where $a_k > 0$) or homogeneous ($\mathbf{T}_k = \mathbf{T}$)), on $\mathbf{R}_k$ (free or homogeneous ($\mathbf{R}_k = \mathbf{R}$)) and on centers $\boldsymbol{\mu}_k$ (vectors $\mathbf{V}_k = \mathbf{T}_k^{-1} \boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$ are free or homogeneous ($\mathbf{V}_k = \mathbf{V}$)). It allows to obtain 11 parsimonious models that are straightforwardly denoted by $[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}_k]$, $[\mathbf{R}, \mathbf{T}, \mathbf{V}_k]$, $[\mathbf{R}_k, a_k \mathbf{T}, \mathbf{V}_k]$...

Scale unit-dependency illustration Considering the very current scale unit transformation (10), corresponding to standard units (mm, cm, standardized...), [8] listed models in each family where invariance holds: 8 EIG among 14 whereas all 11 RTV.

We consider $n = 272$ eruptions of the famous Old Faithful geyser, described by two variables ($d = 2$): Duration (of an eruption) and Waiting (to the next eruption) both measured in minutes, so $\mathbf{id} = (\text{min}, \text{min})$ with notations of the previous sections. This sample from [57] has been subject to many clustering studies (see [2] for an example) and a widespread structure of Old Faithful eruptions in the literature consists of two clusters (often interpreted as short and long eruptions) thus, the number of groups is fixed to $K = 2$ through this numerical illustration. Each model $\mathbf{m}$ of the two families (EIG and RTV), with free mixing proportions π_k, has been inferred on this data set with three different units $\mathbf{u}$: the original units $\mathbf{id}$, new units $\mathbf{u}^{\text{scale}_1}$ for Duration in seconds (noted also sec) while Waiting is unchanged, and standardized (noted also stand) units $\mathbf{u}^{\text{scale}_2}$ for both Duration and Waiting. The following packages are used: the `Rmixmod` R package [27] for EIG and the `mixrtv` Matlab package [8] for RTV.

Respectively to each situation, Table 1 displays the best model $\text{p}^{\mathbf{u}}_{\mathbf{m}}$ of each family according to the $\text{ICL}^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})}$ criterion value (thus ICL is expressed with the unit reference $\mathbf{id}$ for the model $\text{p}^{\mathbf{u}}_{\mathbf{m}}$). We retrieve the fact, demonstrated in [8], that RTV models are invariant to scaling units whereas EIG models $[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$ and $[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$ are not. We observe here that this unit dependency of some EIG models may be an opportunity to build new well suitable while meaningful models. In particular, the ICL criterion proposes to retain the model $\text{p}^{\mathbf{u}^{\text{scale}_1}}_{[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']}$ instead of any other standard model of the EIG family associated to the $\mathbf{id}$ units. This fact illustrates the cheap and meaningful enlargement of the EIG family provided by the unit combination.

Table 1 *The best model within each family (EIG and RTV), inferred on the Old Faithful data ($K = 2$) when measurement units $\mathbf{u} \in \{\mathbf{id}, \mathbf{u}^{scale_1}, \mathbf{u}^{scale_2}\}$ of the couple (Duration, Waiting) vary. For each unit case, the corresponding ICL value is expressed with the $\mathbf{id} = (min, min)$ unit, noted below $ICL^{\mathbf{id}}$ as a shortcut for $ICL^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})}$.*

	$\mathbf{id} = (\text{min}, \text{min})$		$\mathbf{u}^{\text{scale}_1} = (\text{sec}, \text{min})$		$\mathbf{u}^{\text{scale}_2} = (\text{stand}, \text{stand})$	
family	$\mathbf{m}$	$\text{ICL}^{\mathbf{id}}$	$\mathbf{m}$	$\text{ICL}^{\mathbf{id}}$	$\mathbf{m}$	$\text{ICL}^{\mathbf{id}}$
EIG	$[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	1 160.3	$[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	1 158.7	$[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$	1 160.3
RTV	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1 158.8	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1 158.8	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1 158.8

4.2 Non-scaling units for continuous data in a mixed data case

Authors in [20] (see also [35] p. 139–142) considered the clustering of patients on the basis of petrial variates alone for the prostate cancer clinical trial data of [10] which is reproduced in [1] p. 261–274. This data set was obtained from a randomized clinical trial comparing four treatments for $n = 506$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease. As reported by [10], Stage 3 represents local extension of the disease without evidence of distance metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, X-ray evidence, or both. Twelve pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases). There are 62 missing values, so about 1% of the whole sample, and 475 patients have finally no missing data.

The skewed variables “size of primary tumour” (denoted below as SZ) and “serum prostatic acid phosphatase” (denoted below as AP) were considered with two historical units for performing the clustering task. On the one hand, [39] use initial raw units, denoted by **id**. On the other hand, [23] proposed using a square root and a logarithm transformation, respectively $\mathbf{u}_{\text{SZ}} = \sqrt{\cdot}$ and $\mathbf{u}_{\text{AP}} = \ln(\cdot)$, since both SZ and AP are skewed. Other variable unit than SZ and AP of **u** is unchanged compared to **id**.

We propose now to infer which unit, among **id** and **u**, can be retained when involving the model selection principle in model-based clustering. The model **m** we consider is the latent class model for mixed data given by (4), while dealing directly with missing data through the MAR mechanism (1). The model **m** includes also the number of clusters $K \in \{1, \ldots, 4\}$. Estimation is performed through the MixtComp software[2].

Results with the ICL criterion are displayed on Figure 1. It advocates in favour of the transformed units **u** while retaining also two clusters. Note that the raw unit **id** clearly fails to select this number of clusters. Exploring now both two-clusters partitions, Table 2 also indicates that partition obtained with **u** is more correlated to the medical partition (Stage 3 and Stage 4) than the partition obtained with **id** is.

Table 2 Missclassification detail for the prostate data set with the two competitor units **id** and **u** ($K = 2$).

	Estimated $\hat{\mathbf{z}}$ with **id**		Estimated $\hat{\mathbf{z}}$ with **u**	
Medical **z**	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Stage 3	287	5	270	22
Stage 4	52	162	23	191
Missclassified	11%		9%	

[2] MixtComp is a clustering software developped by Biernacki C., Iovleff I. and Kubicki V. and freely available on the MASSICCC web platform https://modal-research-dev.lille.inria.fr/#/

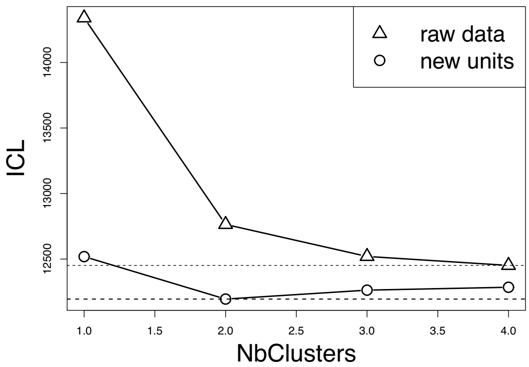


Fig. 1 Selecting both the units and the number of clusters by the ICL criterion on the prostate data set.

4.3 Shifted and scaled count data in mixed data case

We consider the German health registry of the R dataset `rwm1984COUNT` for the year 1984, provided by [46] and studied in [19]. It is composed by $n = 3\ 874$ patients that spent time into German hospitals during year 1984 and patients are described through eleven mixed variables of different kinds. Clustering can be performed by the latent class principle described in (4), involving univariate Gaussian, Poisson and multinomial distributions according to the type of variable. All details on this data set and implied distributions are displayed in Table 3. The MixtComp software can still be used for performing related estimation.

Table 3 Variable types and associated univariate distributions for the dataset `rwm1984COUNT`.

	Variables	Type	Distribution
1	number of visits to doctor during year	count	Poisson
2	number of days in hospital	count	Poisson
3	educational level	categorical	multinomial
4	age	count	Poisson
5	outwork	binary	Bernoulli
6	gender	binary	Bernoulli
7	matrimonial status	binary	Bernoulli
8	kids	binary	Bernoulli
9	household yearly income	continous	Gaussian
10	years of education	count	Poisson
11	self employed	binary	Bernoulli

Four unit systems are sequentially considered (among shifted and scaled units; see Section 3.1) differing over the count data, described as follows:

- **id**: original units for all variables;

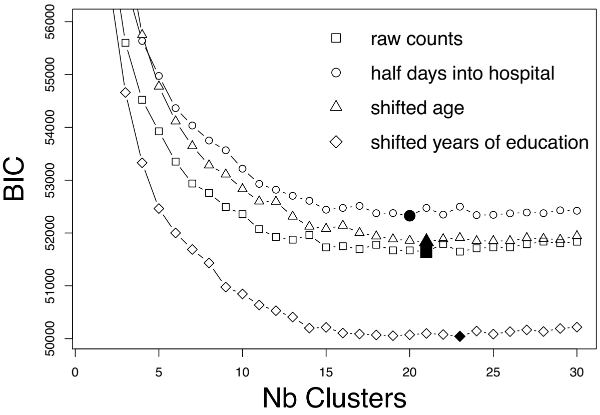


Fig. 2 BIC values for different combinations of units and number of clusters for the German health registry data set.

- $\mathbf{u}^1$: the time spent into hospital (variable number 2 in Table 3) is counted in half days instead of days, other variables being with the original units;
- $\mathbf{u}^2$: the minimum of the age series (variable number 4 in Table 3) is deduced from all ages leading to shifted ages, other variables being with the original units;
- $\mathbf{u}^3$: the minimum of years of education (variable number 10 in Table 3) is deduced from the series leading to shifted years of education.

Then, a BIC criterion is used for selecting the model $\mathrm{p}^{\mathbf{id}}_{\mathbf{u}^{-1}(\mathbf{m})}$, with $\mathbf{u} \in \{\mathbf{id}, \mathbf{u}^1, \mathbf{u}^2, \mathbf{u}^3\}$ and $\mathbf{m}$ being the latent class model with $K \in \{1, \ldots, 30\}$. Figure 2 displays the BIC values for each model. It selects 23 clusters obtained under the shifted years of education units ($\mathbf{u}^3$). Since $\mathbf{u}^3$ provides a high BIC value improvement in comparison to other units, the statistician could be motivated by showing the resulting partition to the practitioner for seeking a potential interesting meaning. Note again that interpretation of model $\mathrm{p}^{\mathbf{u}^3}_{\mathbf{m}}$ is particularly simple by its decomposition into classical units $\mathbf{u}^3$ *vs.* classical model $\mathbf{m}$.

4.4 Domain specific transformation for RNA-seq count data

We consider a sample of RNA-seq gene expressions arising from the rat count table[3], composed by 30 000 genes described by 22 counting descriptors. Genes with low expression are removed with a classical technique of the domain, leading finally to 6 173 genes. Two different processes are involved for dealing with such data:

- **Standard process**: [48] use the initial count units (**id**) and use a Poisson mixture ($\mathbf{m}^1$);

[3] http://bowtie-bio.sourceforge.net/recount/

Table 4 BIC value associated to two combinations of different models and units for the RNA-seq gene expressions data set.

Model	Units	BIC
Poisson ($\mathbf{m}^1$)	raw unit (**id**)	2 615 654
Gaussian ($\mathbf{m}^2$)	transformed ($\mathbf{u}$)	909 190

- **"RNA-seq unit"**: being motivated by genetic considerations, [14] use the transformation $\mathbf{u}_j(x_{ij}) = \ln(\text{scaled normalization}(x_{ij}))$ (see details about the scaling in [14]; note that other scalings exist also in [47]) over which a Gaussian mixture is used ($\mathbf{m}^2$).

As in [14], we compare both models $\mathrm{p}^{\mathbf{id}}_{\mathbf{m}^1}$ and $\mathrm{p}^{\mathbf{u}}_{\mathbf{m}^2}$, with $K = 30$ clusters. The BIC value associated to both is displayed in Table 4. We retrieve here results of [14] since the model $\mathrm{p}^{\mathbf{u}}_{\mathbf{m}^2}$ is clearly retained in our case also, demonstrated once more that a practitioner driven unit, specific to the data domain, may highly contribute to improve the modelling task.

5 Real data sets experiments: the co-clustering case

5.1 A binary data set with *natural* initial units

We consider the SPAM e-mail database[4] which gathers $n = 4\ 601$ e-mails composed by 1 813 "spams" and 2 788 "good e-mails". Each e-mail is described by $d = 54$ continuous descriptors (three other continuous descriptors exist but we do not use them), where 48 of them are percentages that a given *word* appears in an e-mail ("make", "you'. . .) and the remaining 6 are percentages that a given *char* appears in an e-mail (";", "$". . .). In fact, we do not use this initial data set but a transformation of these continuous descriptors into binary descriptors by noting $x_{ij} = 1$ if word/char j appears in e-mail i, $x_{ij} = 0$ otherwise. This data set is displayed on the left of both Figure 3 (a) and (b). For each variable j, two different units are possible :

- $\mathbf{id}_j$: it corresponds to the previous coding;
- $\mathbf{u}_j(\cdot) = 1 - (\cdot)$: it corresponds to the reverse coding where $x_{ij} = 0$ if word/char j appears in e-mail i, $x_{ij} = 1$ otherwise.

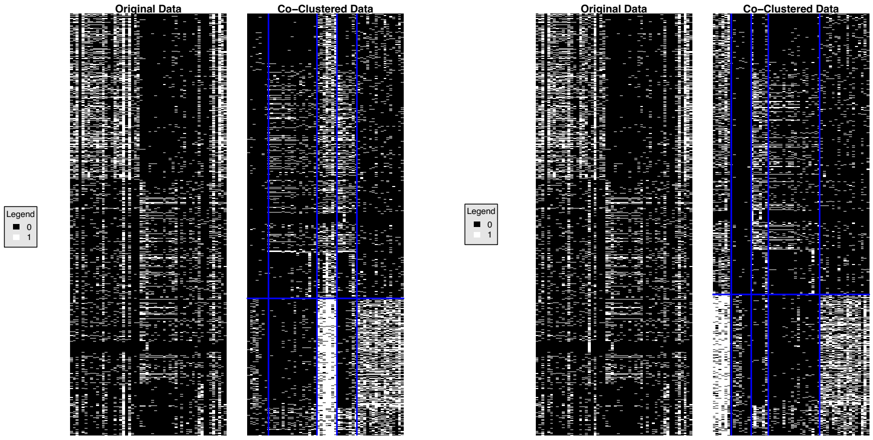
The whole coding is denoted by $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and 2^{54} possibilities of different units for the whole data set exist.

We perform now a co-clustering process (see Section 2.2) essentially for retrieving the initial partition of individuals into "spams" and "good e-mails", thus the partition of variables is just a technical way for reducing the dimension. We fix two individual clusters ($K = 2$) and five variable clusters ($L = 5$), only units $\mathbf{u}$ being to be estimated. Since the number of available units is particularly high and can not be exhaustively browsed, we use a "naive" algorithm

[4] `https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/`

which estimates the co-clustering model for 100 000 randomly chosen different units, and we retain finally the unit **u** providing the best ICL value. We involve the R BlockCluster package [4], available on the CRAN, for performing these estimations.

The co-clustered data set we obtained, associated to the initial unit **id** and also to the unit **u** associated to the best ICL we found, are displayed on the right of Figure 3 (a) and (b), respectively. In both cases, the ICL value and also the empirical error rate (of individuals) are provided. We observe that the best **u** we estimated leads to very slightly better ICL than the initial unit **id** and, moreover, it does not improve the partitioning result. If we have a closer look at these results, we observe that the best unit **u** consists just in recoding one variable ($j = 19$: "you"), all other corresponding to the initial coding. Globally, it thus indicates that the initial coding should be preferred. This conclusion makes sense with the fact that the initial coding acts as a *natural* coding for the practitioner ("appears", "does not appear"): each variable having the same meaning, it can be expected by the practitioner to code them in an identical way. Our proposal to change units just for some variables was quite artificial and technical but it was somewhat reassuring that the ICL criterion tends to refuse it.



(a) Initial units **id**: ICL=92 682.54 and error rate of lines is 0.1984.

(b) Best estimated units **u**: ICL=92 524.57 and error rate of lines is 0.2008.

Fig. 3 Initial data set and co-clustered data set for the initial units **id** and the best estimated units **u** by the ICL criterion for the SPAM e-mail database ($K = 2$ and $L = 5$ are fixed).

Table 5 Variable meaning for the congressional voting records data set.

1. handicapped-infants
2. water-project-cost-sharing
3. adoption-of-the-budget-resolution
4. physician-fee-freeze
5. el-salvador-aid
6. religious-groups-in-schools
7. anti-satellite-test-ban
8. aid-to-nicaraguan-contras
9. mx-missile
10. immigration
11. synfuels-corporation-cutback
12. education-spending
13. superfund-right-to-sue
14. crime
15. duty-free-exports
16. export-administration-act-south-africa

5.2 A categorical data set with *artificial* initial units

We consider now a categorical data set with three levels by variable (non-binary variables). It corresponds to the congressional voting records data set[5] composed by votes for each of the $n = 435$ U.S. House of Representatives Congressmen. Congressmen are divided into two classes (267 are democrats, 168 are republicans), and $d = 16$ votes with $m = 3$ modalities each [50] are provided with the following meaning:

- "yea": voted for, paired for, and announced for;
- "nay": voted against, paired against, and announced against;
- "?": voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known.

Details on variables are displayed on Table 5. This data set is displayed on the left of Figure 4 (a) and (b).

Contrary to the previous data set (spams), levels "yea" and "nea" are arbitrarily coded *between variables*. Indeed, there is no longer a common level reference having a common meaning like "appears" or "does not appear". For instance, variable 3 (see Table 5) could be arbitrarily coded as well with

3. **adoption**-of-the-budget-resolution = "**yes**" ⇔ 3. **rejection**-of-the-budget-resolution = "**no**",

independently of all other variables. However, level "?" has the same meaning for all variables are thus it should be similarly coded for all variables for avoiding the artificial coding effect we have previously seen for spams. Consequently, it is possible to consider the following different units for each variable j:

- $\mathbf{id}_j$: it corresponds to the initial units of the data set

$$x_{ij} = \begin{cases} (1,0,0) \text{ if voted "yea" to vote } j \text{ by congressman } i \\ (0,1,0) \text{ if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1) \text{ if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

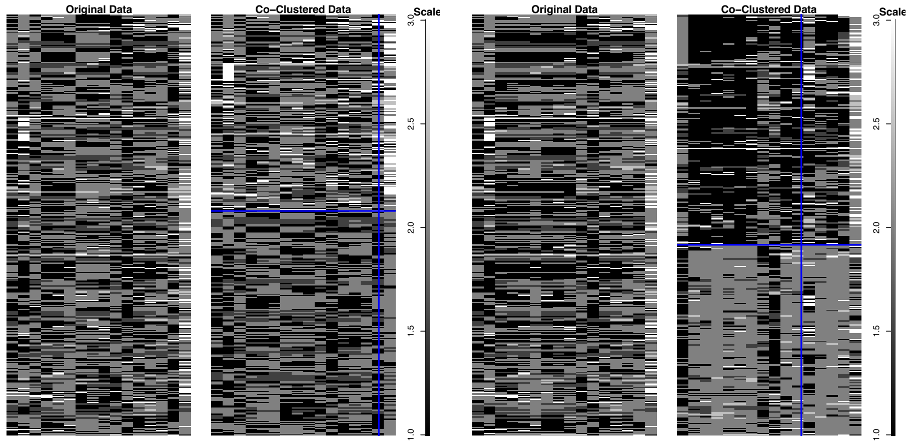
- $\mathbf{u}_j$: it corresponds to reversing the coding *only for "yea" and "nea"*

$$\mathbf{u}_j(x_{ij}) = \begin{cases} (\mathbf{0},\mathbf{1},\mathbf{0}) \text{ if voted "yea" to vote } j \text{ by congressman } i \\ (\mathbf{1},\mathbf{0},\mathbf{0}) \text{ if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1)\ \text{ if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

[5] http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

We perform now a co-clustering process (see Section 2.2) essentially for retrieving the initial partition of individuals into a political party. We fix two individual clusters ($K = 2$) and also two variable clusters ($L = 2$), only units $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ being to be selected. Since the number of available units is not too high here, a comprehensive search to find the best **u** by ICL ($2^{16} = 65\,536$ cases) can be performed with the BlockCluster package. The co-clustered data set we obtained, associated to the initial unit **id** and also to the unit **u** associated to the best ICL value, are displayed on the right of Figure 3 (a) and (b), respectively. In both cases, the ICL value and also the empirical error rate (of individuals) are provided. We observe that the best **u** we estimated leads this time simultaneously to a good improvement of the ICL value (compared to the initial unit **id**) and of the partitioning result. This improvement is wholly corroborated with a view at the co-clustered data set associated to the **id** unit at the right of Figure 4 (a) and at the co-clustered data set associated to the best **u** unit at the right of Figure 4 (b). If we have also a closer look at these results, we observe that the best unit **u** consists in recoding the following five variables: 3. adoption-of-the-budget-resolution, 7. anti-satellite-test-ban, 9. aid-to-nicaraguan-contras, 10. mx-missile, 16. duty-free-exports.

As a concluding remark on this data set, here initial units **id** where *artificially* fixed by the practitioner thus it could make sense to change, and selects, units.



(a) Initial units **id**: ICL=5 916.13 and error rate of lines is 0.1984.

(b) Best estimated units **u**: ICL=5 458.156 and error rate of lines is 0.1034.

Fig. 4 Initial data set and co-clustered data set for the initial units **id** and the best estimated units **u** by the ICL criterion for the congressional voting records data set ($K = 2$ and $L = 2$ are fixed).

6 Concluding remarks

This work aims to alert on the fact that interpretation of ("classical") models is usually unit dependent. From this point of view, models should thus be revisited as a couple (units,models) and it can be an opportunity for cheap, wide and meaningful enlarging of "classical" model families. Several attempts on this topic already exist in literature but this paper provides a formalization of this principle while extending it to non-continuous types of data. We focus on clustering and co-clustering but the idea is extensively valid for other statistical tasks.

Beyond this potentially attractive unification of measurements units and classical models, we have to be aware that some units could be practitioner meaningful and in that case the statistician should restrict his "technical model enlarging" to such a constraint. In counterpart, combinatorial problems may occur since the new model family can be potentially huge trough the unit combination. This problem should be specifically addressed in future works. In addition, possibility to parametrize measurement unit transformation, similarly to the example provided by [61], should be another good opportunity for enlarging the possible units and, consequently, the resulting models.

References

1. Andrews, D.F., Herzberg, A.M.: Data: A Collection of Problems from Many. Fields for the Student and Research Worker. Springer-Verlag (1985)
2. Atkinson, A., Riani, M.: Exploratory tools for clustering multivariate data. Computational Statistics and Data Analysis **52**(1), 272–285 (2007)
3. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)
4. Bhatia, P., Iovleff, S., Govaert, G.: Blockcluster: An r package for model based co-clustering. Journal of Statistical Software (in press) (2015)
5. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(7), 719–725 (2000)
6. Biernacki, C., Jacques, J.: A generative model for rank data based on insertion sort algorithm. Computational Statistics and Data Analysis **58**, 162–176 (2013)
7. Biernacki, C., Jacques, J.: Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. Statistics and Computing **26**(5), 929–943 (2016). URL https://hal.inria.fr/hal-01052447
8. Biernacki, C., Lourme, A.: Stable and visualizable gaussian parsimonious clustering models. Statistics and Computing **24**(6), 953–969 (2014)
9. Bock, H.: Statistical Testing and Evaluation Methods in Cluster Analysis. In: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions, pp. 116–146. Calcutta (1981)
10. Byar, D., Green, S.: The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. Bulletin du Cancer **67**, 477–490 (1980)
11. Celeux, G., Diebolt, J.: The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly **2**(1), 73–92 (1985)
12. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognition **28**(5), 781–793 (1995)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data (with discussion). Journal of the Royal Statistical Society, Series B **39**, 1–38 (1977)

14. Gallopin, M., Rau, A., Celeux, G., Jaffrézic, F.: Transformation des données et comparaison de modèles pour la classification des données rna-seq. 47èmes Journées de Statistique de la SFdS (2015)
15. Ghahramani, Z., Hinton, G.: The em algorithm for factor analyzers. Tech. rep., University of Toronto (1997)
16. Goodman, L.A.: Exploratory latent structure models using both identifiable and unidentifiable models. Biometrika **61**, 215–231 (1974)
17. Govaert, G.: Data Analysis. ISTE-Wiley (2009). URL https://hal.archives-ouvertes.fr/hal-00447855
18. Govaert, G., Nadif, M.: Co-Clustering. Wiley (2013)
19. Hilbe, J.M.: Modeling count data. Cambridge University Press (2014)
20. Hunt, L., Jorgensen, M.: Mixture model clustering: a brief introduction to the multimix program. Australian and New Zealand Journal of Statistics **41**(2), 153–171 (1999)
21. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters **31**, 651–666 (2010)
22. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, New Jersey (1988)
23. Jorgensen, M., Hunt, L.: Mixture model clustering of data sets with categorical and continuous variables. In: Proceedings of the Conference ISIS, pp. 375–384 (1996)
24. Keribin, C., Brault, V., Celeux, G., Govaert, G.: Estimation and selection for the latent block model on categorical data. Statistics and Computing **25**(6), 1201–1216 (2015). DOI 10.1007/s11222-014-9472-2
25. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of measurement (additive and polynomial representations), vol. 1. Academic Press, New York (1971)
26. Law, M.H., Figueiredo, M.A.T., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1154–1166 (2004)
27. Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., Govaert, G.: Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. Journal of Statistical Software **in press** (2015)
28. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edition edn. Wiley (2002)
29. Lomet, A., Govaert, G., Grandvalet, Y.: Model selection in block clustering by the integrated classification likelihood. In: 20th International Conference on Computational Statistics (COMPSTAT 2012), pp. 519–530. Lymassol, France (2012). URL https://hal.archives-ouvertes.fr/hal-00730829
30. Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A.: Foundations of measurement, vol. 3. Academic Press, New York (1990)
31. Manly, B.F.: Exponential data transformations. Statistician **25**(1), 37–42 (1976)
32. Marbac, M., Sedki, M.: Variable selection for model-based clustering using the integrated complete-data likelihood. arXiv:1501.06314 (2015)
33. Maugis, C., Celeux, G., Martin-Magniette, M.: Variable selection for clustering with Gaussian mixture models. Biometrics **65**(3), 701–709 (2009)
34. Maugis, C., Celeux, G., Martin-Magniette, M.L.: Variable selection in model-based clustering: A general variable role modeling. Computational Statistics and Data Analysis **53**, 3872–3882 (2009)
35. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New-York (2000)
36. McLachlan, G., Peel, D.: Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis (41), 379–388 (2003)
37. McNicholas, P., Murphy, T.: Model-based clustering of microarray expression data via latent gaussian mixture models. Bioinformatics **21**(26), 2705–2712 (2010)
38. McNicholas, P.D.: Mixture Model-Based Classification. Chapman and Hall, New York (2016)
39. McParland, D., Gormley, I.C.: Model based clustering for mixed data: clustmd. Advances in Data Analysis and Classification **10**(2), 155–169 (2016)
40. Melnykov, V., Maitra, R.: Finite mixture models and model-based clustering. Statist. Surv. **4**, 80–116 (2010). DOI 10.1214/09-SS053. URL http://dx.doi.org/10.1214/09-SS053

41. Meynet, C.: Sélection de variables pour la classification non supervisée en grande dimension. Ph.D. thesis, Université Paris-Sud 11 (2012)
42. Meynet, C., Maugis-Rabusseau, C.: A sparse variable selection procedure in model-based clustering. Research report (2012). URL http://hal.inria.fr/hal-00734316
43. Moustaki, I., Papageorgiou, I.: Latent class models for mixed variables with applications in archaeometry. Computational Statistics and Data Analysis **48**(3), 65—675 (2005)
44. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research **8**, 1145–1164 (2007)
45. Raftery, A.E., Dean, N.: Variable Selection for Model-Based Clustering. Journal of the American Statistical Association **101**(473), 168–178 (2006)
46. Rao C. R., M.J.P., Rao, D.C.: Handbook of statistics: epidemiology and medical statistics, vol. 27. Elsevier (2007)
47. Rau, A., Maugis-Rabusseau, C.: Transformation and model choice for rna-seq co-expression analysis. bioRxiv (2016). DOI 10.1101/065607. URL http://biorxiv.org/content/early/2016/10/21/065607
48. Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.L., Celeux, G.: Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. Bioinformatics **31**(9), 1420–1427 (2015)
49. Redner, R., Walker, H.: Mixture densities, maximum likelihood and the em algorithm. SIAM Review **26**(2), 195–239 (1984)
50. Schlimmer, J.C.: Concept acquisition through representational adjustment. Ph.D. thesis, Department of Information and Computer Science, University of California, Irvine, CA (1987)
51. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**, 461–464 (1978)
52. Seber, G.A.F., Lee, A.J.: Linear Regression Analysis, second edition edn. John Wiley & Sons, New Jersey (2012)
53. Sedki, M., Celeux, G., Maugis-Rabusseau, C.: SelvarMix: A R package for variable selection in model-based clustering and discriminant analysis with a regularization approach. Research report (2014). URL https://hal.inria.fr/hal-01053784
54. Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A.: Foundations of measurement, vol. 2. Academic Press, New York (1989)
55. Tadesse, M.G., Sha, N., Vannucci, M.: Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association **100**(470), 602–617 (2005)
56. Thomas, I., Frankhauser, P., Biernacki, C.: The morphology of built-up landscapes in wallonia (belgium): a classification using fractal indices. Landscape and Urban Planning **84**, 99–115 (2008)
57. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4 edn. Springer, New York (2002)
58. Wolfe, J.H.: A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, US Naval Personnel Research Activity, San Diego, California (1971)
59. Yeung, K., Fraley, C., Murua, A., Raftery, A., Ruzzo, W.: Model-based clustering and data transformations for gene expression data. Bioinformatics **17**(10), 977–987 (2001)
60. Zhou, H., Pan, W., Shen, X.: Penalized model-based clustering with unconstrained covariance matrices. Electronic Journal of Statistics **3**, 1473–1496 (2009)
61. Zhu, X., Melnykov, V.: Manly transformation in finite mixture modeling. Computational Statistics and Data Analysis **in press** (2016)