CS-BIGS 4(2): 73-82 http://legacy.bentley.edu/csbigs/documents/biernacki.pdf

Simultaneous t-Model-Based Clustering for Time Dependent Data: Application to a Study of the Financial Health of Corporations

Christophe Biernacki
Université Lille 1, CNRS & INRIA, France

Alexandre Lourme
Université de Pau et des Pays de l'Adour, IUT de Génie Biologique, France

Student's t mixture model-based clustering is often used as a robust alternative to the Gaussian model-based clustering. In this paper, we aim to cluster several different datasets at the same time, instead of a single one as is common, in a context where underlying t-populations are not completely unrelated: All individuals are described by the same features and partitions of identical meaning are expected. Justifying from some natural arguments a stochastic linear link between the components of the mixtures associated to each dataset, we propose some parsimonious and meaningful models for a so-called simultaneous clustering method. Maximum likelihood mixture parameters, subject to the linear link constraint, can be easily estimated by a GEM algorithm that we describe. We then propose to apply these models to two financial company time-dependent data sets, consisting of both healthy and bankrupt companies. Our new models point out that the hidden structure could be more complex than generally expected, distinguishing three groups: not only two clear groups of healthy and bankrupt companies but also a third one representing companies with unpredictable health.

Keywords: Stochastic linear link, t-mixture, model-based clustering, EM algorithm, model selection, company failure.

1. Introduction

Clustering aims to split a sample into classes in order to reveal a hidden but meaningful structure in a dataset. In a probabilistic context it is standard practice to suppose that the data arise from a mixture of parametric distributions and to draw a partition by assigning each data point to the prevailing component (see McLachlan and Peel, 2000, for a review). In particular, in the multivariate continuous situation, t-mixture modeling is usually seen as a robust alternative to the Gaussian modeling (Archambeau and Verleysen, 2007) since it is frequently the case that real data have heavier tails than the normal distribution allows for (Bishop and Svensén, 2005). It has found successful applications in such diverse fields as image registration (Gerogiannis et al. 2009) or letter recognition (Chatzis and Varvarigou, 2008) for example. Nowadays, involving t-models for clustering a given dataset could be considered familiar to every statistician as well as to more and more practitioners.

However, in many situations, one needs to cluster several datasets, possibly arising from different populations (instead of a single one) into partitions with identical meaning. For instance, Lourme and Biernacki (2010) extended the standard Gaussian model-based clustering for simultaneous partitioning of three samples of seabirds living in several geographic zones, leading to very different morphological variables and showed that this model outperforms the naïve approach consisting in performing one independent clustering by sample. The proposed model relies on a linear stochastic link between the samples, which can be justified from some simple but realistic assumptions.

This paper proposes to extend this work to the case of multivariate Student's t models (in-short t-models) in order to simultaneously classify several datasets instead of applying several independent t-clustering methods. Similarly to the Gaussian case (Lourme and Biernacki, 2010), a linear stochastic link between the populations from which the samples arise is argumented and established. This link allows us to estimate, by maximum likelihood (ML), all mixture parameters at the same time and consequently allows us to cluster the diverse datasets simultaneously.

In Section 2, starting from the standard approach of some independent t-mixture model-based clustering methods, we present the principle of simultaneous clustering. Some parsimonious and meaningful models on the established stochastic link are then proposed in Section 3 and associated ML estimates are given in Section 4 through a GEM algorithm. Some experiments are finally performed on a large set of companies described by their financial ratios given over two time periods (mixing of data from 2002 and 2003) in order to build a typology over their financial health (Section 5). Finally in Section 6 we make concluding remarks.

2. From independent to simultaneous t-clustering

In simultaneous clustering, the aim is to separate H samples into K groups. Each sample $x^h (h \in \{1, \dots, H\})$ is composed of n^h individuals $x_i^h (i = 1, \dots, n^h)$ of $\mathbb{R}^d$ and arises from a population P^h. In addition, all populations are described by the same $d$continuous variables and we assume that the underlying partitions of each sample have the same meaning.

2.1.Standard solution: Several independent t-clusterings

In a standard t-model-based clustering framework (see McLachlan and Peel, 2000, Chapter 7), the individuals $x_i^h (i = 1, \dots, n^h)$ of each sample x^h are assumed to be independently drawn from the random vector X^h following a K-order t-mixture P^h with probability density function:

$$f(x; \psi^h) = \sum_{k=1}^{K} \pi_k^h t_d(x; \nu_k^h, \mu_k^h, \Sigma_k^h); x \in \mathbb{R}^d.$$

The coefficients $\pi_k^h (k = 1, \dots, K)$ are mixing proportions (for all k $\pi_k^h > 0$ and $\sum_{k=1}^{K} \pi_k^h = 1$) and $t_d(\bullet; \nu_k^h, \mu_k^h, \Sigma_k^h)$ denotes the d-variate t-distribution with degree of freedom $\nu_k^h \in \mathbb{R}_*^+$, with location parameter $\mu_k^h \in \mathbb{R}^d$ and with inner product matrix $\Sigma_k^h \in \mathbb{R}^{d \times d}$ (positive-definite):

$$t_d(x; \nu_k^h, \mu_k^h, \Sigma_k^h) = \frac{\Gamma(\frac{d+\nu_k^h}{2})(\pi\nu_k^h)^{-d/2}|\Sigma_k^h|^{-1/2}}{\Gamma(\nu_k^h/2)\left[1+(x-\mu_k^h)'{\Sigma_k^h}^{-1}(x-\mu_k^h)/\nu_k^h\right]^{(d+\nu_k^h)/2}}.$$

The mixture P^h is then entirely determined by $\psi^h = (\psi_k^h)_{k=1,\dots,K}$ where $\psi_k^h = (\pi_k^h, \nu_k^h, \mu_k^h, \Sigma_k^h)$.

Two kinds of hidden data can be highlighted in this model. First, a binary vector $z_i^h = (z_{i,1}^h, \dots, z_{i,K}^h)$ indicates whether the data point x_i^h has been generated $(z_{i,k}^h = 1)$ or not $(z_{i,k}^h = 0)$ by the k-th t-component C_k^h of P^h mixture. The vector $z_i^h$is assumed to arise from the K-variate multinomial distribution of order 1 and of parameter $(\pi_1^h, \dots, \pi_K^h)$. Second, if x_i^h has been generated by C_k^h, then it can be assumed equivalently that x_i^harises from the normal d-variate distribution $N_d(\mu_k^h, \Sigma_k^h / u_i^h)$, where $u_i^h \in \mathbb{R}_*^+$ denotes some hidden data arising from the gamma distribution $\gamma_{\nu_k^h/2, \nu_k^h/2}$ (see McLachlan and Peel, 2000, p. 223).

So the complete data model assumes that $(x_i^h, u_i^h, z_i^h)_{i=1,\dots,n^h}$ are realizations of independent random vectors identically distributed according to (X^h, U^h, Z^h) in $\mathbb{R}^d \times \mathbb{R}_*^+ \times \{0,1\}^K$ where $Z^h = (Z_1^h, \dots, Z_K^h)$ is a binary vector from the multinomial distribution with parameter $(\pi_k^h; k = 1, \dots, K)$, $U^h \mid Z_k^h = 1$ is a random variable distributed as $\gamma_{\nu_k^h/2, \nu_k^h/2}$ and $(X^h \mid U^h = u, Z_k^h = 1)$ is normal with mean μ_k^h and covariance matrix Σ_k^h / u.

Estimating $\psi = (\psi^h)_{h=1,\dots,H}$ by maximizing its log-likelihood function $\ln L(\psi; x) = \sum_{h=1}^{H} \sum_{i=1}^{n^h} \ln\left[f(x_i^h; \psi^h)\right] = \sum_{h=1}^{H} \ln L^h(\psi^h; x^h)$ computed on the observed data leads to maximizing independently each log-likelihood function $\ln L^h(\psi^h; x^h)$ of the parameter ψ^h computed on the sample x^h. Several avatars of the EM-algorithm can perform the maximization. Two examples are available in McLachlan and Peel (2000) (p. 224 – 229) and in Wang and Hu (2009).

The observed data point x_i^h is then allocated by the Maximum A Posteriori principle (MAP) to the group corresponding to the highest estimated posterior probability of membership computed at the ML estimate $\hat{\psi}$:

$$t_{i,k}^h(\hat{\psi}) = E(Z_k^h \mid X^h = x_i^h; \hat{\psi}).$$

Since the partition estimated by independent clustering is arbitrarily numbered, the practitioner must, if necessary, renumber some clusters in order to assign the same index to clusters having the same meaning for all populations. The simultaneous clustering method that we now present aims both to improve the partition estimation and to automatically give the same numbering to clusters with identical meaning.

2.2.Proposed solution: Using a linear stochastic link between the populations

From the beginning the groups that have to be discovered consist in a same meaning partition of each sample and samples are described by the same features. In a similar case (but in a Gaussian mixture model-based clustering context), when populations were so related, we proposed in Lourme and Biernacki (2010) to establish a distributional relationship between the components sharing identical labels. We take up here, in a t-model-based clustering context, this idea on which the so-called simultaneous clustering method is based. Then we assume below that the conditional populations are related by a stochastic link and we specify this link thanks to three additional hypotheses H_1, $H_{2,}$, H_3.

For all $(h, h') \in \{1, \dots, H\}^2$ and all $k \in \{1, \dots, K\}$, a map $\xi_k^{h,h'}: \mathbb{R}^d \to \mathbb{R}^d$ is assumed to exist, so that:

$$(X^{h'} \mid Z_k^{h'} = 1) \sim \xi_k^{h,h'}(X^h \mid Z_k^h = 1). \qquad (1)$$

This model implies that individuals from some t-component C_k^h are stochastically transformed (via $\xi_k^{h,h'}$) into individuals of $C_k^{h'}$. In addition, as samples are described by the same features, it is natural, in many practical situations, to expect that a variable in some population depends mainly on the same feature, in another population. So we assume (Hypothesis H_1) that the j-th ($j \in \{1, \dots, d\}$) component $(\xi_k^{h,h'})^{(j)}$ of $\xi_k^{h,h'}$ depends only on the j-th component $x^{(j)}$ of its variable $x \in \mathbb{R}^d$.

In other words, $(\xi_k^{h,h'})^{(j)}$ corresponds to a map from $\mathbb{R}$ into $\mathbb{R}$ that transforms, in distribution, the conditional t-covariate $(X^h \mid Z_k^h = 1)^{(j)}$ into the corresponding conditional t-covariate $(X^{h'} \mid Z_k^{h'} = 1)^{(j)}$. Assuming moreover that $(\xi_k^{h,h'})^{(j)}$ is continuously differentiable for all j (Hypothesis H_2) then the only possible transformation is an affine map. Indeed it is proved in Biernacki et al. (2002) that there exist exactly two continuously differentiable maps from $\mathbb{R}$ into $\mathbb{R}$ which transform some real-valued normal non-degenerate variable into another one, and that these two maps are both affine. This theoretical result does not concern normal distributions only but it can be extended to any couple of real-valued variables with support $\mathbb{R}$ as $(X^h \mid Z_k^h = 1)^{(j)}$ and $(X^{h'} \mid Z_k^{h'} = 1)^{(j)}$, admitting a symmetric distribution (see Appendix A for a proof).

As a consequence, for all $(h, h') \in \{1, \dots, H\}^2$ and $k \in \{1, \dots, K\}$, there exist $D_k^{h,h'} \in \mathbb{R}^{d \times d}$ diagonal, and $b_k^{h,h'} \in \mathbb{R}^d$ so that:

$$(X^{h'} \mid Z_k^{h'} = 1) \sim D_k^{h,h'}(X^h \mid Z_k^h = 1) + b_k^{h,h'}. \qquad (2)$$

Relation (2) is the affine form of the distributional relationship (1), obtained from both hypotheses H_1 and H_2. It implies on the one hand that inner product matrices and location parameters are linked respectively by:

$$\Sigma_k^{h'} = D_k^{h,h'} \Sigma_k^h D_k^{h,h'} \text{ and } \mu_k^{h'} = D_k^{h,h'} \mu_k^h + b_k^{h,h'}. \qquad (3)$$

Relation (2) implies on the other hand that degrees of freedom are equal through the populations:

$$\nu_k^1 = \dots = \nu_k^H; k \in \{1, \dots, K\}. \qquad (4)$$

As inner product matrices are invertible, the matrices $D_k^{h,h'}$ are non singular. Let us assume henceforward that any couple of corresponding conditional covariables $(X^h \mid Z_k^h = 1)^{(j)}$ and $(X^{h'} \mid Z_k^{h'} = 1)^{(j)}$ are positively correlated. That assumption (Hypothesis H_3) implies that the matrices $D_k^{h,h'}$ are positive, and means that the covariable correlation signs, within some conditional population, remain the same through the populations.

Thus, any couple of identically labeled component parameters, ψ_k^h and $\psi_k^{h'}$, now has to satisfy (4) and there exists a diagonal positive-definite matrix $D_k^{h,h'} \in \mathbb{R}^{d \times d}$ and a vector $b_k^{h,h'} \in \mathbb{R}^d$ which satisfy (3). (Let us note that then $D_k^{h,h'} = D_k^{h',h^{-1}}$ and that $b_k^{h,h'} =- D_k^{h,h'} b_k^{h',h}$.)

The whole parameter space Ψ of ψ is characterized henceforth by both (3) and (4). The so-called simultaneous clustering method relies (in a t-mixture model-based clustering framework) on inference on the parameter ψ on its constrained parameter space.

3. Parsimonious models

Parsimonious models can now be established by combining classical assumptions within each mixture on both mixing proportions and t-parameters (intrapopulation models), with meaningful constraints on the parametric link (3) between the conditional populations (interpopulation models).

3.1.Intrapopulation models

Inspired by Gaussian parsimonious mixtures one can envisage several models of constraints on each t-mixture parameter. Inner product matrices within $P^h (h = 1, \dots, H)$ may be homogeneous ($\Sigma_{\mathrm{k}}^{\mathrm{h}} = \Sigma^{\mathrm{h}}$) or heterogeneous, mixing proportions may be equal (π) or free (π_{k}), degrees of freedom may be homogeneous (ν) or free (ν_{k}). These models will be called *intrapopulation models*.

Although they are not considered here, some other intrapopulation models based on an eigenvalue decomposition of inner product matrices (see Celeux and Govaert, 1995) can be envisaged as an immediate extension of our intrapopulation models.

Remark. Homogeneous inner product matrices in a t-mixture should not be mistaken for homoscedasticity. A t-random vector has finite moments of 2nd order if and only if $\nu > 2$. In this case, the covariance matrix is obtained by multiplying the inner product matrix by $\nu/(\nu-2)$. The homoscedasticity of a t-mixture is then a consequence of assuming (for instance) that (i) inner product matrices are homogenous and (ii) degrees of freedom are both homogeneous and greater than 2.

3.2.Interpopulation models

In the most general case the matrices $D_k^{h,h'}$ are diagonal positive-definite and the vectors $b_k^{h,h'}$ are unconstrained. We can also consider component independent situations for $D_k^{h,h'}$ $(D_k^{h,h'} = D^{h,h'})$ and/or on $b_k^{h,h'}$ $(b_k^{h,h'} = b^{h,h'})$. Other constraints on $D_k^{h,h'}$ and $b_k^{h,h'}$ can be easily proposed but are not considered in this paper (see Lourme and Biernacki, 2010). We can also assume that the mixing proportion vectors $(\pi_1^{\mathrm{h}}, \dots, \pi_{\mathrm{K}}^{\mathrm{h}})$ $(h = 1, \dots, H)$ are either free (π^{h}) or equal (π). These models will be called *interpopulation models* and they have to be combined with some intrapopulation model.

Remark. We can see here that some of the previous constraints cannot be set simultaneously on the transformation matrices and on the translation vectors. When the vectors $b_k^{h,h'}$ do not depend on k for example, then neither do the matrices $D_k^{h,h'}$. Indeed, from (3), we obtain the expression $\mu_k^h = (D_k^{h,h'})^{-1} \mu_k^{h'} - (D_k^{h,h'})^{-1} b_k^{h,h'}$, and consequently $b_k^{h',h} =- (D_k^{h,h'})^{-1} b_k^{h,h'}$ depends on $k$once $D_k^{h,h'}$ or $b_k^{h,h'}$ does.

3.3.Combining inter and intrapopulation models

The most general model of simultaneous clustering is noted $(\pi^h, D_k^{h,h'}, b_k^{h,h'}; \nu_k, \pi_k, \Sigma_k^h)$. It assumes that mixing proportion vectors may be different between populations (so the coefficients π_k^h are free on h), the matrices $D_k^{h,h'}$ are just diagonal positive-definite, the vectors $b_k^{h,h'}$ are unconstrained, and that each mixture has heterogeneous product matrices with free mixing proportions (thus the coefficients π_k^h are also free on k) and non-homogeneous degrees of freedom. The model $(\pi, D^{h,h'}, b^{h,h'}; \nu, \pi, \Sigma^h)$ in another example assumes all mixing proportions to be equal to $1/K$, the matrices $D_k^{h,h'}$ and the vectors $b_k^{h,h'}$ to be component independent and each mixture to have both homogeneous product matrices and homogeneous ν_k.

Since a simultaneous clustering model consists of a combination of some intra and interpopulation models, one will have to pay attention to un-allowed combinations. It is impossible for example to assume both that mixing proportion vectors are free across the diverse populations, and that each of them has equal components.

A model $(\pi^h, ., .; ., \pi, .)$ is therefore not allowed. In the same way, we cannot assume – it is straightforward from the relationship between Σ_k^h and $\Sigma_k^{h'}$in (3) – that both the transformation matrices $D_k^{h,h'}$are free and at the same time that each mixture has homogeneous inner product matrices. A model $(., D_k^{h,h'}, .; ., ., \Sigma^h)$ is therefore prohibited.

Table 1 displays all allowed combinations of intra and interpopulation models, leading to 30 models and Table 2 indicates the associated number δ of free parameters.

Table 1. Allowed intra/interpopulation model combinations and identifiable models. We note by ' · ' non-allowed combinations of intra and interpopulation models, by '°' allowed but non-identifiable models, and by ' • ' both allowed and identifiable models.

			Intrapopulation models							
			ν				ν_k			
			π		π_k		π		π_k	
Interpopulation models			Σ^h	Σ_k^h	Σ^h	Σ_k^h	Σ^h	Σ_k^h	Σ^h	Σ_k^h
$\pi(\pi^h)$	$D^{h,h\prime}$	$b^{h,h\prime}$	• (·)	• (·)	• (•)	• (•)	• (·)	• (·)	• (•)	• (•)
		$b_k^{h,h\prime}$	°(·)	• (·)	• (•)	• (•)	• (·)	• (·)	• (•)	• (•)
	$D_k^{h,h\prime}$	$b_k^{h,h\prime}$	· (·)	• (·)	· (·)	• (•)	· (·)	• (·)	· (·)	• (•)

Table 2. Dimension δ of the parameter ψ in simultaneous clustering in case of both equal mixing proportions and homogeneous conditional degrees of freedom.

		$\boldsymbol{\Sigma^h}$	$\boldsymbol{\Sigma_k^h}$
$D^{h,h\prime}$	$b^{h,h\prime}$	$\beta+\gamma+2d(H-1)$	$\beta+K\gamma+2d(H-1)$
	$b_k^{h,h\prime}$	$\beta+\gamma+d(K+1)(H-1)$	$\beta+K\gamma+d(K+1)(H-1)$
$D_k^{h,h\prime}$	$b_k^{h,h\prime}$	•	$\beta+K\gamma+2dK(H-1)$

Note: $\beta=Kd+1$ denotes the dimension of the parameter component set $\{\mu_1^1,\dots,\mu_K^1\}\cup\{\nu_1^1\}$ and $\gamma=(d^2+d)/2$ is the size of the Σ_1^1 parameter component. If mixing proportions π_k^h are free on both hand k (resp. free on k only), then one must add $H(K-1)$ (resp. $K-1$) to the indicated dimensions below. If degrees of freedom are allowed to vary among the components, then $K-1$ must be added to the indicated dimensions.

Remark. All proposed models are identifiable except one of them $(\pi,D^{h,h\prime},b_k^{h,h\prime};\nu,\pi,\Sigma^h)$, since the latter authorizes different component label permutations depending on the population, and, as a consequence, some crossing of the link between the t-components. Indeed, it is easy to show that in this model, any component may be linked to any other one.

However, assuming the data arise from this unidentifiable model must not be rejected since it just leads to combinatorial possibilities in constituting groups of identical labels from the components C_k^h. In this case, simultaneous clustering provides a partition of the data, but the practitioner keeps some freedom in renumbering the components in each population.

4. Parameter estimation

Notations. In the following sections, indices i and h respectively vary across $\{1,\dots,n^h\}$ and $\{1,\dots,H\}$, and both j and k across $\{1,\dots,K\}$, unless otherwise stated.

4.1. *A useful reparameterization*

The parametric link between the location parameters and the inner product matrices (3) allows for a new parameterization of the model at hand, which is both useful and meaningful for estimating ψ. It is easy to check that for any identifiable model, each matrix $D_k^{h,h\prime}$ is unique as well as each vector $b_k^{h,h\prime}$. As a consequence it makes sense to define for any value of the parameter ψ the following vectors: $\theta^1=\psi^1$ and $\theta^h=[(\pi_k^h,D_k^h,b_k^h);k=1,\dots,K]$ $(h=2,\dots,H)$, where $D_k^h=D_k^{1,h}$ and $b_k^h=b_k^{1,h}$. Let us denote by Θ the space spanned by the vector $\theta=(\theta^1,\dots,\theta^H)$ when ψ scans the parameter space Ψ. There exists a canonical bijective map between Ψ and Θ. Thus θ constitutes a new parameterization of the model at hand, and estimating ψ or θ by maximizing their likelihood, respectively on Ψ or Θ, is equivalent.

The parameter θ^1 appears to be a 'reference population parameter' whereas $(\theta^2,\dots,\theta^H)$ corresponds to a 'link parameter' between the reference population and the other ones. But in spite of appearance the estimated model does not depend on the initial choice of the population P^1. Indeed the bijective correspondence between the parameter spaces θ and ψ ensures that the model inference is invariant by relabelling the populations.

4.2. *Invoking a GEM algorithm*

The loglikelihood of the new parameter θ, computed on the observed data, has no explicit maximum, neither does its expected completed loglikelihood. But Dempster et al. (1977) showed that an EM algorithm is not required to converge to a local maximum of the parameter likelihood in an incomplete data structure. The conditional expectation of its completed loglikelihood has just to increase at each M-step instead of being maximized. This algorithm, called GEM (Generalized EM), can be easily implemented here[1]. Starting from some initial value of

[1] The Matlab code can be obtained from the authors on request.

the parameter θ, the two following steps alternate. The algorithm stops either when reaching stationarity of the likelihood or after a given number of iterations.

- E-step: From the current value $\tilde{\theta}$ of the parameter, the average of $\left(\mathrm{U}^{\mathrm{h}} \mid \mathrm{X}^{\mathrm{h}} = \mathrm{x}_{\mathrm{i}}^{\mathrm{h}}, \mathrm{Z}_{\mathrm{k}}^{\mathrm{h}} = 1 \right)$ is computed with:

$$\begin{aligned} \mathrm{u}_{\mathrm{i,k}}^{\mathrm{h}} &= \mathrm{E}\left(\mathrm{U}^{\mathrm{h}} \mid \mathrm{X}^{\mathrm{h}} = \mathrm{x}_{\mathrm{i}}^{\mathrm{h}}, \mathrm{Z}_{\mathrm{k}}^{\mathrm{h}} = 1; \tilde{\theta} \right) \\ &= \frac{\mathrm{d} + \tilde{\nu}_{\mathrm{k}}^{\mathrm{h}}}{\left[\tilde{\nu}_{\mathrm{k}}^{\mathrm{h}} + \| \mathrm{x}_{\mathrm{i}}^{\mathrm{h}} - \tilde{\mu}_{\mathrm{k}}^{\mathrm{h}} \|^2_{\tilde{\Sigma}_{\mathrm{k}}^{\mathrm{h}}{}^{-1}} \right]} \end{aligned}$$

and the average of its logarithm is given by:

$$\begin{aligned} w_{i,k}^{h} &= E\left(\ln U^{h} \mid X^{h} = x_i^{h}, Z_k^{h} = 1; \tilde{\theta} \right) \\ &= \ln u_{i,k}^{h} + \psi_0 \frac{d + \tilde{\nu}_k^{h}}{2} - \ln \frac{d + \tilde{\nu}_k^{h}}{2}, \end{aligned}$$

where ψ_0 stands for the digamma function.

The expected component memberships are then computed according to:

$$\begin{aligned} t_{i,k}^{h} &= E(Z_k^{h} \mid X^{h} = x_i^{h}; \tilde{\theta}) \\ &= \tilde{\pi}_k^{h} t_d(x_i^{h}; \tilde{\nu}_k^{h}, \tilde{\mu}_k^{h}, \tilde{\Sigma}_k^{h}) / \sum_j \tilde{\pi}_j^{h} t_d(x_i^{h}; \tilde{\nu}_j^{h}, \tilde{\mu}_j^{h}, \tilde{\Sigma}_j^{h}). \end{aligned}$$

- GM-step: The expectation of θ conditional on the completed loglikelihood can be alternatively maximized with respect to the two following component sets of the parameter θ: $\{\pi_{\mathrm{k}}^1, \nu_{\mathrm{k}}^1, \mu_{\mathrm{k}}^1, \Sigma_{\mathrm{k}}^1\}$ and $\{\pi_{\mathrm{k}}^{\mathrm{h}}, \mathrm{D}_{\mathrm{k}}^{\mathrm{h}}, \mathrm{b}_{\mathrm{k}}^{\mathrm{h}}\}$ ($\mathrm{h} = 2, \dots, \mathrm{H}$). It provides the estimator θ^{+} that is used as $\tilde{\theta}$ at the next iteration of the current GM-step. The detail of the GM-step is given in the following two subsections since it depends on the intra and interpopulation model at hand.

4.3.Estimation of the reference population parameter $\boldsymbol{\theta^1}$

From now on, we adopt the convention that for all k, D_k^1 is the identity matrix of $GL_d(\mathbb{R})$ and b_k^1 is the null vector of $\mathbb{R}^d$.

Mixing proportions π_k^1. Setting $\hat{n}_k^h = \sum_i t_{i,k}^h$ and $\hat{n}_k = \sum_h \hat{n}_k^h$, we obtain ${\pi_k^1}^{+} = \hat{n}_k^1 / n^1$ when assuming that mixing proportions are free, ${\pi_k^1}^{+} = \hat{n}_k / n$ when they only depend on the component, and ${\pi_k^1}^{+} = 1/K$ when they depend neither on the component nor on the population.

Degrees of freedom ν_k^1. We recall here that under the conditional linear stochastic link (2), degrees of freedom are homogeneous throughout the populations: $\nu_k^1 = \cdots = \nu_k^H$. When degrees of freedom are allowed to be heterogeneous on k, each ${\nu_k^1}^{+}$ is a solution to the equation:

$$\frac{\partial}{\partial \nu_k} \left[\sum_{i,h} t_{i,k}^h - \ln \Gamma \frac{\nu_k}{2} + \frac{\nu_k}{2} \ln \frac{\nu_k}{2} + \frac{\nu_k}{2} w_{i,k}^h - u_{i,k}^h \right] = 0.$$

Otherwise when degrees of freedom are constrained to be also homogeneous on k, ${\nu_k^1}^{+}$ is solution of:

$$\frac{\partial}{\partial \nu_k} \left[\sum_{i,j,h} t_{i,j}^h - \ln \Gamma \frac{\nu_j}{2} + \frac{\nu_j}{2} \ln \frac{\nu_j}{2} + \frac{\nu_j}{2} w_{i,j}^h - u_{i,j}^h \right] = 0.$$

Location parameters μ_k^1. The component location parameters in the reference population are estimated by:

$${\mu_k^1}^{+} = \sum_{i,h} t_{i,k}^h u_{i,k}^h \Delta_{i,k,h} \,/\, \sum_{i,h} t_{i,k}^h u_{i,k}^h,$$

where $\Delta_{i,k,h} = {\tilde{D}_k^h}{}^{-1} x_i^h - \tilde{b}_k^h$.

Inner product matrices Σ_k^1. If the inner product matrices are allowed to be heterogeneous within each t-mixture, they are then estimated in the reference population by:

$${\Sigma_k^1}^{+} = (1/\hat{n}_k) \sum_{i,h} t_{i,k}^h u_{i,k}^h \left[\Delta_{i,k,h} - {\mu_k^1}^{+} \right] \left[\Delta_{i,k,h} - {\mu_k^1}^{+} \right]'.$$

Otherwise, when assuming that each mixture has homogeneous inner product matrices, those of P^1 are estimated by:

$${\Sigma_k^1}^{+} = (1/n) \sum_{i,j,h} t_{i,j}^h u_{i,j}^h \left[\Delta_{i,j,h} - {\mu_j^1}^{+} \right] \left[\Delta_{i,j,h} - {\mu_j^1}^{+} \right]'.$$

4.4.Estimation of the link parameters $\boldsymbol{\theta^h (h \geq 2)}$

Mixing proportions π_k^h. We have ${\pi_k^h}^{+} = \hat{n}_k^h / n^h$ when assuming that mixing proportions are free, ${\pi_k^h}^{+} = \hat{n}_k / n$ when they only depend on the component, and ${\pi_k^h}^{+} = 1/K$ when they depend neither on the component nor on the population.

Translation vectors b_k^h. When the vectors $b_k^h (k = 1, \dots, K)$ are assumed to be free for any $h \in \{2, \dots, H\}$, they are estimated by:

$${b_k^h}^{+} = \sum_i t_{i,k}^h u_{i,k}^h x_i^h \,/\, \sum_i t_{i,k}^h u_{i,k}^h - \tilde{D}_k^h {\mu_k^1}^{+},$$

and by :

$$
\begin{aligned}
b_k^{h+} \quad = \quad & \left[\sum_{i,j} t_{i,j}^h u_{i,j}^h \left[\tilde{D}_j^h \Sigma_j^{1+} \tilde{D}_j^h\right]^{-1}\right]^{-1} \\
\times \quad & \sum_{i,j} t_{i,j}^h u_{i,j}^h \left[\tilde{D}_j^h \Sigma_j^{1+} \tilde{D}_j^h\right]^{-1} x_i^h - \tilde{D}_j^h \mu_j^{1+}
\end{aligned}
$$

when assuming they are equal.

Matrices D_k^h. The transformation matrices D_k^h cannot be estimated explicitly but, as the expectation of θ conditional on the completed loglikelihood is concave with respect to $D_k^{h^{-1}}$ (whatever are $h \in \{2, \dots, H\}$ and $k \in \{1, \dots, K\}$), we obtain $D_k^{h^+}$ by any convex optimization algorithm.

Remark. Until now we have assumed that the matrices D_k^h were positive. If that assumption is weakened by simply fixing the sign of each coefficient of the matrix D_k^h to be positive or negative, then, first, the identifiability of the model is preserved (whatever the model at hand), and second the expectation of θ conditional on the completed loglikelihood, remains concave with respect to $D_k^{h^{-1}}$ on the parameter space Θ.

We will then always be able to obtain $D_k^{h^+}$ at the GM-step of the GEM algorithm, numerically at least.

5. Companies financial health

5.1.The data

The prediction of a company's ability to satisfy its financial obligations is an important question that requires a strong knowledge of the mechanism leading to bankruptcy. Du Jardin and Séverin (2010) proposed a study of bankruptcy trajectories over the years for a deeper understanding of this process. The original first sample (year 2002) is made up of 250 healthy firms and 250 bankrupt ones. The second sample (year 2003) is made up of 260 healthy and 260 bankrupt companies. The first sample was used to select variables. Forty one variables commonly used in the literature were retained, including forty ratios and one variable representing a balance sheet statement. The ratios were divided into six groups; the first represents the performance of the firms (such as for instance EBITDA/Total assets), the second their efficiency (such as for instance Value added/Total sales), the third their financial distress (such as for instance financial expenses/Total sales), the fourth their financial structure (such as for instance Total debt/Total equity), the fifth their liquidity (such as for instance quick ratio) and the sixth (and last) their rotation (such as for instance Accounts payable/Total sales). Here, we propose to use simultaneous t-model-based clustering in order to obtain both a typology of the financial health of the companies and a quantitative measure of the evolution of that typology over a time period.

We selected a subsample of Du Jardin and Séverin (2010): some outliers are discarded and only the more discriminant variables are kept. The new sample is now made up of a first subsample x^1 of $n^1 = 428$ companies in 2002 (216 healthy and 212 bankrupt companies) and of a second sample x^2 of $n^2 = 461$ companies in 2003 (241 healthy and 220 bankrupt companies). Concerning the variables, only four financial ratios ($d = 4$) expected to provide some meaningful information about the health of the companies, are retained: EBITDA/Total Assets, Value Added/Total Sales, Quick Ratio, Accounts Payable/Total Sales. Figure 1 displays both datasets ($H = 2$) in the canonical plane [EBITDA/Total assets, Quick Ratio].

Note that conditions for using simultaneous clustering are all satisfied. First, both samples are described by the same set of variables. Second, we expect to obtain two partitions (one per year) with the same financial meaning over the years, only their descriptive features having evolved.

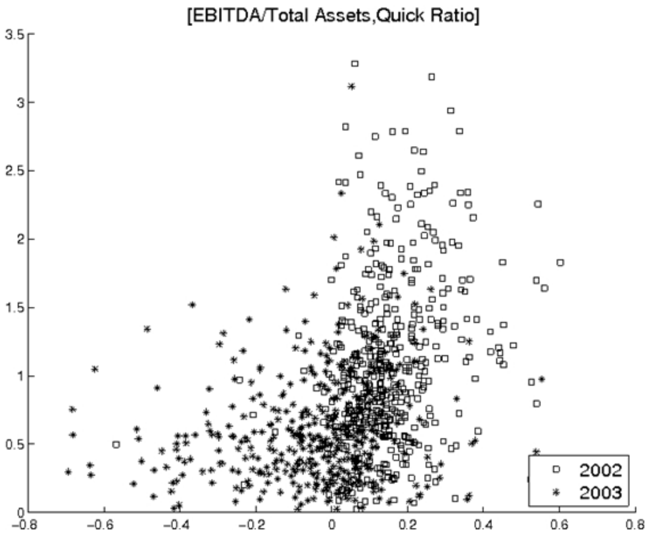


Figure 1. Financial data in the canonical plane [EBITDA/Total Assets, Quick Ratio] for years 2002 and 2003.

5.2.Results of simultaneous vs. independent clustering

We applied on both financial subsamples each of the 30 allowed models of simultaneous clustering displayed in

Table 1 for different numbers of clusters ($K = 1, \dots, 5$) and with the GEM algorithm (5 trials for each procedure, 500 iterations and 5 directional maximizations at each GM step). The selection of the couple (*model*, K) is performed by retaining the greatest value of the ICL (Integrated Completed Likelihood) information criterion (Biernacki et al. 2000) defined by:

$$\mathrm{ICL} = \ln L(\hat{\psi}; x) - \frac{\delta}{2}\ln(n) + \sum_{i,k,h} \hat{z}_{i,k}^{h} \ln t_{i,k}^{h}(\hat{\psi}),$$

where $L(\hat{\psi}; x)$ denotes the maximized likelihood of the parameter ψ computed on the observed data x, δ the dimension of ψ, n the sample size ($n = \sum_h n^h$) and $\hat{z}_{i,k}^{h}$ the MAP of $t_{i,k}^{h}(\hat{\psi})$. Here the ICL criterion is preferred to the BIC (Schwarz, 1978; Lebarbier and Mary-Huard, 2006) since it favors well separated groups, a particularly interesting property for obtaining potentially meaningful clusters.

Table 3 displays the best ICL criterion value among all models for simultaneous clustering strategy. We notice that ICL retains a three clusters ($K = 3$) solution. Table 4 gives the associated confusion table of this obtained partition in comparison to the bankruptcy and healthy specifications.

We see that estimated Clusters 1 and 2 are highly correlated respectively to failed and not-failed companies, whereas Cluster 3 is clearly a group where failed and not-failed companies are indistinguishable. This new typology sheds a new light on the financial health of companies by indicating that it is easy to clearly identify healthy and unhealthy companies (see Figure 2) for a small number of cases (Clusters 1 and 2 have respectively mixing proportions equal to 0.07 and 0.13) whereas it is expected to be a very hard task for most of them (Cluster 3 has a mixing proportion of 0.80). By using the *t*-parameters of each cluster, it is obviously possible to draw a synthetic description of each of them (classical analysis in model-based clustering so not reported here) but we focus on the specificity of simultaneous clustering which provides information about the evolution of the groups over the years. The retained best model is $(\pi, D^{h,h'}, b^{h,h'}; \nu, \pi_k, \Sigma^h)$ which means that (i) the

Table 3. Best ICL values, over all models, obtained in simultaneous and independent clustering with different number of clusters.

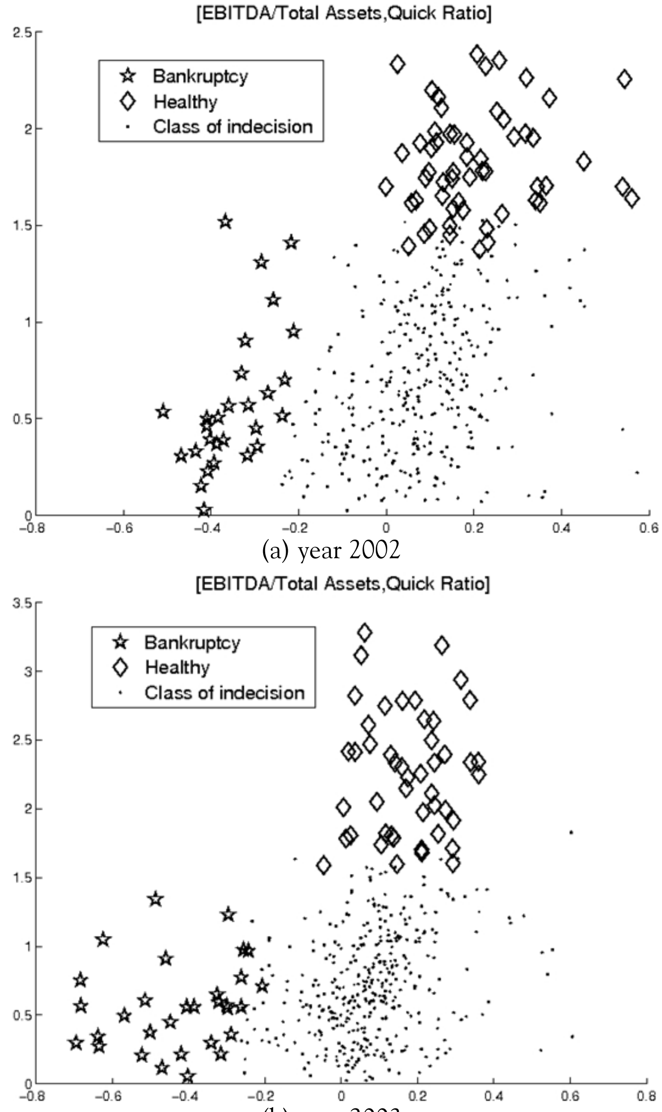
K	1	2	3	4	5
Simultaneous	1169.7	1191.3	**1202.0**	1183.4	1131.3
Independent	1154.6	**1163.6**	1072.1	1127.7	1098.3

Table 4. Confusion table associated to the partition provided by the best simultaneous clustering model retained by ICL.

	Cluster 1	Cluster 2	Cluster 3
Healthy	3	94	360
Bankruptcy	56	10	366

Table 5. Confusion table associated to the partition provided by the best independent clustering model retained by ICL.

	Cluster 1	Cluster 2
Healthy	228	229
Bankrutpcy	289	143



(a) year 2002

(b) year 2003

Figure 2. Estimated partition of companies (Healthy, Bankruptcy, Indecision) for the two consecutive years (2002, 2003), obtained by a simultaneous t-mixture model-based clustering methodology.

mixing proportion of each cluster is invariant between 2002 and 2003 and also (ii) other cluster features uniformly evolved over the years. More precisely, the associated estimated transition parameters are given by

$$\hat{D}^{1,2} = diag(1.12,0.95,1.20,0.93)$$
$$\hat{b}^{1,2} = 10^{-3}.(-18.2,2,-102,-1)',$$

thus the clusters from 2002 and 2003 appear to vary only through the two variables EBITDA/Total Assets and Quick Ratio.

This result is meaningful since the two main variables able to predict bankruptcy are the liquidity and the performance. The change of financial structure is a consequence of the evolution of these two variables. We can assume that the problems of firms arise from several steps. The liquidity ratio collapses before the performance ratio. In some cases, even if we can highlight a decrease in these ratios, the situation still remains good because the other variables (such as financial structure) are strong enough to bear the difficulties the firm faces.

For comparison, Table 3 displays also the best ICL criterion value among all models for independent clustering. We notice now that $K = 2$ clusters are retained and the associated confusion table (Table 5) indicates that estimated clusters yield poor information about the health of companies in comparison to the three-component solution given by simultaneous clustering. In addition, independent clustering does not allow for an easy interpretation of the evolution of the groups over the years. Finally, it is worth noting that ICL prefers the simultaneous solution to the independent one.

6. Concluding remarks

Simultaneous model-based clustering aims to model not only a partitioning of data but also an evolution of it over different subsamples. It was illustrated in the t-case on data related to the financial health of companies over two years. A meaningful three-cluster solution was selected, which was not the case with the classical independent clustering procedure. We also quantified the estimated evolution between the two years. It appears to be moderate in this example but it would be interesting to extend the study to a larger number of years ($H = 4$ or more) for accessing to possibly more changes over more distant years.

Acknowledgements The authors wish to thank E. Séverin and P. Du Jardin for authorizing them to work on their financial datasets and also for their advice.

References

Archambeau, C. and Verleysen, M. 2007. Robust Bayesian clustering. *Neural Networks,* 20(1):129 – 138.

Biernacki, C., Celeux, G. and Govaert, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(7): 719 – 725.

Biernacki, C., Beninel, F. and Bretagnolle, V. 2002. A Generalized Discriminant Rule when Training Population and Test Population Differ on their Descriptive Parameters. *Biometrics*, 58(2):387 – 397.

Bishop, C.M. and Svensén, M. 2005. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235 – 252.

Celeux, G. and Govaert, G. 1995. Gaussian Parsimonious Clustering Models. *Pattern Recognition*, 28(5):781 – 793.

Chatzis, S. and Varvarigou, T. 2008. Robust fuzzy clustering using mixtures of Student's-t distributions. *Pattern Recogn. Lett.*, 29(13):1901 – 1905.

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society* B, 39:1 – 38.

Du Jardin, P. and Séverin, E. 2010. Dynamic analysis of the business failure process: a study of bankruptcy trajectories. In: *Portuguese Finance Network*, Ponte Delgada, Portugal.

Gerogiannis, D., Nikou, C. and Likas, A. 2009. The mixture of Student's t-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9): 1285 – 1294.

Lebarbier, E. and Mary-Huard, T. 2006. Le critère BIC, fondements théoriques et interprétation. *Journal de la Société Francaise de Statistique*, 1:39 – 57.

Lourme, A. and Biernacki, C. 2010. Simultaneous Gaussian Models-Based Clustering for Samples of Multiples Origins. *Pub. IRMA*, 70-VII, University Lille 1, Lille.

McLachlan, G.J. and Peel, D. 2000. Finite Mixture Models. *Wiley*, New York.

Schwarz, G. 1978. Estimating the number of components in a finite mixture model. *Annals of Statistics*, 6:461 – 464.

Wang, H. and Hu, Z. 2009. Estimation for Mixture of Multivariate t-Distributions. *Neural Process. Lett.*, 30(3):243 – 256.

Appendix A

Theorem 1. (Extension of some theoretical result of Biernacki et al. 2002)

Let X and Y be two real-valued, absolutely continuous and symmetric random variables with support $\mathbb{R}$. If some affine map from $\mathbb{R}$ into $\mathbb{R}$ stochastically transforms X into Y, then there exists another such affine map. In this case, these two affine maps are the only C^1-class maps from $\mathbb{R}$ into $\mathbb{R}$ which stochastically transform X into Y.

Proof. Let us assume that there exists a couple $(a, b) \in \mathbb{R}^* \times \mathbb{R}$ such that $Y \sim aX + b$. As X is symmetric, there exists a real number ω such that $(X - \omega)$ and $(\omega - X)$ are identically distributed. It follows that $(aX + b)$ and $(-aX + 2a\omega + b)$ are identically distributed.

Let φ be a map of class C^1 from $\mathbb{R}$ into $\mathbb{R}$ such that $Y \sim \varphi(X)$. Since Y is absolutely continuous, φ is strictly monotonic. Indeed if φ were not strictly monotonic, φ' would be null at some point c and the probability density function of Y would be infinite at $\varphi(c)$.

In addition, as the support of Y is $\mathbb{R}$, φ is surjective from $\mathbb{R}$ into $\mathbb{R}$. Hence φ is a bijection of class C^1 from $\mathbb{R}$ to $\mathbb{R}$.

Let us assume that φ is strictly increasing on $\mathbb{R}$ and let us denote by F_Y the cumulative distribution function of Y. For all real γ, $[X \leq \gamma]$ amounts to $[\varphi(X) \leq \varphi(\gamma)]$ and to $[\, aX + b \leq a\gamma + b \,]$. Since $\varphi(X)$ and $(aX + b)$ are identically distributed as Y it follows that $F_Y(\varphi(\gamma)) = F_Y(a\gamma + b)$. Morevoer since F_Y is a bijection from $\mathbb{R}$ into $(0,1)$, $\varphi(\gamma) = a\gamma + b$.

Let us assume now that φ is strictly decreasing on $\mathbb{R}$. For all real γ, $[X \geq \gamma]$ amounts to $[\varphi(X) \leq \varphi(\gamma)]$ and to $[-aX + 2a\omega + b \leq -a\gamma + 2a\omega + b\,]$. Since $\varphi(X)$ and $(-aX + 2a\omega + b)$ are identically distributed as Y it follows that $F_Y(\varphi(\gamma)) = F_Y(-a\gamma + 2a\omega + b)$ and so $\varphi(\gamma) = -a\gamma + 2a\omega + b$.

Corollary 1.

If X and Y are two real-valued random variables with t-distributions and identical degrees of freedom, there exist exactly two C^1-class maps from $\mathbb{R}$ into $\mathbb{R}$ which transform stochastically X into Y and these two maps are both affine.

Proof. This is an immediate consequence of Theorem 1 since the affine group of $\mathbb{R}$ acts transitively on any family of univariate *t*-distributions with identical degrees of freedom.

Correspondence:
Christophe.Biernacki@math.univ-lille1.fr