Enhancing the selection of a model-based clustering with external qualitative variables

Jean-Patrick Baudry, Margarida Cardoso, Gilles Celeux, Maria-José Amorim, Ana Sousa Ferreira

Enhancing the selection of a model-based clustering with external qualitative variables

Jean-Patrick Baudry [*] Margarida Cardoso [†] Gilles Celeux [‡]
Maria José Amorim [§] Ana Sousa Ferreira [¶]

November 2, 2012

Abstract

In cluster analysis, it could be useful to interpret the obtained partition with respect to external qualitative variables. An approach is proposed in the model-based clustering context to select a model and a number of clusters in order to get a partition which both provides a good fit with the data and is well related to the external variables. This approach makes use of the integrated joint likelihood of the data and the partitions at hand, namely the model-based partition and the partitions associated to the external variables. It is worth noticing that the known partitions are only used to select a relevant mixture model. Each mixture model is fitted by the maximum likelihood methodology from the data. Numerical experiments illustrate the promising behaviour of the derived criterion.

Keywords Mixture models; Model-based clustering; Number of clusters; Penalised criteria; Qualitative variables; BIC; ICL

[*]LSTA, Université Pierre et Marie Curie – Paris VI
[†]BRU-UNIDE, ISCTE-IUL
[‡]INRIA Saclay-Île-de-France
[§]ISEL, ISCTE-Lisbon University Institute
[¶]ProjAVI (MEC), BRU-UNIDE & CEAUL

1 Introduction

In model selection, assuming that the data arose from one of the models in competition is often somewhat unrealistic and could be misleading. However this assumption is implicitly made when using standard model selection criteria such as AIC or BIC. This "true model" assumption could lead to overestimating the model complexity in practical situations. On the other hand, a common feature of standard penalized likelihood criteria such as AIC and BIC is that they do not take into account the modelling purpose. Our opinion is that taking account the modelling purpose when selecting a model leads to more flexible criteria favoring useful and parsimonious models. This point of view could be exploited in many statistical learning situations. Here, it is developed in a model-based clustering context to choose a sensible partition of the data, eventually favoring partitions leading to a relevant interpretation with respect to external qualitative variables. The paper is organised as follows. In Section 2, the framework of model-based clustering is described. Our new penalised likelihood criterion is presented in Section 3. Numerical experiments on simulated and real data sets are presented in Section 4 to illustrate the behavior of this criterion and highlight its possible interest. A short discussion section ends the paper.

2 Model-based clustering

Model-based clustering consists of assuming that the data set to be classified arises from a mixture distribution, trying to recover it at best and associating each cluster with one of the mixture components. Embedding cluster analysis in this precise framework is useful in many aspects. In particular, it allows to choose the number K of classes (i.e. the number of mixture components) in a proper way.

2.1 Finite mixture models

Please refer to McLachlan and Peel (2000) for a comprehensive introduction to finite mixture models.

Data to be classified $\mathbf{y}$ in $\mathbf{R}^{nd}$ are assumed to arise from a mixture

$$\mathbf{f}(\mathbf{y}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k)$$

where the p_k's are the mixing proportions and $\phi(\cdot \mid \mathbf{a}_k)$ denotes the mixture probability density function (as the d-dimensional Gaussian density) with

parameter $\mathbf{a}_k$, and $\theta_K = (p_1, \ldots, p_{K-1}, \mathbf{a}_1, \ldots, \mathbf{a}_K)$. The corresponding parameter space is denoted by Θ_K. A mixture model can be regarded as a latent structure model involving unknown label data $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ which are binary vectors with $z_{ik} = 1$ if and only if $\mathbf{y}_i$ arises from component k. Those indicator vectors define a partition $P = (P_1, \ldots, P_K)$ of the data $\mathbf{y}$ with $P_k = \{\mathbf{y}_i \mid z_{ik} = 1\}$. Each model is usually fitted through maximum likelihood estimation. The corresponding estimator, denoted from now on by $\hat{\theta}_K$, is generally derived with the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997). From a density estimation perspective, a classical way for choosing a mixture model is to select the model maximising the integrated likelihood,

$$\mathbf{f}(\mathbf{y} \mid K) = \int_{\Theta_K} \mathbf{f}(\mathbf{y} \mid \theta_K)\pi(\theta_K)d\theta_K,$$

$$\mathbf{f}(\mathbf{y} \mid \theta_K) = \prod_{i=1}^{n} f(\mathbf{y}_i \mid \theta_K),$$

$\pi(\theta_K)$ being a weakly informative prior distribution on θ_K. For n large enough, it can be approximated with the BIC criterion (Schwarz, 1978)

$$\log \mathbf{f}(\mathbf{y} \mid K) \approx \log \mathbf{f}(\mathbf{y} \mid \hat{\theta}_K) - \frac{\nu_K}{2} \log n,$$

with $\hat{\theta}_K$ the maximum likelihood estimator and ν_K the number of free parameters in the mixture model with K components. Numerical experiments (see for instance Roeder and Wasserman, 1997) show that BIC works well at a practical level for mixture models.

2.2 Choosing K from the clustering view point

In the model-based clustering context, an alternative to the BIC criterion is the ICL criterion (Biernacki et al., 2000) which aims at maximising the integrated likelihood of the complete data $(\mathbf{y}, \mathbf{z})$

$$\mathbf{f}(\mathbf{y}, \mathbf{z} \mid K) = \int_{\Theta_K} \mathbf{f}(\mathbf{y}, \mathbf{z} \mid \theta_K)\pi(\theta_K)d\theta_K,$$

It can be approximated with a BIC-like approximation:

$$\log \mathbf{f}(\mathbf{y}, \mathbf{z} \mid K) \approx \log \mathbf{f}(\mathbf{y}, \mathbf{z} \mid \hat{\theta}_K^*) - \frac{\nu_K}{2} \log n$$

$$\hat{\theta}_K^* = \arg\max_{\theta_K} \mathbf{f}(\mathbf{y}, \mathbf{z} \mid \theta_K).$$

But $\mathbf{z}$ and $\hat{\theta}_K^*$ are unknown. Arguing that $\hat{\theta}_K \approx \hat{\theta}_K^*$ if the mixture components are well separated for n large enough, Biernacki et al. (2000) replace $\hat{\theta}_K^*$ by $\hat{\theta}_K$ and the missing data $\mathbf{z}$ with $\hat{\mathbf{z}} = \text{MAP}(\hat{\theta}_K)$ defined by

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \text{argmax}_\ell \, \tau_i^\ell(\hat{\theta}_K) = k \\ 0 & \text{otherwise,} \end{cases}$$

$\tau_i^k(\hat{\theta}_K)$ denoting the conditional probability that $\mathbf{y}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$):

$$\tau_i^k = \frac{p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k)}{\sum_{\ell=1}^K p_\ell \phi(\mathbf{y}_i \mid \mathbf{a}_\ell)}. \tag{1}$$

Finally the ICL criterion is

$$\text{ICL}(K) = \log \mathbf{f}(\mathbf{y}, \hat{\mathbf{z}} \mid K, \hat{\theta}_K) - \frac{\nu_K}{2} \log n. \tag{2}$$

Roughly speaking ICL is the criterion BIC decreased by the estimated mean entropy

$$E(K) = -\sum_{k=1}^K \sum_{i=1}^n \tau_i^k(\hat{\theta}_K) \log \tau_i^k(\hat{\theta}_K) \geq 0.$$

This is apparent if the estimated labels $\hat{\mathbf{z}}$ are replaced in the definition (2) by their respective conditional expectation $\tau_i^k(\hat{\theta}_K)$.

Because of this additional entropy term, ICL favors values of K giving rise to partitioning the data with the greatest evidence. The derivation and approximations leading to ICL are questioned in Baudry (2009, Chapter 4). However, in practice, ICL appears to provide a stable and reliable estimate of K for real data sets and also for simulated data sets from the clustering view point. ICL, which is not aiming at discovering the true number of mixture components, can underestimate the number of components for simulated data arising from mixtures with poorly separated components Biernacki et al. (2000). It concentrates on selecting a relevant number of classes.

Remark that, for a given number of components K and a parameter θ_K, the class of each observation $\mathbf{y}_i$ is assigned according to the MAP rule defined above.

3 A particular clustering selection criterion

Suppose that the problem is to classify observations described with vectors $\mathbf{y}$'s. But, in addition, a known classification $\mathbf{u}$ on the population, associated to a qualitative variable not directly related to the variables defining

the vector $\mathbf{y}$, is available. Relating the classification $\mathbf{z}$ and the classification $\mathbf{u}$ could be of interest to get a suggestive and simple interpretation of the classification $\mathbf{z}$. With this purpose in mind, it is possible to define a penalized likelihood criterion which selects a model providing a good compromise between the mixture model fit and its ability to lead to a clear classification of the observations well related to the external classification $\mathbf{u}$. Ideally, it is wished that $\mathbf{y}$ and $\mathbf{u}$ should be conditionally independent knowing $\mathbf{z}$, as holds if $\mathbf{u}$ can be written as a function of $\mathbf{z}$. Let us consider the following heuristics. The problem is to find the mixture model m maximising the integrated completed likelihood

$$p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m) = \int p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m, \theta_m)\pi(\theta_m)d\theta_m.$$

Note that, since a mixture model m is not only characterized with the number of components K, but also with assumptions on the proportions and the component variance matrices (see Celeux and Govaert, 1995), it is indexed with m rather than K in the following.

Using a BIC-like approximation as in Biernacki et al. (2000),

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m) \approx \log p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m, \hat{\theta}^*_m) \\ - \frac{\nu_m}{2} \log n, \end{aligned} \tag{3}$$

with

$$\hat{\theta}^*_m = \arg\max_{\theta_m} p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m, \theta_m).$$

An approximation analogous to that leading to ICL is done: $\hat{\theta}^*_m$ is replaced by $\hat{\theta}_m$, the maximum likelihood estimator. The unknown labels $\mathbf{z}$ are then replaced by the labels deduced from the MAP rule with this estimator. Assuming moreover that $\mathbf{y}$ and $\mathbf{u}$ are conditionally independent knowing $\mathbf{z}$, which should hold at least for mixtures with enough components, it can be written

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m, \hat{\theta}^*_m) &= \log p(\mathbf{y}, \mathbf{u} \mid \mathbf{z}, m, \hat{\theta}^*_m) \\ &\quad + \log p(\mathbf{z} \mid m, \hat{\theta}^*_m) \\ &= \log p(\mathbf{y} \mid \mathbf{z}, m, \hat{\theta}^*_m) \\ &\quad + \log p(\mathbf{z} \mid m, \hat{\theta}^*_m) \\ &\quad + \log p(\mathbf{u} \mid \mathbf{z}, m, \hat{\theta}^*_m) \end{aligned} \tag{4}$$

using the conditional independence assumption. From (3) this yields

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{u}, \mathbf{z} \mid m) \approx \log p(\mathbf{y}, \mathbf{z} \mid m, \hat{\theta}_m) \\ + \log p(\mathbf{u} \mid \mathbf{z}, m, \hat{\theta}_m) \\ - \frac{\nu_m}{2} \log n,\end{aligned}$$

and the estimation of $\log p(\mathbf{u} \mid \mathbf{z}, \hat{\theta}_m)$ is derived from the contingency table $(n_{k\ell})$ relating the qualitative variables $\mathbf{u}$ and $\mathbf{z}$: for any $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, U_{\max}\}$, $U_{\max}$ being the number of levels of the variable $\mathbf{u}$,

$$n_{k\ell} = \operatorname{card}\big\{i : z_{ik} = 1 \text{ and } u_i = \ell\big\}.$$

$\log p(\mathbf{u} \mid \mathbf{z}, \hat{\theta}_m)$ is then estimated by

$$\sum_{i=1}^{n} \log \frac{n_{z_i u_i}}{n_{z_i \cdot}} = \sum_{l=1}^{U_{max}} \sum_{k=1}^{K} n_{kl} \log \frac{n_{kl}}{n_{k\cdot}},$$

where $n_{k\cdot} = \sum_{\ell=1}^{U} n_{k\ell}$,

Finally, this leads to the Supervised Integrated Completed Likelihood (SICL) criterion

$$SICL(m) = \mathrm{ICL}(m) + \sum_{\ell=1}^{U_{max}} \sum_{k=1}^{K} n_{k\ell} \log \frac{n_{k\ell}}{n_{k\cdot}}.$$

The last additional term $\sum_{\ell=1}^{U_{max}} \sum_{k=1}^{K} n_{k\ell} \log \frac{n_{k\ell}}{n_{k\cdot}}$ quantifies the strength of the link between the qualitative variables $\mathbf{u}$ and $\mathbf{z}$.

Taking several external variables into account The same kind of derivation enables to derive a criterion that takes into account several external variables $\mathbf{u^1}, \ldots, \mathbf{u^r}$. Suppose that $\mathbf{y}, \mathbf{u^1}, \ldots, \mathbf{u^r}$ are conditionally independent knowing $\mathbf{z}$. Then (4) gets

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{u^1}, \ldots, \mathbf{u^r}, \mathbf{z} \mid m, \hat{\theta}^*_m) = \log p(\mathbf{y} \mid \mathbf{z}, m, \hat{\theta}^*_m) \\ + \log p(\mathbf{z} \mid m, \hat{\theta}^*_m) \\ + \log p(\mathbf{u^1} \mid \mathbf{z}, m, \hat{\theta}^*_m) \\ + \ldots \\ + \log p(\mathbf{u^r} \mid \mathbf{z}, m, \hat{\theta}^*_m),\end{aligned} \tag{5}$$

with $\hat{\theta}_m^* = \arg\max_{\theta_m} p(\mathbf{y}, \mathbf{u^1}, \ldots, \mathbf{u^r}, \mathbf{z} \mid m, \theta_m)$. As before, we assume that $\hat{\theta}_m \approx \hat{\theta}_m^*$ and apply the BIC-like approximation. Finally,

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{u^1}, \ldots, \mathbf{u^r}, \mathbf{z} \mid m) \approx \log p(\mathbf{y}, \mathbf{z} \mid m, \hat{\theta}_m) \\ + \log p(\mathbf{u^1} \mid \mathbf{z}, m, \hat{\theta}_m) \\ + \ldots \\ + \log p(\mathbf{u^r} \mid \mathbf{z}, m, \hat{\theta}_m) \\ - \frac{\nu_m}{2} \log n,\end{aligned}$$

and as before, the estimation of $\log p(\mathbf{u^j} \mid \mathbf{z}, \hat{\theta}_m)$ is derived from the contingency table $(n_{k\ell}^j)$ relating the qualitative variables $\mathbf{u^j}$ and $\mathbf{z}$: for any $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, U_{\max}^j\}$, $U_{\max}^j$ being the number of levels of the variable $\mathbf{u^j}$,

$$n_{k\ell}^j = \text{card}\big\{i : z_{ik} = 1 \text{ and } u_i^j = \ell\big\}.$$

Finally, with $n_{k\cdot} = \sum_{\ell=1}^{U^j} n_{k\ell}^j$ for any j and k (this does not depend on j), we get the multiple external variables criterion "multi-SICL":

$$\text{SICL}(m) = \text{ICL}(m) + \sum_{j=1}^{r} \sum_{\ell=1}^{U_{max}^j} \sum_{k=1}^{K} n_{k\ell}^j \log \frac{n_{k\ell}^j}{n_{k\cdot}}.$$

4 Numerical experiments

We first present two simple applications to show that the SICL criterion is doing the job it is expected to do. The first example is an application to the Iris data set (Fisher, 1936) which consists of 150 observations of four measurements ($\mathbf{y}$) for three species of Iris ($\mathbf{u}$). Those data are depicted in Figure 1 and the variations of criteria BIC, ICL and SICL in function of K are provided in Figure 2. While BIC and ICL choose two classes, SICL selects the three-component mixture solution which is closely related to the species of Iris, as attested by the contingency table between the two partitions (Table 1).

For the second experiment, we simulated 200 observations from a Gaussian mixture in R^2 depicted in Figure 3 and the variable $\mathbf{u}$ corresponds exactly to the mixture component from which each observation arises. Diagonal mixture models (i.e. with diagonal variance matrices) are fitted. The variations of the criteria BIC, ICL and SICL in function of K are provided in Figure 4. We repeated this experiment with 100 different simulated data

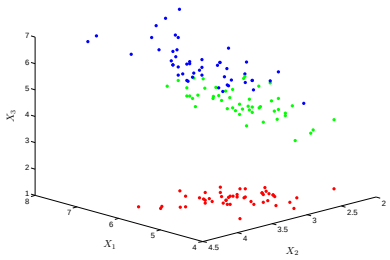

Figure 1: Iris data set

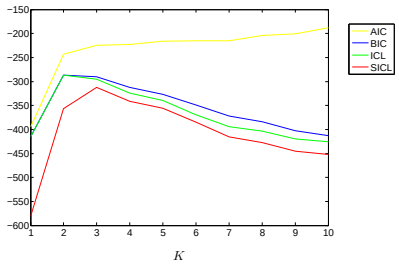

Figure 2: Information criteria for the Iris data set

Table 1: Iris data. Contingency table between the "species" variable and the classes derived from the three-component mixture.

Species \ k	1	2	3
Setosa	0	50	0
Versicolor	45	0	5
Virginica	0	0	50

sets. BIC almost always recovers the four Gaussian components, while ICL almost always selects three because of the two very overlapping ones (the "cross"). Since the solution obtained through MLE with the four-component mixture model yields classes nicely related to the considered $\mathbf{u}$ classes, SICL favors the four-component solution more than ICL does. But since it also takes the overlapping into account, it still selects the three-component model about half of the times (56 times out of 100 in our experiments), and selects the four-component model in almost all the remaining cases (40 out of 100). Actually, as illustrated in Figure 4 for a given data set, SICL hesitates between three and four clusters. In this case, this suggests considering both solutions.

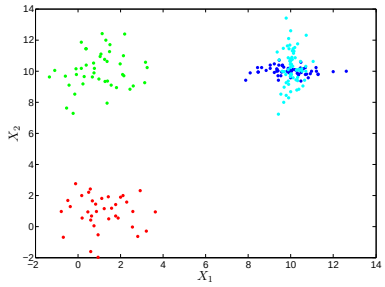

Figure 3: "Cross" data set

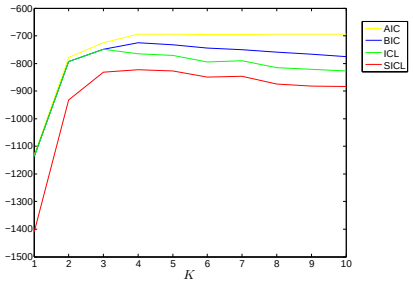

Figure 4: Information criteria for the "Cross" data set

In the next two experiments, we illustrate that SICL does not interfere with the model selection when $\mathbf{u}$ cannot be related with the mixture distributions at hand. At first, we consider a situation where $\mathbf{u}$ is a two-class partition which has no link at all with a four-component mixture data. In Figure 5 the classes of $\mathbf{u}$ are in red and in blue. As is apparent from Figure 6, SICL does not change the solution $K = 4$ provided by BIC and ICL.

Then we consider a two-component mixture and a two-class $\mathbf{u}$ partition "orthogonal" to this mixture. In Figure 7 the classes of $\mathbf{u}$ are in red and in blue. As is apparent from Figure 8, SICL does not change the solution $K = 2$ provided by BIC and ICL despite this solution has no link at all with the $\mathbf{u}$ classes.

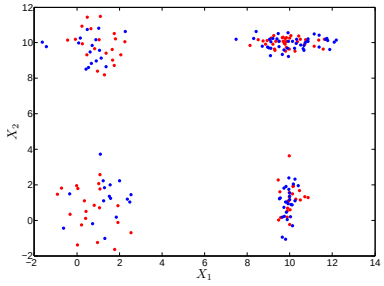

Figure 5: Simulated data set

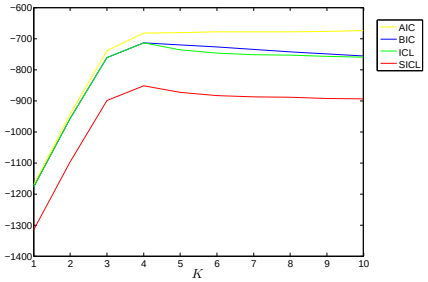

Figure 6: Information criteria for this simulated data set

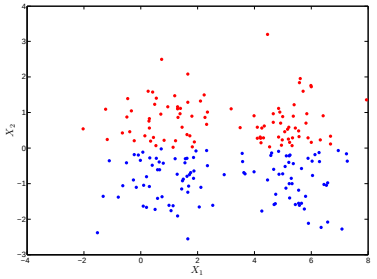

Figure 7: Simulated data set

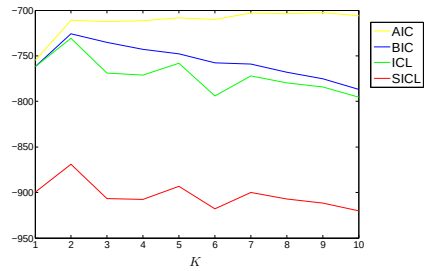

Figure 8: Information criteria for this simulated data set

4.1 Real data set: wholesale customers

The segmentation of customers of a wholesale distributor is performed to illustrate the performance of the SICL criterion. The data set refers to 440 customers of a wholesale: 298 from the Horeca (Hotel/Restaurant/Café) channel and 142 from the Retail channel. They are distributed into two large Portuguese cities regions (Lisbon and Oporto) and a complementary region.

Table 2: Distribution of the Region variable

Region	Frequency	Percentage
Lisbon	77	17.5
Oporto	47	10.5
Other region	316	31.8
Total	440	100

The wholesale data concerns customers. It includes the annual spending in monetary units (m.u.) on product categories: fresh products, milk products, grocery, frozen products, detergents and paper products, and delicatessen. These variables are summarized in Table 3.

Table 3: Product categories sales (m.u.).

	Mean	Std. Deviation
Fresh products	12000	12647
Milk products	5796	5796
Grocery	7951	9503
Frozen	3072	4855
Detergents and Paper	2881	4768
Delicassen	1525	2820

Data also includes responses to a questionnaire intended to evaluate possible managerial actions with potential impact on sales such as improving the store layout, offering discount tickets or extending products' assortment. The customers were asked whether the referred action would have impact on their purchases in the wholesale and their answers were registered in the scale: 1-*Certainly no*; 2-*Probably no*; 3-*Probably yes*; 4-*Certainly yes*. A Gaussian mixture model has been fitted on the continuous variables described in Table 3 with Rmixmod Lebret et al. (2012)(Lebret *et al.*, 2012).

The results are presented in Figure 9.

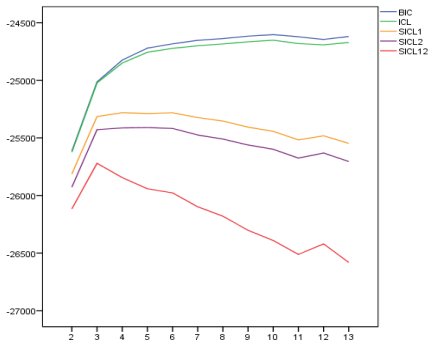

Figure 9: Information criteria for the wholesale dataset

The SICL values based on the Channel, Region, Channel and Region external variables are indicated by SICL1, SICL2 and SICL12 respectively. BIC and ICL select a useless nine-cluster solution, with no clear interpretation. SICL1 selects a four-cluster solution, SICL2 a five-cluster solution and SICL12 a three-cluster solution.

The five-cluster solution is less usable than the alternatives (see Figure 10). Figure 11 highlights the link between the four-cluster solution and the Channel external variable. The product categories spending patterns associated to each cluster are displayed in Figure 12. The cluster 3 is small but includes customers that spend a lot and tend to be particularly sensitive to the potential extension of the products' assortment (see Figure 14).

SICL12 provides the most clear-cut selection (see Figure 9) and parsimonious solution. As a matter of fact, this three-cluster solution is well linked with the external variables (see Figures 15 and 16) while the clusters remain easily discriminated by the product categories' spendings: in particular, cluster 2 (resp. 3) includes a majority of Horeca (resp. Retail) customers buying a lot of fresh products (resp. grocery) (see Figure 13). Cluster 3 is slightly more sensitive to the offering of discount tickets while cluster 2 is slightly more prone to react to improvement of the store layout (see Figure 17).

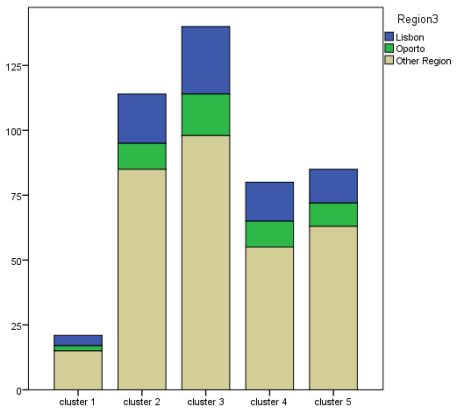

Figure 10: Distribution of the variable Region on the SICL2 solution

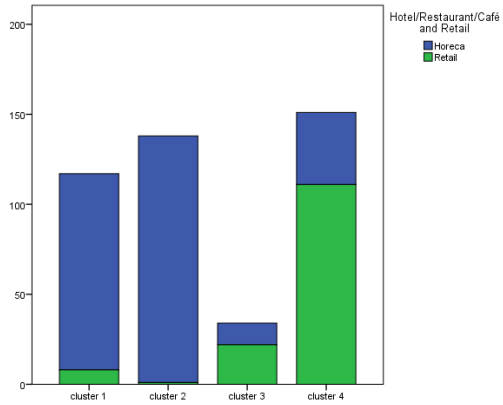

Figure 11: Distribution of the variable Channel on the SICL1 solution

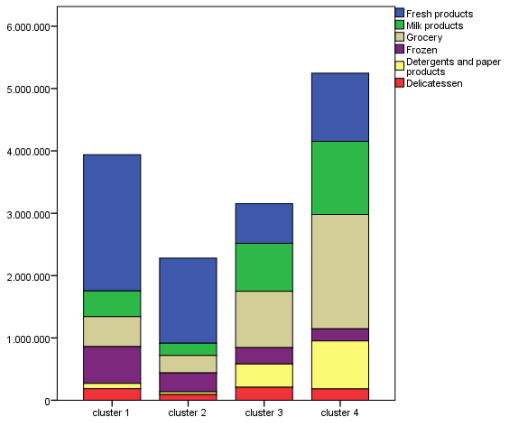

Figure 12: Distribution of the product categories on the SICL1 solution

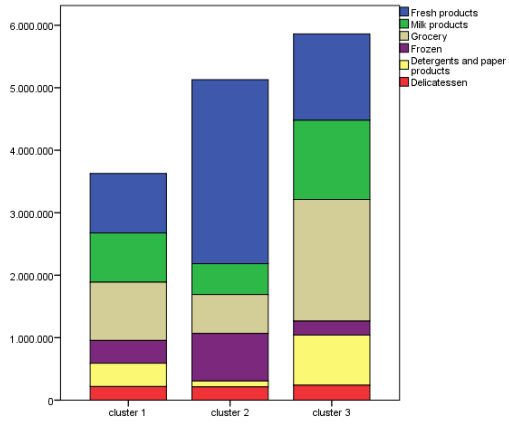

Figure 13: Distribution of the product categories on the SICL12 solution

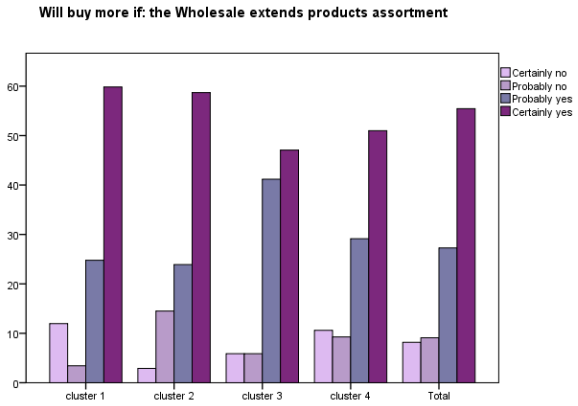

Figure 14: SICL1 solution and managerial actions

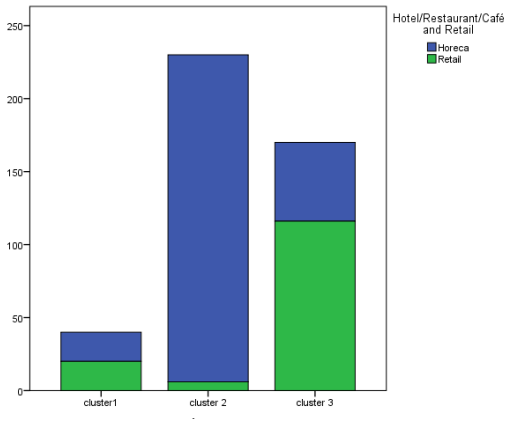

Figure 15: Distribution of the Channel variable on the SICL12 solution

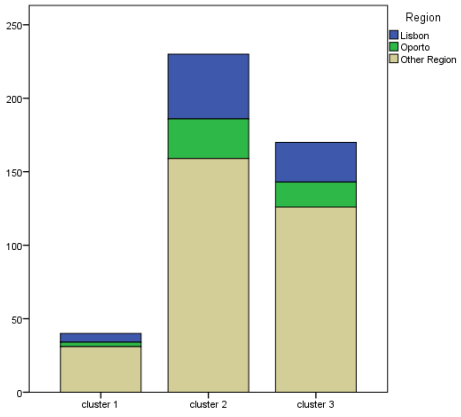

Figure 16: Distribution of the Region variable on the SICL12 solution

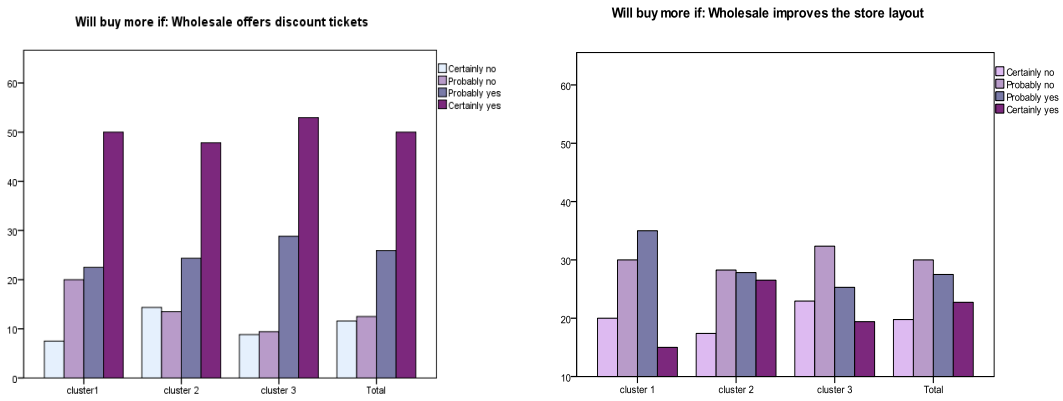

Figure 17: SICL12 solution and managerial actions

5 Discussion

The criterion SICL has been conceived in the model-based clustering context to choose a sensible number of classes possibly well related to an external qualitative variable or a set of external qualitative variables of interest (variables other than the variables on which the clustering is based). This criterion can be useful to draw attention to a well-grounded classification related to this external qualitative variables. It is an example of a model selection criterion taking into account the modeler purpose to choose a useful and stable model. From our experience, in many situations, SICL selects the same models as the criteria ICL or BIC. But when SICL provides a different answer than ICL or BIC, it could shed light to a quite interesting clustering as illustrated in the numerical experiments. It seems that SICL could be expected to select a different partition than ICL particularly when several external variables are considered. Thus, SICL could highlight partitions of special interest with respect to external qualitative variables. Therefore, we think that SICL deserves to enter in the toolkit of model selection criteria for clustering. In most cases, it will propose a sensible solution and when it points out an original solution, it could be of great interest for practical purposes.

References

Baudry, J.-P. (2009). *Model Selection for Clustering. Choosing the Number of Classes.* PhD thesis, Univ. Paris-Sud. `http://tel.archives-ouvertes.fr/tel-00461550/fr/`.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. PAMI*, 22:719–725.

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

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781 – 793.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2012). Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. `http://cran.r-project.org/web/packages/Rmixmod/index.html`.

McLachlan, G. and Krishnan, T. (1997). *The EM-algorithm and Extensions.* New York : Wiley.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York : Wiley.

Roeder, K. and Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

Jean-Patrick Baudry
Université Pierre et Marie Curie - Paris VI
Boîte 158, Tour 15-25, 2^{e} étage
4 place Jussieu, 75252 Paris Cedex 05
France.
Jean-Patrick.Baudry@upmc.fr
`http://www.lsta.upmc.fr/Baudry`