Integrated Classification Likelihood for Model selection in Block Clustering

Aurore Lomet, Gérard Govaert and Yves Grandvalet
CNRS, Heudiasyc UMR 7253
Université de Technologie de Compiègne
60205 Compiègne, France

March 15, 2012

Abstract

Block clustering (or co-clustering or simultaneous 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 modelling of data tables whose blocks arise from row and column clusters. For continuous data, each table entry is typically assumed to follow a Gaussian distribution whose parameters are common to all entries belonging to the same block, that is, with identical row and column classes. Several candidate models can be adjusted to a given data table: they may differ in the numbers of clusters or more generally 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. We develop here a criterion based on an approximation of the integrated classification likelihood (ICL) of block models, and propose a BIC-like criterion derived from the form obtained. The proposed criteria are illustrated by experiments on simulated data, where their performances are shown to be best reliable for medium to large data tables with well-separated clusters. Keywords: co-clustering, latent block model, model selection, ICL, BIC, simulated data.
1 Introduction

These last years have seen an increased interest in research and use of block clustering of data tables, in domains of machine learning, statistics, data mining or genomics (Berkhin, 2006). One approach is block clustering whose objective is to organize a data table in homogeneous blocks (clusters) by using an estimation of a probabilistic model, the latent block model proposed by Govaert and Nadif (2009). This method can lead to several models defined, for example, for different numbers of clusters or with different constraints on the parameters. Thus, the problem of model selection well-known in clustering (Fraley and Raftery, 1998; Burnham and Anderson, 2004) become a “critical point” to obtain a meaningful classification.

To solve this problem, different solutions have been proposed for latent block models (Dijk et al., 2009; Wyse and Friel, 2010). The first one proposes a Bayesian estimation procedure of the latent block model and use Bayes factors to compare several models (and select the numbers of clusters). The second one proposes an algorithm using a Markov chain Monte Carlo (MCMC) whose the number of row and column clusters need not be known in advance. The model selection is incorporated in the algorithm but need a large running time. We consider here the same approach of Govaert and Nadif (2009) but we propose an other procedure for model selection: we develop an adaption of criteria used for mixture and graph models (Biernacki et al., 2002; Daudin et al., 2008).

A commonly used criterion for model selection is the Bayesian information criterion (Schwarz, 1978). In the Bayesian framework, BIC aims to select the best model among a finite set of models $M$ by choosing the model $M_h$ with highest posterior probability:

$$M_h = \arg\max_{M \in M} P(M|X) ,$$

where $X$ denotes the observed data table.

Assuming that the prior probability $P(M)$ of each model is constant, maximizing $P(M|X)$ is equivalent to maximizing the integrated likelihood $p(X|M)$. Thus, selecting the most probable model consists in selecting the one with the highest integrated likelihood. BIC is a correct approximation of this integrated likelihood under regularity conditions and having an i.i.d. sample. Unfortunately, this criterion can not be used for the latent block model due to regularity problems and the inter-dependence between rows and columns.
To solve this problem, we propose an adaptation of Integrated Classification Likelihood (ICL) (Biernacki et al., 2002) for the latent block model applied for continuous data. This criterion aims to approximate the integrated likelihood conditionally to the row/columns partitions. Under this conditioning, the data are independent and identically distributed.

In this paper, we first introduce the different notions involved in the development of ICL by detailing the latent block model for Gaussian distributions and its classification likelihood. Afterward, we present the main theoretical result, our adaptation of the approximate Integrated Classification Likelihood for the latent block model. The last section reports the performances of $ICL_{BIC}$ obtained with different 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 denoted $X = (x_1^T, \ldots, x_n^T)^T$, with $(x_i)_j = x_{ij}$, and $x_{ij} \in \mathcal{X}$ may be a real, a positive integer or a binary 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 $X$ will be denoted $x^j$, so that $X = (x^1, \ldots, x^m)$. The row labeling in $g$ groups, which is denoted $z = (z_1, \ldots, z_n)$, takes its values in $Z = \{1, \ldots, g\}^n$. Similar notations are given for the column labeling in $m$ groups, with $w \in W = \{1, \ldots, m\}^d$. Probabilities will be denoted by $\mathbb{P}(\cdot)$, and probability distributions, on either discrete or continuous variables by $p(\cdot)$.

3 The Latent Block Model

3.1 Latent-Block Model

The latent-block model (Govaert and Nadif, 2003) is a probabilistic model employed in order to classify a data table in homogeneous blocks. The rows and columns of the data table are organized to reveal similarities and differences between groups of entries. This model generalizes mixture models
whose the density is defined by:

\[ p(X) = \sum_{(z,w)\in Z \times W} p(z,w)p(X|z,w) . \] (1)

The latent block model is based on two assumptions: the model assumes that all entries belonging to the same block are independent and identically distributed (assumption 1) and the prior independence between the row and column labels \(Z\) and \(W\) (assumption 2).

**Assumption 1** All entries belonging to the same block are independent and identically distributed:

\[ p(X|z,w) = \prod_{i,j} p(x_{ij}; \alpha_{z_i,w_j}) \]

where \(\alpha_{z_i,w_j}\) is the parameters of the unidimensional distribution of \(x_{ij}\).

**Assumption 2** All labels are independent. The row and column labels are independent:

\[ p(z,w) = p(z)p(w) \]

The row labels are also independent, and symmetrically the column labels are independent:

\[ p(z) = \prod_i p(z_i) , \]
\[ p(w) = \prod_j p(w_j) . \]

Under assumptions 1 and 2, the probability distribution of the latent block model is defined by:

\[ p(X; \theta) = \sum_{(z,w)\in Z \times W} p(z; \pi) p(w; \rho) p(X|z,w; \alpha) , \] (2)

\[ = \sum_{(z,w)\in Z \times W} \prod_{i,j} \pi_{z_i} \rho_{w_j} p(x_{ij}; \alpha_{z_i,w_j}) \] (3)

where \(\pi = (\pi_1, \ldots, \pi_g)\) is the probability \(p(z = k) = \pi_k\) of row labels, \(\rho = (\rho_1, \ldots, \rho_m)\) is the probability \(p(w = l) = \rho_l\) of column labels, \(\alpha\) is the
set of parameters describing \((X|z,w)\) that will be detailed in (3), and \(\theta = (\pi, \rho, \alpha, n, d)\) is a shortcut notation for all parameters.

As seen from this definition, the model assumes independence between row and column labels \(Z\) and \(W\). We also assume that \(\{Z_i\}_{i=1}^n\) and \(\{W_j\}_{j=1}^d\) are independent. The assumption 2 may easily be misinterpreted, we stress here that the unconditional independence of \(Z\) and \(W\) does not imply their conditional independence knowing the data, that is \(p(z, w|X) \neq p(z|X) p(w|X)\). For example, in market analysis, consumer and product segments can be considered as independent variables, but the purchase of a given product may convey some information about the buyer.

The first assumption of the model implies that the \(n \times d\) unidimensional random variables \(X_{ij}\) are assumed to be independent once \(Z\) and \(W\) are fixed. Under this assumption, their distribution parameters \(\alpha_{z_iw_j}\) are only indexed by the row and columns labels \((z_i, w_j)\). In the following, the data will be supposed to be continuous and we have chosen an unidimensional Gaussian. In this case, \(\alpha_{z_iw_j}\) is the parameter vector \((\mu_{kl}, \sigma^2_{kl})\) where \(\mu_{kl}\) and \(\sigma^2_{kl}\) are respectively the mean and the variance of the row cluster \(k\) and the column cluster \(l\). Considering these parameters, several parsimonious models can be defined:

- the proportions \((\pi, \rho)\) and the variances \(\sigma^2\) of each cluster differ;
- the proportions are equal and the variances differ;
- the proportions differ and the variances are equal;
- the proportions and the variances of each cluster are equal.

Moreover, for each type, several models can be also defined by changing the cluster numbers.

### 3.2 The Complete Data

The latent block model is an incomplete data structure model as mixture models. In this case, the complete data are \((X, z, w)\) where \(X\) is the observed data table and the unobserved vectors \(z\) and \(w\) are the row/column labels. The complete data log-likelihood \(L_c(\theta; X, z, w) = \log f(X, z, w; \theta)\) is then written:

\[
L_c(X, z, w; \theta) = \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,l} w_{jl} \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log p(x_{ij}; \alpha_{z_iw_j}) .
\]
This complete data likelihood is used in the estimation procedure by the VEM algorithm (Nadif and Govaert, 2008). At the end of the algorithm, we obtain the variational approximation of the log-likelihood $\hat{L}$, its maximizer $\hat{\theta}$ and the row and column labeling $(\hat{z}, \hat{w})$ which are estimated by the maximum a posteriori (MAP). An approximation of the classification likelihood $L_c(X, \hat{z}, \hat{w}; \hat{\theta})$ can thus be computed.

4 Integrated Classification Likelihood for the Gaussian latent block model

The objective of model selection is to choose the “best model”. In the Bayesian framework, the best model is the the most probable one. When the models are equiprobable, this is equivalent to choosing the model $M_h$ maximizing the integrated likelihood:

$$M_h = \arg\max_{M} p(X|M),$$

with

$$p(X|M) = \int_\Theta p(X|M; \theta)p(\theta|M)d\theta,$$

where $\Theta$ is the parameter space and $p(\theta|M)$ is the prior probability of the parameter $\theta$ for the model $M$.

A common way to approximate this integrated likelihood is to use the Bayesian information criterion (BIC). Nevertheless, as presented in introduction and by Biernacki et al. (2002), this criterion can not use for the block clustering due to regularity conditions and the inter-dependence between rows and columns. To overcome these problems and to take into account the block structure, we look for maximizing the integrated classification likelihood. The selected model $M_c$ is then defined by:

$$M_c = \arg\max_{M} p(X, z, w|M),$$

with

$$p(X, z, w|M) = \int_\Theta p(X, z, w|M; \theta)p(\theta|M)d\theta.$$

The aim of the ICL criterion is then to evaluate the integrated classification likelihood: $ICL(M) = \log p(X, z, w|M)$ which the contribution of the
missing data \((z, w)\) has to be isolated to avoid the same problems as for BIC.

Under assumption 2, the probability of the integrated classification likelihood can be rewritten as:

\[
p(X, z, w|M) = p(X|z, w, M)p(z, w|M),
\]

and, then, \(ICL\) becomes:

\[
ICL(M) = \log p(X|z, w, M) + \log p(z|M) + \log p(w|M).
\]

To approximate this integrated completed likelihood, a BIC-like approximation is used for the first term under the assumption of the prior independence between the parameters \(p(\theta) = p(\alpha)p(\pi)p(\rho)\). For the two other and under the same assumption, a non-informative Jeffreys prior distribution is applied on the row and column parameters \((\pi, \rho)\) and a Stirling approximation is used. Then, the ICL criterion can be approximated (see appendix) by:

\[
ICL(M) \approx \log p(X, z, w|M; \theta^*) - \lambda \frac{1}{2} \log(nd) - \frac{g - 1}{2} \log n - \frac{m - 1}{2} \log d,
\]

where \(\theta^* = \arg \max_\theta \log p(X, z, w|M; \theta)\).

The parameter \(\theta^*\) and the unobserved vectors of labels \((z, w)\) being unknown, they are replaced by the estimated parameter \(\hat{\theta}\) (estimated by the VEM algorithm) and the partitions \((\hat{z}, \hat{w}) = MAP(\hat{\theta})\). Finally, we propose the ICL criterion:

\[
ICL_{BIC}(M) = \log p(X, \hat{z}, \hat{w}|M; \hat{\theta}) - \lambda \frac{1}{2} \log(nd) - \frac{g - 1}{2} \log n - \frac{m - 1}{2} \log d,
\]

\[
= L_c(X, \hat{z}, \hat{w}|M; \hat{\theta}) - \lambda \frac{1}{2} \log(nd) - \frac{g - 1}{2} \log n - \frac{m - 1}{2} \log d.
\]

This criterion is a penalized classification likelihood whose the penalty is the sum of a penalty on the parameters of Gaussian distributions, a penalty
on the parameters of the row proportions and a penalty on the parameters of the columns proportions.

Using the link which exists between $BIC$ and $ICL$ in the classical mixture model, we propose, in an analogical way, the follow $BIC_{like}$ criterion:

$$BIC_{like}(M) = \log p(X|M, \theta^*) - \frac{\lambda}{2} \log(n d) - \frac{g - 1}{2} \log n - \frac{m - 1}{2} \log d,$$

Because $\log p(X|M, \theta^*)$ is hard to compute (Nadif and Govaert, 2008), this quantity is replaced by $\tilde{L}(X|M, \hat{\theta})$ estimated by the VEM algorithm:

$$BIC_{like}(M) = \tilde{L}(X|M, \hat{\theta}) - \frac{\lambda}{2} \log(n d) - \frac{g - 1}{2} \log n - \frac{m - 1}{2} \log d.$$

5 Numerical Experiments

In this section, the performances of $ICL_{BIC}$ and $BIC_{like}$ are studied on simulated data from the latent block model defined in section 3.2, whose the advantage is to have a gold standard to assess model selection. These criteria are compared to an other traditional criterion, the Akaike Information Criterion, adapted for the latent block model by:

$$AIC_{like}(M) = \tilde{L}(X|M, \hat{\theta}) - \frac{\lambda + g + m - 2}{2}.$$  

5.1 Simulations experiments

The performances of $ICL_{BIC}$, $BIC_{like}$ and $AIC_{like}$ are evaluated for its ability to select the correct model (the model used for the simulation of data). This model is defined by two arguments whose the first one is the number of clusters and the second is the type of model (with common or different variances and proportion). Thus, two series of experiments and a series of a cross-experiment (selection of the type of model and cluster) are carried out.

5.1.1 Experimental Setup

We consider the three above experimental setups, where the model is either modified by adjusting the number of clusters or by adjusting the type of
model or the both. In the first series of experiments, the objective is to evaluate the ability of $ICL_{BIC}$, $BIC_{like}$ and $AIC_{like}$ to find the correct number of clusters (that is, of the true generative model). The type of model is fixed but the number of parameters varies through the number of clusters (in rows and columns). In the second series of experiments, we test the performance of criteria regarding its ability to select the correct type of model assuming the number of clusters to be known. In the last one, the performances of criteria are evaluated for its ability to find simultaneously the correct cluster number and the correct type of model.

In these three series of experiments, the data are simulated from a latent block model whose means and variances differ and whose proportions are equal (three row clusters and three column clusters). To avoid unbalanced errors between clusters, the means are equidistant. We consider three degrees of cluster overlap: 5%, 12% and 20% of classification error according to conditional Bayes risk, which are respectively considered as well separated, moderately separated and ill-separated as represented in Figure 5.1.1. For each configuration, the experiments are carried out on tables of different sizes: $50 \times 50$, $200 \times 200$, $500 \times 500$ and $1000 \times 1000$. To obtain this three degrees of overlap for these different table size, the parameter distribution is modified by varying the variance. For the twelve possibilities (three degrees of cluster overlap $\times$ four table sizes), twenty data sets are simulated. The latent block models of each data table are estimated by the VEM algorithm.

![Figure 1: Projections of the rows of two data tables on the two first column eigenvectors.](image)

well-separated clusters  moderately-separated clusters  ill-separated clusters
5.1.2 Results

In this section, the performances of $ICL_{BIC}$, $BIC_{like}$ and $AIC_{like}$ are evaluated regarding its ability to recover:

- the cluster number knowing the type of model (A),
- the type of model knowing the cluster number (B),
- the cluster number and type of model (C).

For each configuration described above, the criteria are calculated for the 16 candidate models obtained by considering all number pairs of row and columns clusters between one and four, and for the four different types of model.

The selected model is the one which has the lowest value of the criterion (to respect tradition on criteria for model selection, all criteria presented before are multiply by -2 in the experiments). For the first type of experiments (A), the selection is declared correct if the model has three row and three column clusters. For the second (B), the selection is declared correct if the type of model selected corresponds to the one used for simulations (different means and variances, equal row and column proportions). Finally, for the last series of experiments (C), the selection is declared correct if the lowest value of a criterion is obtained for the $3 \times 3$ clusters and the model whose the proportions are equal and the variances differ.

Table 1 to 4 reports the number of successes among 20 repetitions of each setup for different table sizes. Generally, $AIC_{like}$ overestimates the cluster number (for example, $3 \times 4$ or $4 \times 4$) although it can occasionally underestimate it (for instance, $3 \times 2$). $BIC_{like}$ and $ICL_{BIC}$ underestimate the cluster number (for example, $2 \times 2$ or $2 \times 3$). Regarding the selection of model type, the three criteria choose the model whose the proportions and the variances of each cluster differ or the correct model.

Some results of other simulations with different models confirmed that the $BIC_{like}$ criterion has better results than the other criteria. Moreover, the quality of $BIC_{like}$ and $ICL_{BIC}$ improves with the increase of data size and the decrease of the overlap degree. The results of $AIC_{like}$ showed that this criterion is less efficient than the other criteria and unstable: its behavior depends on the model and not necessary on the overlap degree.
<table>
<thead>
<tr>
<th>Degree of overlap (%)</th>
<th>( AIC_{like} )</th>
<th>( BIC_{like} )</th>
<th>( ICL_{BIC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection type</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
</tr>
<tr>
<td>cluster</td>
<td>16 16 16</td>
<td>3 0 0</td>
<td>2 0 0</td>
</tr>
<tr>
<td>model</td>
<td>19 10 10</td>
<td>20 18 18</td>
<td>20 16 11</td>
</tr>
<tr>
<td>cluster and model</td>
<td>12 4 2</td>
<td>2 0 0</td>
<td>1 0 0</td>
</tr>
</tbody>
</table>

Table 1: Number of correct selections over 20 trials for a data size 50 × 50.

<table>
<thead>
<tr>
<th>Degree of overlap (%)</th>
<th>( AIC_{like} )</th>
<th>( BIC_{like} )</th>
<th>( ICL_{BIC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection type</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
</tr>
<tr>
<td>cluster</td>
<td>15 15 16</td>
<td>16 11 2</td>
<td>16 4 1</td>
</tr>
<tr>
<td>model</td>
<td>15 11 9</td>
<td>17 14 18</td>
<td>17 13 8</td>
</tr>
<tr>
<td>cluster and model</td>
<td>12 7 2</td>
<td>16 0 0</td>
<td>16 0 0</td>
</tr>
</tbody>
</table>

Table 2: Number of correct selections over 20 trials for a data size 200 × 200.

<table>
<thead>
<tr>
<th>Degree of overlap (%)</th>
<th>( AIC_{like} )</th>
<th>( BIC_{like} )</th>
<th>( ICL_{BIC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection type</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
</tr>
<tr>
<td>cluster</td>
<td>20 20 20</td>
<td>20 20 20</td>
<td>20 20 2</td>
</tr>
<tr>
<td>model</td>
<td>18 16 14</td>
<td>20 20 20</td>
<td>20 20 14</td>
</tr>
<tr>
<td>cluster and model</td>
<td>10 9 10</td>
<td>20 20 7</td>
<td>20 17 0</td>
</tr>
</tbody>
</table>

Table 3: Number of correct selections over 20 trials for a data size 500 × 500.

<table>
<thead>
<tr>
<th>Degree of overlap (%)</th>
<th>( AIC_{like} )</th>
<th>( BIC_{like} )</th>
<th>( ICL_{BIC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection type</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
<td>5% 12% 20%</td>
</tr>
<tr>
<td>cluster</td>
<td>20 20 20</td>
<td>20 20 20</td>
<td>20 20 12</td>
</tr>
<tr>
<td>model</td>
<td>20 20 12</td>
<td>20 20 10</td>
<td>20 20 14</td>
</tr>
<tr>
<td>cluster and model</td>
<td>13 15 10</td>
<td>20 20 10</td>
<td>20 20 0</td>
</tr>
</tbody>
</table>

Table 4: Number of correct selections over 20 trials for a data size 1000 × 1000.
6 Conclusion

The Latent Block Model is employed in statistics to organize a data table in homogeneous blocks. A critical point of this method is to select the correct model in order to obtain the meaningful co-clustering. Because we use an probabilistic model, this problem can be seen as a problem of model selection where a model is defined by a number of row and column clusters and a type of model (different or equal variances or proportions).

A current solution in model selection in statistics is the Bayesian Information Criterion. However, this criterion cannot be applied for the latent block model due to regularity problems and the inter-dependence between the rows and the columns of the data table. To solve it, we have proposed an adaptation of the Integrated Classification Likelihood for latent block models applied on continuous data, defined by:

\[ ICL_{BIC}(M) = \log p(X, \hat{z}, \hat{w}|M, \hat{\theta}) - \frac{\lambda}{2} \log(nd) - \frac{g-1}{2} \log n - \frac{m-1}{2} \log d . \]

Derived of this criteria, we propose also a \( BIC_{like} \) criterion:

\[ BIC_{like}(M) = \hat{L}(X|M, \hat{\theta}) - \frac{\lambda}{2} \log(nd) - \frac{g-1}{2} \log n - \frac{m-1}{2} \log d . \]

These criteria have the advantage to have no use for long additional calculation. Indeed, in a first time, the MLE \( \hat{\theta} \) are estimated by the VEM algorithm. Then, the row/columns labels \( (\hat{z}, \hat{w}) \) are computed by an iterative algorithm of the Maximum A Posteriori (MAP) which converges toward a local minimum in finite time. After the procedure of estimation, the criteria \( ICL_{BIC} \) and \( BIC_{like} \) are directly computed. Moreover, the user does not need to enter a lot of parameters at the beginning of his experiments; he only needs to know how many clusters and type of models he wants to test.

To evaluate the performances of these criteria, we tested it on simulated data and we compared the results to another criterion generally used \( AIC_{like} \). The results on simulated data showed that \( AIC_{like} \) is unstable and less successful than \( BIC_{like} \) and \( ICL_{BIC} \) which are efficient to select the correct model when the overlap degree is not too large. Moreover, the quality of these criteria improves with the increase of the data size. A study on real data remains to realize.
Acknowledgment This research was supported by the CLasSel ANR project ANR-08-EMER-002.

A Appendix: Development of $ICL_{BIC}$

The criterion $ICL$ can be written in three terms:

$$ICL(M) = \log p(X|z, w, M) + \log p(z|M) + \log p(w|M) .$$

Under the assumption of the prior independence between the parameters $p(\theta) = p(\alpha)p(\pi)p(\rho)$, these three terms becomes:

$$\log p(X|z, w, M) = \int_\Theta p(X|z, w, M, \theta)p(\theta|M)d\theta ,$$

$$= \int_\Theta p(X|z, w, M, \theta)p(\alpha|M)p(\pi|M)p(\rho|M)d\alpha d\pi d\rho ,$$

$$= \int_A p(X|z, w, M, \alpha)p(\alpha|M)d\alpha \int_{P \times R} p(\pi|M)p(\rho|M)d\pi d\rho = 1 ,$$

$$\log p(z|M) = \int_\Theta p(z|M, \theta)p(\theta|M)d\theta ,$$

$$= \int_P p(z|M, \pi)p(\pi|M)p(\alpha|M)p(\rho|M)d\alpha d\pi d\rho ,$$

$$= \int_P p(z|M, \pi)p(\pi|M)d\pi \int_{A \times R} p(\alpha|M)p(\rho|M)d\alpha d\rho = 1 ,$$

$$\log p(w|M) = \int_\Theta p(w|M, \theta)p(\theta|M)d\theta ,$$

$$= \int_R p(w|M, \rho)p(\rho|M)p(\alpha|M)p(\pi|M)d\alpha d\pi d\rho ,$$

$$= \int_P p(w|M, \rho)p(\rho|M)d\rho \int_{A \times P} p(\alpha|M)p(\pi|M)d\alpha d\rho = 1 .$$

For the first term ($\log p(X|z, w, M)$), the BIC-like approximation can be
used (conditionally to the row/column partitions, the variables are i.i.d.):

\[
\log p(X|z, w, M) = \log \int_A p(X|z, w, M, \alpha)p(\alpha|M)d\alpha, \\
\approx \max_\alpha \log p(X|z, w, M, \alpha) - \frac{\lambda}{2} \log(nd),
\]

where \(\lambda\) is the dimension vector \(\alpha\) in the space \(A\).

For the others terms \(\log p(z|M)\) (et \(\log p(w|M)\)), the BIC-like approximation is unnecessary to calculate \(\log p(z|M)\) and \(\log p(w|M)\). By taking a flat prior distribution on \(\pi\) and \(\rho\) (a Dirichlet distribution \(D(\delta, \ldots, \delta)\)), we obtain:

\[
p(z|M) = \int_\mathcal{P} \prod_{k=1}^{g} \pi_{k}^{n_k+1} \frac{\Gamma(g\delta)}{\Gamma(\delta) \cdots \Gamma(\delta)} \prod_{k=1}^{g} \pi_{k}^{n_k+1} \sum_{k=1}^{g} \pi_{k} = \frac{\Gamma(g\delta) \Gamma(\delta + n_1) \cdots \Gamma(\delta + n_g)}{\Gamma(\delta)^g \Gamma(n + g\delta)},
\]

where \(n_k\) is the number of rows in the cluster \(k\).

Using the non-informative Jeffreys prior distribution for the proportion parameters \((\delta = 1/2)\), the log-likelihood of the prior distributions of the partitions become:

\[
\log p(z|M) = \log \Gamma\left(\frac{g}{2}\right) + \sum_{k=1}^{g} \log \Gamma(n_k + 1) - g \log \Gamma\left(\frac{1}{2}\right) - \log(n + g), \\
\log p(w|M) = \log \Gamma\left(\frac{m}{2}\right) + \sum_{l=1}^{m} \log \Gamma(d_l + 1) - m \log \Gamma\left(\frac{1}{2}\right) - \log(d + m).
\]

When \(n_k\) and \(d_l\) are large enough, the approximation of the Gamma function by the Stirling formula \(\Gamma(t+1) \approx t^{t+1/2} \exp(-t)(2\pi)^{1/2}\) can be used. Thus, applying this approximation and neglecting terms of order \(O(1)\), the
log-likelihood of the prior distributions of the partitions are:

\[
\log p(z|M) \approx \sum_{k=1}^{g} n_k \log n_k - n \log n \\
+ (\delta - \frac{1}{2}) \sum_{k=1}^{g} n_k - (g\delta - \frac{1}{2}) \log n ,
\]

\[
= -\frac{1}{2} (g-1) \log n \text{ for } \delta = 1/2
\]

\[
\log p(w|M) \approx \sum_{l=1}^{m} d_l \log d_l - d \log d \\
+ (\delta - \frac{1}{2}) \sum_{l=1}^{m} d_l - (m\delta - \frac{1}{2}) \log d .
\]

\[
= -\frac{1}{2} (m-1) \log m \text{ for } \delta = 1/2
\]

In addition, \( \sum_{k=1}^{g} n_k \log n_k = \max_\pi \log p(z|M, \pi) \) and \( \sum_{l=1}^{m} d_l \log d_l = \max_\rho \log p(w|M, \rho) \). For \( \delta = 1/2 \), we obtain:

\[
\log p(z|M) \approx \max_\pi \log p(z|M, \pi) - \frac{g-1}{2} \log n ,
\]

\[
\log p(w|M) \approx \max_\rho \log p(w|M, \rho) - \frac{m-1}{2} \log d .
\]

Then, the ICL criterion can be approached by:

\[
ICL(M) = \log p(X|z, w, M) + \log p(z|M) + \log p(w|M) ,
\]

\[
\approx \max_{\alpha} \log p(X|z, w, M, \alpha) - \frac{\lambda}{2} \log (nd) \\
+ \max_\pi \log p(z|M, \pi) - \frac{g-1}{2} \log n \\
+ \max_\rho \log p(w|M, \rho) - \frac{m-1}{2} \log d ,
\]

\[
\approx \max_{\alpha, \pi, \rho} \log p(X, z, w|M, \alpha, \pi, \rho) \\
- \frac{\lambda}{2} \log (nd) - \frac{g-1}{2} \log n - \frac{m-1}{2} \log d ,
\]

\[
\approx \max_\theta \log p(X, z, w|M; \theta) \\
- \frac{\lambda}{2} \log (nd) - \frac{g-1}{2} \log n - \frac{m-1}{2} \log d .
\]
Bibliography


