Model selection in block clustering by the integrated classification likelihood

Aurore Lomet, Gérard Govaert, Yves Grandvalet

Model Selection in Block Clustering by the Integrated Classification Likelihood

Aurore Lomet, Gérard Govaert and Yves Grandvalet, *Université de Technologie de Compiègne – CNRS UMR 7253 Heudiasyc*, {aurore.lomet,gerard.govaert,yves.grandvalet}@utc.fr

Abstract. Block clustering (or co-clustering) aims at simultaneously partitioning the rows and columns of a data table to reveal homogeneous block structures. This structure can stem from the latent block model which provides a probabilistic modeling of data tables whose block pattern is defined from the row and column classes. For continuous data, each table entry is typically assumed to follow a Gaussian distribution. For a given data table, several candidate models are usually examined: they may differ in the numbers of clusters or in the number of free parameters. Model selection then becomes a critical issue, for which the tools that have been derived for model-based one-way clustering need to be adapted. In one-way clustering, most selection criteria are based on asymptotical considerations that are difficult to render in block clustering due to dual nature of rows and columns. We circumvent this problem by developing a non-asymptotic criterion based on the Integrated Classification Likelihood. This criterion can be computed in closed form once a proper prior distribution has been defined on the parameters. The experimental results show steady performances for medium to large data tables with well-separated and moderately-separated clusters.

Keywords. Block clustering, integrated classification likelihood, model selection, Gaussian data.

1 Introduction

These last years have seen an increased interest in research and use of co-clustering of data tables, in domains of machine learning, statistics, data mining [1] or genomics [7]. We consider block clustering, whose objective is to organize a data table in homogeneous blocks (clusters). Latent block models [6] enable to define a series of solutions to the block clustering problem, which may differ in their numbers of clusters or in the number of free parameters. Then, model selection [4], becomes a "critical point" to obtain a meaningful classification.

Different model selection techniques have been adapted to block clustering. The first approach combines one-way clustering criteria (like Silhouette index) applied to the row and column clusters, thus ignoring the specific nature of the co-clustering problem [3, 11]. For the latent block model, an extension of AIC-3 has been devised to select the number of clusters, but it still lacks a clear probabilistic interpretation [12]. Indeed, for AIC-3 and other asymptotic criteria, the asymptotic regime needs to be clearly defined which is tricky for the latent block model: a data table is a single element of the set of tables, which can be characterized by its number of row vectors, its number of column vectors, or its number of entries. In the Bayesian setup, a fitting algorithm incorporating the selection of the number of clusters has also been proposed [13]. This computer intensive algorithm relies on Markov chain Monte Carlo to estimate the model with maximum posterior probability (for a given prior distribution on the number of row and column clusters).

We propose, in this paper, another model selection procedure, based on the Integrated Classification Likelihood (ICL) [2]. Our proposal, derived for the latent block model, takes into account the co-clustering structure, is well-motivated from the probabilistic viewpoint, and requires marginal post hoc computations. It does not rely on asymptotic derivations, which is a particularly welcomed property in the block clustering context where asymptotic are not clearly defined.

In this paper, the different notions involved in the development of ICL are first introduced by detailing the Gaussian latent block model and its classification likelihood. Then, our main contribution is presented, that is, the adaptation of the so-called "exact ICL criterion" to the latent block model. The last section compares the performances of this criterion to AIC-3 on a series of numerical experiments.

2 Notations

Throughout this paper, we will use boldface lowercase for vectors, boldface uppercase for matrices, calligraphic uppercases for sets, and medium uppercase for random variables, whatever their type. The $n \times d$ data table to be processed is noted $\mathbf{X} = (\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_n^\mathsf{T})^\mathsf{T}$, with $(\mathbf{x}_i)_j = x_{ij}$, and $x_{ij} \in \mathcal{X}$ will be here a real variable. We will systematically use i as a row index and j as a column index and, when not detailed in sums or products, i goes from 1 to n and j goes from 1 to d. Column j of $\mathbf{X}$ will be noted $\mathbf{x}^j$, so that $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^m)$. The row labeling in g groups, which is noted $\mathbf{z} = (z_1, \ldots, z_n)$, takes its values in $\mathcal{Z} = \{1, \ldots, g\}^n$. Similar notations are given for the column labeling in m groups, with $\mathbf{w} \in \mathcal{W} = \{1, \ldots, m\}^d$. Row and column clusters will respectively be indexed by k and ℓ. Probability distributions, on either discrete or continuous variables are noted $p(\cdot)$.

3 Latent Block Model

Derivation of the Model

The latent-block model [6] is a probabilistic model enabling to classify a data table in homogeneous blocks. The rows and columns of the data table are clustered to reveal similarities and differences between groups of entries. This model generalizes mixture models, its density being

a mixture of block components pertaining to the latent row and column classes:

$$p(\mathbf{X}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z},\mathbf{w})\, p(\mathbf{X}|\mathbf{z},\mathbf{w}) \ .$$

The latent block model is then derived from the two following assumptions:

Assumption 1.
Conditionally on the row and column labels, all table entries are independent, and all the entries belonging to the same block are identically distributed:

$$p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha}) = \prod_{i,j} p(x_{ij}|\boldsymbol{\alpha}_{z_i,w_j}) \ ,$$

where $\boldsymbol{\alpha}$ is a shorthand notation for all $\boldsymbol{\alpha}_{z_i,w_j}$ parameters that characterize the distribution of x_{ij} given (z_i, w_j). For Gaussian data, $\boldsymbol{\alpha}_{k\ell}$ is the mean $\mu_{k\ell}$ and the variance $\sigma^2_{k\ell}$ of block (k,ℓ).

Assumption 2.
All latent variables are independent. The row labels are independent from the column labels:

$$p(\mathbf{z},\mathbf{w}) = p(\mathbf{z})\, p(\mathbf{w}) \ .$$

The row labels are independent and identically distributed, and correspondingly, the column labels are independent and identically distributed:

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\pi}) &= \prod_i p(z_i|\boldsymbol{\pi}) \ , \\ p(\mathbf{w}|\boldsymbol{\rho}) &= \prod_j p(w_j|\boldsymbol{\rho}) \ , \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots \pi_g)$ is the prior probability of row labels and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ is the prior probability of column labels.

Under these assumptions, the probability distribution of the model is:

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z}|\boldsymbol{\pi})\ p(\mathbf{w}|\boldsymbol{\rho})\ p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha}) \ , \\ &= \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,j} \pi_{z_i}\rho_{w_j} p(x_{ij}|\boldsymbol{\alpha}_{z_i,w_j}) \ , \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}, n, d)$ is a shorthand notation for all parameters.

Model Types

A family of models can be defined by considering several numbers of classes in rows and columns. We additionally consider the options where the class priors $(\boldsymbol{\pi}, \boldsymbol{\rho})$ are free or equal and where the variance parameters $\boldsymbol{\sigma}^2$ are free or equal for all clusters . Thus, four models are defined for fixed numbers of clusters by combining these two options.

Complete Data

The latent block model considers that the distribution of the observed data table $\mathbf{X}$ depends on the unobserved latent class variables $(\mathbf{z}, \mathbf{w})$. The complete data refers thus to the triplet $(\mathbf{X}, \mathbf{z}, \mathbf{w})$, and the complete data log-likelihood for the latent block model that assumes that $(\mathbf{X}, \mathbf{z}, \mathbf{w})$ is observed is then:

$$\begin{aligned} L_c(\mathbf{X}, \mathbf{z}, \mathbf{w}|\boldsymbol{\theta}) &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell \\ &\quad + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log p(x_{ij}|\boldsymbol{\alpha}_{k,\ell}) \ , \end{aligned}$$

where, slightly abusing notations, the double subscript class indicators z_{ik} and $w_{j\ell}$ are such that $z_{ik} = 1$ if $z_i = k$ and $z_{ik} = 0$ otherwise.

The complete data likelihood plays a central role in a variational expectation-maximization (VEM) algorithm [9] which is used to the estimate model parameters. As the regular EM algorithm, VEM maximizes the log-likelihood iteratively by maximizing the conditional expectation of the complete data log-likelihood given a current estimate of $\boldsymbol{\theta}$ and the observed $\mathbf{X}$. This maximization relies on the computation of $p(\mathbf{z}, \mathbf{w}|\mathbf{X}, \boldsymbol{\theta})$ at the E-step. Without further assumptions, this computation is intractable for the latent block model, requiring $n^g \times d^m$ operations. This problem is circumvented thanks to a variational approximation assuming the conditional independence of $\mathbf{z}$ and $\mathbf{w}$: $p(\mathbf{z}, \mathbf{w}|\mathbf{X}) = p(\mathbf{z}|\mathbf{X})p(\mathbf{w}|\mathbf{X})$. Both terms of the factorized distribution are easily dealt with, requiring respectively $n \times g$ and $d \times m$ operations. The M-step then proceeds as for the regular EM algorithm. At convergence, this algorithm returns an approximate log-likelihood $\tilde{L}$, its maximizer $\hat{\boldsymbol{\theta}}$, and the estimation of the corresponding most probable row and column labeling $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$. An approximate classification likelihood $L_c(\mathbf{X}, \hat{\mathbf{z}}, \hat{\mathbf{w}}|\hat{\boldsymbol{\theta}})$ can also be computed.

4 ICL for the Gaussian Latent Block Model

The objective of model selection is to choose the "best model". In the Bayesian framework, the best model may be defined as the most probable one given the observed data and a prior distribution over models. When this prior is flat on the finite set of models under consideration, the best model maximizes the integrated likelihood:

$$p(\mathbf{X}|M) = \int_{\mathcal{T}} p(\mathbf{X}|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M) d\boldsymbol{\theta} \ ,$$

where $\mathcal{T}$ is the parameter space and $p(\boldsymbol{\theta}|M)$ is the prior distribution over parameter $\boldsymbol{\theta}$ for model M. This integrated likelihood is commonly approximated by the Bayesian information criterion (BIC). Nevertheless, this approximation is not valid for block clustering due to the inter-dependence between rows and columns and the violation of the necessary regularity conditions [2].

The integrated classification likelihood (ICL) [2] is constructed on the complete data to overcome the problems pertaining to regularity; it then relies on the ability of clustering models to provide evidence with regards to the unobserved data:

$$p(\mathbf{X}, \mathbf{z}, \mathbf{w}|M) = \int_{\mathcal{T}} p(\mathbf{X}, \mathbf{z}, \mathbf{w}|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M) d\boldsymbol{\theta} \ ;$$

for the latent block model described in Section 3, it also solves the issues raised by the dependencies among table entries.

For convenience, the ICL criterion evaluates the logarithm of the integrated classification likelihood:

$$\begin{aligned} ICL(M) &= \log p(\mathbf{X}, \mathbf{z}, \mathbf{w}|M) \\ &= \log p(\mathbf{X}|\mathbf{z}, \mathbf{w}, M) + \log p(\mathbf{z}|M) + \log p(\mathbf{w}|M) \ , \end{aligned} \tag{1}$$

where the decomposition follows from Assumption 2, supposing the independence of priors $p(\boldsymbol{\theta}) = p(\boldsymbol{\alpha})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\rho})$.

Then, $ICL(M)$ can be expanded in several forms according to the model type, as listed in Section 3. For the first term, assuming that the variances differ, we choose a Gaussian prior distribution of parameters $(\mu_0, \sigma^2_{k\ell}/\kappa_0)$ for the conditional mean of a block $\mu_{k\ell}|\sigma^2_{k\ell}$ and a scaled inverse χ^2 prior distribution of parameters (ν_0, σ^2_0) is then appropriate for the variance parameter $\sigma^2_{k\ell}$. The first term of (1) then becomes (see appendix):

$$\begin{aligned} \log p(\mathbf{X}|\mathbf{z}, \mathbf{w}, M) &= -\frac{nd}{2}\log\pi + \frac{gm\nu_0}{2}\log(\nu_0\sigma_0^2) - gm\log\Gamma(\frac{\nu_0}{2}) + \frac{gm}{2}\log\kappa_0 \\ &\quad + \sum_{k,\ell}\Bigg\{ \log\Gamma(\frac{\nu_0 + n_k d_\ell}{2}) - \frac{1}{2}\log(\kappa_0 + n_k d_\ell) \\ &\quad - \frac{\nu_0 + n_k d_\ell}{2}\log\Big(\nu_0\sigma_0^2 + (n_k d_\ell - 1)s^{2\star}_{k\ell} + \frac{\kappa_0 n_k d_\ell}{\kappa_0 + n_k d_\ell}(\bar{x}_{k\ell} - \mu_0)^2\Big)\Bigg\} \ , \end{aligned} \tag{2}$$

where $\bar{x}_{k\ell} = \frac{1}{n_k d_\ell}\sum_{i,j} z_{ik} w_{j\ell} x_{ij}$, $s^{2\star}_{k\ell} = \frac{1}{n_k d_\ell - 1}\sum_{i,j} z_{ik} w_{j\ell}(x_{ij} - \bar{x}_{k\ell})^2$, n_k and d_ℓ are respectively the mean, the unbiased sample variance, the number of rows and the number of columns of the block (k, ℓ).

When the variances are assumed to be equal, a Gaussian prior distribution of parameters $(\mu_0, \sigma^2/\kappa_0)$ may be postulated for $\mu|\sigma^2$ and a scaled inverse χ^2 prior distribution of parameters (ν_0, σ^2_0) for the variance parameter σ^2. The first term of (1) is then:

$$\begin{aligned} \log p(\mathbf{X}|\mathbf{z}, \mathbf{w}, M) &= -\frac{nd}{2}\log\pi + \frac{\nu_0}{2}\log(\nu_0\sigma_0^2) - \log\Gamma(\frac{\nu_0}{2}) + \frac{gm}{2}\log\kappa_0 \\ &\quad + \log\Gamma(\frac{\nu_0 + nd}{2}) - \frac{1}{2}\sum_{k,\ell}\big\{\log(\kappa_0 + n_k d_\ell)\big\} \\ &\quad - \frac{\nu_0 + nd}{2}\log\Big(\nu_0\sigma_0^2 + (nd - gm)s^{2\star}_w + \sum_{k,\ell}\frac{\kappa_0 n_k d_\ell}{\kappa_0 + n_k d_\ell}(\bar{x}_{k\ell} - \mu_0)^2\Big) \ , \end{aligned} \tag{3}$$

where $s^{2\star}_w = \frac{1}{nd - gm}\sum_{k\ell}(n_k d_\ell - 1)s^{2\star}_{k\ell}$ is the unbiased within-cluster sample variance.

When the proportions $(\boldsymbol{\pi}, \boldsymbol{\rho})$ are assumed to differ, a symmetric Dirichlet prior distribution of parameters $(\delta_0, ..., \delta_0)$ is postulated, and the last two terms of (1) respectively become:

$$\log p(\mathbf{z}|M) = \log\Gamma(g\delta_0) + \sum_k \log\Gamma(n_k + \delta_0) - g\log\Gamma(\delta_0) - \log\Gamma(n + g\delta_0) \ , \tag{4}$$

$$\log p(\mathbf{w}|M) = \log\Gamma(m\delta_0) + \sum_\ell \log\Gamma(d_\ell + \delta_0) - m\log\Gamma(\delta_0) - \log\Gamma(d + m\delta_0) \ . \tag{5}$$

Parameters		$ICL(M)$ expansion		
Variances	Proportions	$\log p(\mathbf{X}\|\hat{\mathbf{z}}, \hat{\mathbf{w}}, M)$	$\log p(\hat{\mathbf{z}}\|M)$	$\log p(\hat{\mathbf{w}}\|M)$
σ^2	π, ρ	(3)	(6)	(7)
$\sigma^2_{k\ell}$	π_k, ρ_ℓ	(2)	(4)	(5)
σ^2	π_k, ρ_ℓ	(3)	(4)	(5)
$\sigma^2_{k\ell}$	π, ρ	(2)	(6)	(7)

Table 1: Variants of the $ICL(M)$ expansion, for a single or multiple variance parameters, and for free or fixed proportion parameters.

When the parameters $(\boldsymbol{\pi}, \boldsymbol{\rho})$ are fixed (and equal), $p(\mathbf{z}|M)$ and $p(\mathbf{w}|M)$ are multinomial variables (of parameter $1/g$ and $1/m$ respectively), and the last two terms of (1) respectively become:

$$\log p(\mathbf{z}|M) = n \log g \ , \tag{6}$$

$$\log p(\mathbf{w}|M) = d \log m \ . \tag{7}$$

As the latent variables $\mathbf{z}$ and $\mathbf{w}$ are unobserved, we replace them by their most probable inferred values, $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$, as estimated by the VEM algorithm [9]. Thus, we obtain four variations of $ICL(M)$, whose lengthy expressions are summarized in Table 1.

5 Numerical Experiments

In this section, the performances of ICL are studied on simulated data, which were generated from the latent block model defined in Section 3, so as to have a known ground truth to assess model selection. We use block models with 3 row and 3 column clusters from two model types, either with free or fixed proportions, and both with multiple variances parameters. These two intermediate model types allow for under and over-estimation of the number of free parameters. To avoid unbalanced errors between clusters, the row and column means are equidistant. We consider three degrees of cluster overlap (5%, 12% and 20% of classification error according to conditional Bayes risk) and four data sizes (50×50, 200×200, 500×500 and 1000×1000) (see [8] for thorough details). For each setup, twenty data sets are simulated, and for each data set, 64 candidate latent block models are estimated by the VEM algorithm: the numbers of row and column clusters are varied from 1 to 4, and for each such number, the four model types are considered. Note that these 64 models define the set of possible models over which a flat prior will be assumed to compute the ICL criterion.

Our ICL criterion requires specifying the five hyperparameters $(\delta_0, \mu_0, \kappa_0, \nu_0, \sigma_0^2)$. For the Dirichlet distribution on the proportions parameters, we chose to use a non-informative prior $(\delta_0 = \frac{1}{2})$. For the prior on block means, μ_0 is set to the overall mean of the data table entries. The hyperparameter κ_0 is set to $\sqrt{nd/gm}$: its increase with the number of entries in each cluster encodes a prior assumption on the difficulty of the problem, and the chosen rate reflects our observations on the decrease error rates as table size increases. For the prior on the block variances, the hyperparameters ν_0 and σ_0^2 are defined such as the mode $(\frac{\nu_0 \sigma_0^2}{\nu_0 + 2})$ is equal to the overall variance. The hyperparameter ν_0 is set to 10^{-1} in order to allow for large deviations from this modal value. Therefore, σ_0^2 is equal to $20 s^{2\star}$ where $s^{2\star}$ is the unbiased sample variance.

This type of empirical Bayes method is commonly used to set the value of some of the hyperparameters, such as the ones related to the mean distribution [13]. They may even be

	Table Size	50×50			200×200			500×500			1000×1000		
	Bayes' risk	5%	12%	20%	5%	12%	20%	5%	12%	20%	5%	12%	20%
AIC_3	Cluster	0	0	0	0	0	0	0	0	0	0	0	0
	Model Type	8	8	15	19	20	20	20	20	20	20	20	20
	Cluster and Type	0	0	0	0	0	0	0	0	0	0	0	0
ICL	Cluster	9	0	0	18	2	0	20	17	0	20	20	0
	Model Type	5	5	15	17	20	20	20	20	20	20	20	20
	Cluster and Type	2	0	0	16	2	0	20	17	0	20	20	0

Table 2: Number of correct selections over 20 trials for different table sizes and cluster overlaps. The simulated model has multiple variances and free proportions.

employed in more elaborate construction of priors, such as in hierarchical models [10] . That being said, the mathematical properties of this approach are little known and have been little investigated in general and in co-clustering in particular. In this paper, we fix the values of the hyperparameters as stated above and study their sensitivity to evaluate their effects on model selection.

The ICL criterion is compared to the extension of Akaike Information Criterion 3 proposed by [12]:

$$AIC_3(M) \quad = \quad -2\tilde{L}(\mathbf{X}|\hat{\boldsymbol{\theta}}, M) + 3 \times (\eta_{\boldsymbol{\alpha}} + n(g-1) + d(m-1) + g + m - 2) \ ,$$

where $\tilde{L}$ is the variational approximation of the likelihood, $\hat{\boldsymbol{\theta}}$ is the set of parameters estimated by VEM and $\eta_{\boldsymbol{\alpha}}$ denotes the number of free parameters in $\boldsymbol{\alpha}$. This criterion is adapted to the four variants of the latent block model by adjusting $\eta_{\boldsymbol{\alpha}}$ and dropping the last three terms when proportions are fixed.

The competing criteria aim to retrieve the correct model, which was used for generating the data table. This model is defined by two arguments: its number of row and column clusters and the model type (with single or multiple variance parameters and free or fixed proportions). We tested the ability to select the correct number of clusters knowing model type, the model type knowing the correct number of clusters and to retrieve both model arguments.

Tables 2 and 3 report the number of successes among the 20 repetitions for each setup (respectively for free and fixed proportions). For each criterion three types of selection are studied: the correct number of clusters knowing the type of model, the correct type of model knowing the number of clusters and the correct number of clusters and type of model. Overall, ICL performs better than AIC_3, though AIC_3 is usually slightly better at recovering the model type for known number of clusters. The main weakness of AIC_3 is its inability to retrieve the correct number of clusters even in the most favorable cases (with well-separated clusters and large data tables). This behavior contrasts with the performances of ICL that improve when table size increases and when the degree of overlap decreases, eventually selecting systematically both model types and number of clusters in the most favorable cases tested here. When these criteria fail to select the correct model, AIC_3 and ICL underestimate the number of clusters, while they pick a more complex model type (with free proportions for example).

An empirical study of the sensitivity of the hyperparameters highlights that wrong choices may badly influence the results of the model selection realized by the ICL criterion, but that

	Table Size	50×50			200×200			500×500			1000×1000		
	Bayes' risk	5%	12%	20%	5%	12%	20%	5%	12%	20%	5%	12%	20%
AIC_3	Cluster	0	0	0	0	0	0	0	0	0	0	0	0
	Model Type	19	13	16	16	13	13	20	20	17	20	20	19
	Cluster and Type	0	0	0	0	0	0	0	0	0	0	0	0
ICL	Cluster	13	1	0	16	13	2	20	20	14	20	20	19
	Model Type	20	17	16	16	13	4	20	20	15	20	20	16
	Cluster and Type	12	0	0	16	9	1	20	20	0	20	20	1

Table 3: Number of correct selections over 20 trials for different table sizes and cluster overlaps. The simulated model has multiple variances and fixed proportions.

a wide range of values perform about equally well in this regard. Figure 1 shows the influence of each hyperparameter on the number of correct selections (clusters and models). Each graph represents this number versus one hyperparameter (in log-scale), while the other ones are fixed as stated above. For the hyperparameter μ_0, different values were tested and revealed that the mean of the whole data table is a well-advised choice. The other hyperparameters behave well, with wide plateaus of nearly optimal values. Note that this univariate analysis only provide a partial picture of sensitivity, as we observed a strong coupling between some of them, like σ_0^2 and ν_0 which govern the distribution of the prior on the variance parameter(s) of the model.

6 Conclusion

The Latent Block Model is a probabilistic model that enables to organize a data table in homogeneous blocks. A critical point of the overall fitting procedure is to select a correct model, in order to obtain a meaningful co-clustering.

We proposed a model selection strategy based on the adaptation of the exact integrated classification likelihood criterion to the Gaussian latent block model. This non-asymptotic criterion allows to avoid asymptotical problems due to dual nature of rows and columns in block clustering. Moreover, ICL does not require costly additional calculations. Indeed, once the clusters have been estimated, this criterion can be computed in closed form once a proper prior distribution has been defined on the parameters.

The experiments on simulated data targeted two types of model selection, either related to the number of row and column clusters, or to the model type, or both. Model type is defined from parameter restraints, such as considering equal variances or proportions for all clusters. Our results show that AIC_3 is unable to select the correct number of clusters and that it is less successful than ICL regarding model type. The latter perform well when the overlap between clusters is small to moderate, and its performances improve with the size of the data table. The study of the sensitivity of hyperparameters highlights that adjusting the values of some of them allows to improve the performances of ICL.

Beyond the Gaussian latent block model, this approach can be used for other model types such as the latent block model for binary or contingency data, by modifying appropriately the prior of the conditional distribution on the data.

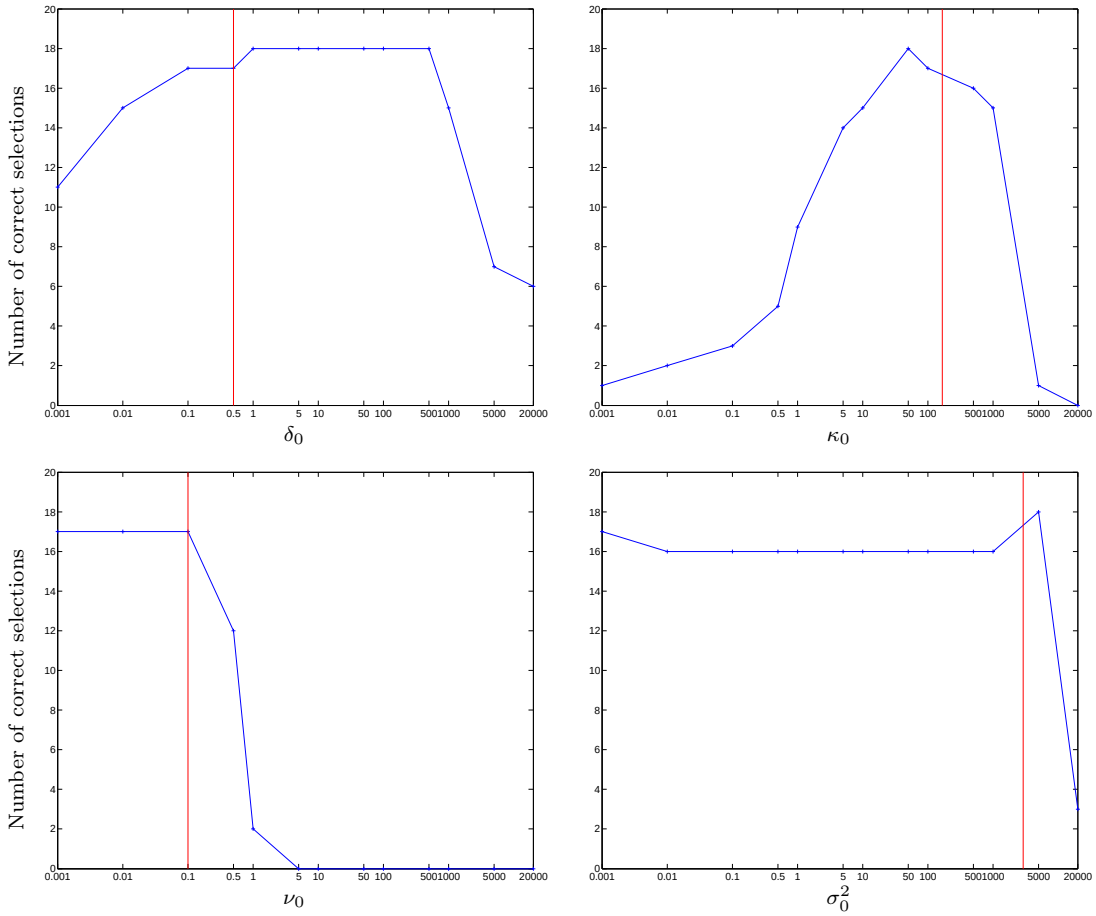

Figure 1: Number of correct selections (clusters and model) according to variations in hyper-parameter values. Here, the data table size is 500×500, the latent block model has different proportions and variances and the clusters are moderately-separated (Bayes' error at 12% for 3×3 clusters).

Acknowledgment This research was partially supported by the French National Research Agency (ANR) under grant ClasSel ANR-08-EMER-002, the PASCAL2 Network of Excellence, and the European ICT FP7 under grant No 247022 - MASH. We would also like to thank the reviewers for their constructive and detailed comments.

1 Appendix: Derivation of ICL

The first term of the expansion (1) is rewritten using the following decomposition:

$$p(\mathbf{X}|\mathbf{z},\mathbf{w},M) \quad = \quad \frac{p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha},M)p(\boldsymbol{\alpha}|M)}{p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{z},\mathbf{w},M)} \ ,$$

where $p(\boldsymbol{\alpha}|M)$ and $p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{z},\mathbf{w},M)$ are respectively the prior and posterior distributions of $\boldsymbol{\alpha}$.

For the latent block model with different variances, given the row and column labels, the entries x_{ij} of each block are independent and identically distributed. We thus apply the standard

results for Gaussian samples [5], where the distributions are defined by:

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha},M) &= \prod_{i,j,k,\ell} \left\{N(x_{ij};\mu_{k\ell},\sigma^2_{k\ell})\right\}^{z_{ik}w_{j\ell}} \ , \\
p(\boldsymbol{\alpha}|M) &= \prod_{k,\ell} \left\{ N(\mu_{k\ell};\mu_0,\frac{\sigma^2_{k\ell}}{\kappa_0}) \times \text{Inv-}\chi^2(\sigma^2_{k\ell};\nu_0,\sigma^2_0) \right\} \ , \\
p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{z},\mathbf{w},M) &= \prod_{k,\ell} \left\{ N(\mu_{k\ell};\frac{\kappa_0\mu_0+n_k d_\ell \bar{x}_{k\ell}}{\kappa_0+n_k d_\ell},\frac{\sigma^2}{\kappa_0+n_k d_\ell}) \right. \\
&\qquad \left. \times \text{Inv-}\chi^2(\sigma^2_{k\ell};\nu_0+n_k d_\ell, \frac{\nu_0\sigma^2_0+(n_k d_\ell-1)s^{2\star}_{k\ell}+\frac{\kappa_0 n_k d_\ell}{\kappa_0+n_k d_\ell}(\bar{x}_{k\ell}-\mu_0)^2}{\nu_0+n_k d_\ell}) \right\} \ .
\end{aligned}
$$

Using the definitions of these distributions, the first term of the expansion (1),

$$
\log p(\mathbf{X}|\mathbf{z},\mathbf{w},M) = \log p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha},M) + \log p(\boldsymbol{\alpha}|M) - \log p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{z},\mathbf{w},M) \ ,
$$

is identified, after some calculations, as (2).

For the latent block model with equal variances, the standard results need to be adapted to account for the shared parameter σ^2. The prior distributions are now defined as follows:

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha},M) &= \prod_{i,j,k,\ell} \left\{N(x_{ij};\mu_{k\ell},\sigma^2)\right\}^{z_{ik}w_{j\ell}} \ , \\
p(\boldsymbol{\alpha}|M) &= \prod_{k,\ell} \left\{ N(\mu_{k\ell};\mu_0,\frac{\sigma^2}{\kappa_0}) \right\} \times \text{Inv-}\chi^2(\sigma^2;\nu_0,\sigma^2_0) \ .
\end{aligned}
$$

The posterior distribution is then computed from the Bayes formula

$$\begin{aligned}
p(\boldsymbol{\mu},\sigma^2|\mathbf{X}) &\propto p(\boldsymbol{\mu},\sigma^2)p(\mathbf{X}|\boldsymbol{\mu},\sigma^2)\\
&\propto (\sigma^2)^{-(\frac{\nu_0}{2}+1)}\exp(-\frac{1}{2\sigma^2}\nu_0\sigma_0^2)\times\prod_{k,\ell}\left\{(\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\kappa_0(\mu_{k\ell}-\mu_0)^2\right)\right\}\\
&\quad\times(\sigma^2)^{-\frac{nd}{2}}\exp\Bigg(-\frac{1}{2\sigma^2}\underbrace{\sum_{i,j,k,\ell}z_{ik}w_{j\ell}(x_{ij}-\mu_{k\ell})^2}_{(nd-gm)s_w^{2\star}+\sum_{k,\ell}n_kd_\ell(\mu_{k\ell}-\bar{x}_{k\ell})^2}\Bigg),\\
&= (\sigma^2)^{-(\frac{\nu_0+nd}{2}+1)}\exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2+(nd-gm)s_w^{2\star})\right)\\
&\quad\times\prod_{k,\ell}\left\{(\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(\kappa_0(\mu_{k\ell}-\mu_0)^2+n_kd_\ell(\mu_{k\ell}-\bar{x}_{k\ell})^2\right)\right)\right\}\\
&= (\sigma^2)^{-(\frac{\nu_0+nd}{2}+1)}\exp\left(-\frac{\nu_0\sigma_0^2+(nd-gm)s_w^{2\star}}{2\sigma^2}\right)\\
&\quad\times\prod_{k,\ell}(\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{\kappa_0+n_kd_\ell}{2\sigma^2}\left(\mu_{k\ell}-\frac{\kappa_0\mu_0+n_kd_\ell\bar{x}_{k\ell}}{\kappa_0+n_kd_\ell}\right)^2-\frac{(\kappa_0n_kd_\ell)(\bar{x}_{k\ell}-\mu_0)^2}{2\sigma^2(\kappa_0+n_kd_\ell)}\right)\\
&= \prod_{k,\ell}(\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{\kappa_0+n_kd_\ell}{2\sigma^2}\left(\mu_{k\ell}-\frac{\kappa_0\mu_0+n_kd_\ell\bar{x}_{k\ell}}{\kappa_0+n_kd_\ell}\right)^2\right)\\
&\quad\times(\sigma^2)^{-(\frac{\nu_0+nd}{2}+1)}\exp\left(-\frac{\nu_0\sigma_0^2+(nd-gm)s_w^{2\star}+\sum_{k,\ell}\frac{\kappa_0n_kd_\ell}{\kappa_0+n_kd_\ell}(\bar{x}_{k\ell}-\mu_0)^2}{2\sigma^2}\right).
\end{aligned}$$

This probability can be factorized:

$$p(\boldsymbol{\mu},\sigma^2|\mathbf{X}) = p(\boldsymbol{\mu}|\sigma^2,\mathbf{X})p(\sigma^2|\mathbf{X})$$

Thus, the posterior distribution is defined (assuming the posterior independence of $\mu_{k\ell}$):

$$\begin{aligned}
p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{z},\mathbf{w},M) =& \prod_{k,\ell}\left\{N\left(\mu_{k\ell};\frac{\kappa_0\mu_0+n_kd_\ell\bar{x}_{k\ell}}{\kappa_0+n_kd_\ell},\frac{\sigma^2}{\kappa_0+n_kd_\ell}\right)\right\}\\
&\times\text{Inv-}\chi^2\left(\sigma^2;\nu_0+nd,\frac{\nu_0\sigma_0^2+(nd-gm)s_w^{2\star}+\sum_{k,\ell}\frac{n_kd_\ell\kappa_0}{n_kd_\ell+\kappa_0}(\bar{x}_{k\ell}-\mu_0)^2}{\nu_0+nd}\right).
\end{aligned}$$

For the terms related to the proportions, when the proportions are free, we assume a symmetric Dirichlet prior distribution of parameters $(\delta_0,...,\delta_0)$ for the row and column parameters $(\boldsymbol{\pi},\boldsymbol{\rho})$, so that:

$$\begin{aligned}
p(\mathbf{z}|M) &= \int_{\mathcal{P}}\pi_1^{n_1}...\pi_g^{n_g}\frac{\Gamma(g\delta_0)}{\Gamma(\delta_0)...\Gamma(\delta_0)}\mathbb{1}_{\sum_k\pi_k=1}d\boldsymbol{\pi},\\
&= \frac{\Gamma(g\delta_0)}{\Gamma(\delta_0)^g}\frac{\Gamma(\delta_0+n_1)...\Gamma(\delta_0+n_g)}{\Gamma(n+g\delta_0)}.
\end{aligned}$$

Details can be found in [2].

Bibliography

[1] P. Berkhin. *A survey of clustering data mining techniques*. Springer, 2006.

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

[3] M. Charrad, Y. Lechevallier, G. Saporta, and M. Ben Ahmed. Détermination du nombre de classes dans les méthodes de bipartitionnement. In *17ème Rencontres de la Société Francophone de Classification*, pages 119–122, Saint-Denis de la Réunion, Île de la Réunion, June 2010. SFC fellowship to attend SFC'2010 conference.

[4] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2004.

[6] G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36:463–473, 2003.

[7] M. Jagalur, C. Pal, E. Learned-Miller, R. T. Zoeller, and D. Kulp. Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics*, 8(Suppl 10):S5, 2007.

[8] A. Lomet, G. Govaert, and Y. Grandvalet. Design of Artificial Data Tables for Co-Clustering Analysis. Technical report, Université de Technologie de Compiègne, 2012.

[9] M. Nadif and G. Govaert. Algorithms for Model-based Block Gaussian Clustering. In *DMIN'08, the 2008 International Conference on Data Mining*, Las Vegas, Nevada, USA, July 14-17 2008.

[10] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[11] J. Schepers, E. Ceulemans, and I. Van Mechelen. Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. *Journal of Classification*, 25(1):67–85, 2008.

[12] B. Van Dijk, J. Van Rosmalen, and R. Paap. A Bayesian approach to two-mode clustering. Technical Report 2009-06, Econometric Institute, 2009.

[13] J. Wyse and N. Friel. Block clustering with collapsed latent block models. *Statistics and Computing*, pages 1–14, 2010.