Examining the robustness of a model selection procedure in the binary latent block model through a language placement test data set

Vincent Brault, Frédérique Letué, Marie-José Martinez

Examining the robustness of a model selection procedure in the binary latent block model through a language placement test data set

Vincent Brault, Frédérique Letué, Marie-José Martinez

Univ. Grenoble Alpes, CNRS, Grenoble INP[1]*, LJK, 38000 Grenoble, France*

Abstract

When entering French university, the students' foreign language level is assessed through a placement test. In this work, we model the placement test results using binary latent block models which allow to simultaneously form homogeneous groups of students and of items. However, a major difficulty in latent block models is to select correctly the number of groups of rows and the number of groups of columns. The first purpose of this paper is to tune the number of initializations needed to limit the initial values problem in the estimation algorithm in order to propose a model selection procedure in the placement test context. Computational studies based on simulated data sets and on two placement test data sets are investigated. The second purpose is to investigate the robustness of the proposed model selection procedure in terms of stability of the students groups when the number of students varies.

Keywords: Latent block model, Model selection, Robustness, Placement test data.

1. Introduction

When entering French university, students may have different foreign language levels. They need to be evaluated before being directed in an adequate foreign language class. To assess their level, universities use different placement tests. The SELF test developed in Grenoble is one of the most used ([1, 2]). At the end of this test, each student get an aggregated score which

[1]Institute of Engineering Univ. Grenoble Alpes

corresponds to the course level where he/she has to register in. He/she also gets a mark for each of the three evaluated skills (oral comprehension, written comprehension and written expression). Moreover, the test creators need to check the relevance of the questions, in particular whether they are discriminating or not. It may be useful to form groups of items more or less difficult per skill.

Results of such a placement test can be displayed as a matrix where each row corresponds to one student and each column to one item. Element (i, j) of the matrix equals 1 if student i answers correctly question j, and 0 otherwise. Latent block models ([3]) aim to simultaneously achieve a clustering of the rows and the columns and thus turn out to be particularly useful to form homogeneous groups of students and of items in the placement test context.

Classical algorithms to estimate parameters in latent block models, namely Variational or Stochastic Expectation Maximization ([3, 4, 5]), are extensions of the EM-algorithm ([6]). Although these algorithms can give satisfactory estimates, they appear to be quite sensitive to starting values and have a marked tendency to provide empty clusters. To overcome these limitations, Keribin et al. [5] proposed several algorithms through Bayesian inference using Gibbs sampling.

A major difficulty in latent block models is to select correctly the number of groups of rows and the number of groups of columns. For this purpose, penalized likelihood criteria such as Akaike Information Criterion (AIC, [7, 8]) or Bayesian Information Criterion (BIC, [9]) are not directly usable since computing the maximized likelihood is not possible. Another widely used criterion is the Integrated Completed Likelihood criterion (ICL) defined by Biernacki et al. [10]. This criterion has been extended to the latent block model for binary data in [11] and for categorical data in [5].

The first purpose of this paper is to tune the number of initializations needed to limit the initial values problem in the estimation algorithm in order to propose a model selection procedure for latent block models on binary data in the context of placement tests. Two computational studies based on simulated data sets and on the two placement test data sets described in Section 2 are investigated. The second purpose of this paper is to investigate the robustness of the proposed model selection procedure. The robustness is here assessed in terms of the stability of the students groups when the number of students varies.

The paper is organized as follows. The two placement test data sets used in this paper are described in the next section. In Section 3, we recall the statistical model, the notations and the estimation algorithm and we present the model selection procedure based on ICL criterion considered in this work. Section 4 deals with the effect of initialization strategies on the model selection procedure. The robustness of the proposed procedure is investigated in Section 5. Finally, a conclusion presenting a perspective for future work ends this paper.

2. Data sets

The data sets considered in this paper have been obtained from the SELF placement test developed at Université Grenoble Alpes in 6 languages (English, French as a Foreign Language, Italian, Japanese, Mandarin Chinese and Spanish) and used at a number of partner universities in France (see [1]). The SELF test is a semi-adaptive multi-stage test. The first stage (the initial testlet) is common to all test takers, but the items in the second stage depend on test takers' results in the first. Results in the second stage are used to refine the estimation of learners' level and arrive at placement results expressed in Common European Framework of Reference (CEFR) levels that are as reliable as possible. In this paper, the results of a Japanese and an English SELF placement tests are considered. The Japanese (resp. English) SELF test has been taken by 137 (resp. 228) students entering Université Grenoble Alpes in 2019. We focus here only on the first stage composed with 33 (resp. 36) items common to all test takers. These items are labelled with three skills: oral comprehension, written comprehension and written expression.

Figure 1 displays the English SELF placement test results for the 228 students (lines) and the 36 items (columns). Each element of the matrix is colored in white when the student answers correctly the question, and in black otherwise.

3. Statistical framework

In this section, the statistical model, the notations and the estimation algorithm are first described. Then, the model selection procedure based on ICL criterion is presented.

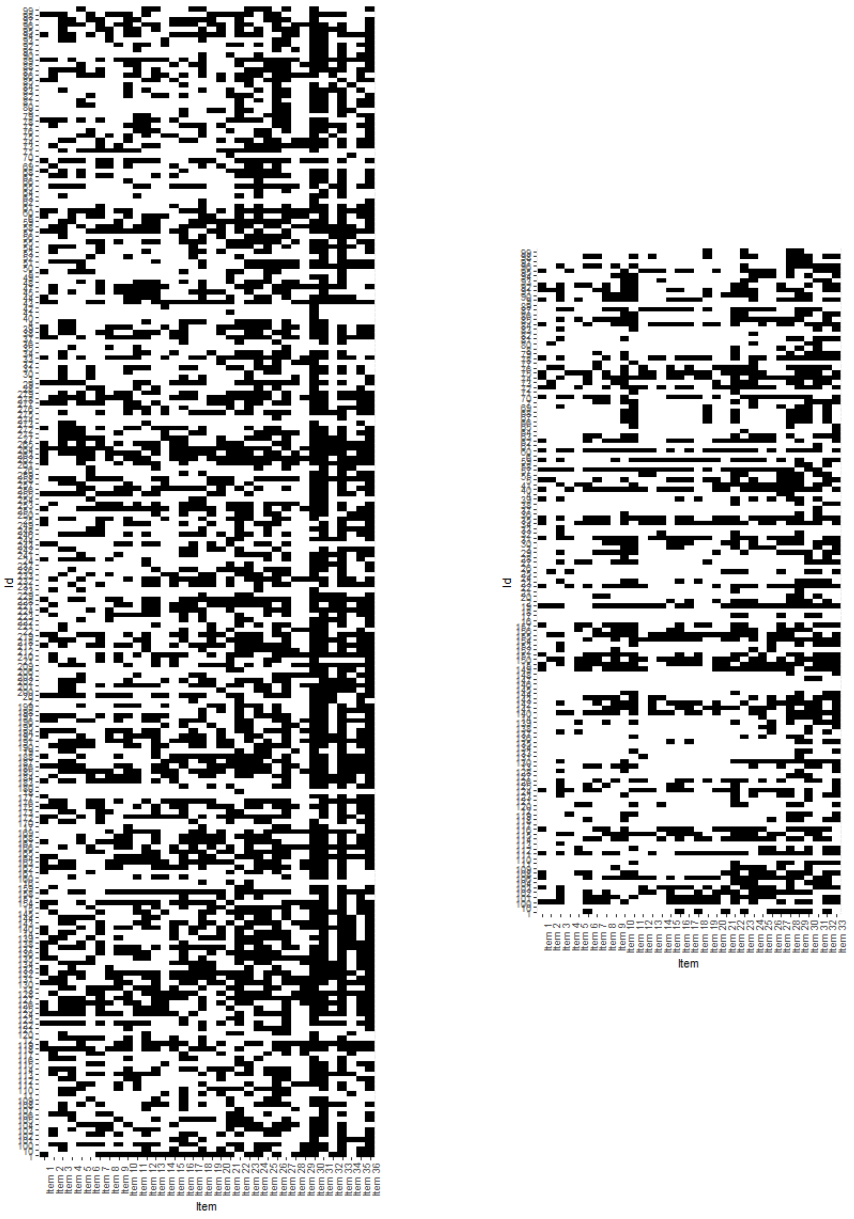

Figure 1: Results of the English SELF placement test (on left) for 228 students (lines) and 36 items (columns) and the Japanese SELF placement test (on right) for 137 students (lines) and 33 items (columns). A white cell corresponds to a correct answer.

3.1. Statistical model and notations

Let n be the number of students, g the number of students groups, q the number of items and m the number of items groups. We define $Z_i, i = 1 \cdots n$, as the independent random variables modelling the students group which student i belongs to. The $Z_i, i = 1 \cdots n$ variables follow the multinomial $\mathcal{M}(1; \pi_1, \ldots, \pi_g)$ distribution. We also define $W_j, j = 1 \cdots q$, as the independent random variables modelling the items group which item j belongs to. The $W_j, j = 1 \cdots q$ variables follow the multinomial $\mathcal{M}(1; \rho_1, \ldots, \rho_m)$ distribution.

Responses are assumed to be independent. Given that student i belongs to group k and item j belongs to group l, the response of student i to item j denoted by Y_{ij} follows the Bernoulli distribution with parameter α_{kl}:

$$P(Y_{ij} = 1 | Z_i = k, W_j = l) = \alpha_{kl}.$$

3.2. Estimation algorithm

A classical method to estimate parameters in latent variables models is the EM-algorithm. [3, 4] show that the EM-algorithm cannot directly be used in practice. Indeed, it would require to calculate a sum over all possible couples $(z_i, w_j), 1 \leq i \leq n, 1 \leq j \leq q$, which would have a too high computational cost. To overcome this problem, [4] propose a variational approximation of the EM-algorithm (VEM) based on a decomposition of the log-likelihood as the sum of the free energy and the Kullback-Leibler divergence ([12]). Since the Kullback-Leibler divergence is expected to be small around the log-likelihood maximum, they propose to only maximise the free energy. In [13], the authors show that the free energy maximum estimator is consistent.

As pointed by Govaert and Nadif [4] and Keribin et al. [5], the estimated parameters obtained by the VEM-algorithm highly depend on its initial values. Moreover, Keribin et al. [5] show that this algorithm tends to provide empty groups. To overcome these problems, they propose a V-Bayes algorithm to avoid empty groups, combined with a Gibbs sampler to limit the initial values problem. In this bayesian approach, proper and independent informative prior distributions are considered for the parameters. The mixing proportions $\pi = (\pi_1, \ldots, \pi_g)$ and $\rho = (\rho_1, \ldots, \rho_m)$ are assumed to be Dirichlet-distributed with parameters $(a, \ldots, a)$. The parameters $\alpha_{11}, \ldots, \alpha_{gm}$ are assumed to be Beta-distributed with parameter (b, b).

In Section 4, hyperparameters a and b in the estimation algorithm are chosen to be equal to 4 and 1 respectively, as advised by Keribin et al. [5]. In this paper, we use the combined V-Bayes Gibbs algorithm to get estimators $\widehat{\pi}, \widehat{\rho}, \widehat{\alpha}$ of the parameters for a given number of students groups g and a given number of items groups m. Furthermore, in order to limit the initial values problem, this algorithm is run T times and we finally keep the estimators $\widehat{\pi}, \widehat{\rho}, \widehat{\alpha}$ which maximize the free energy over the T runs.

3.3. Model selection procedure and the ICL criterion

In this subsection, we investigate the ICL model selection criterion in order to propose a model selection procedure.

The ICL criterion has been defined in [10] as the logarithm of the integrated completed likelihood in a mixture model context. It has been extended to the latent block model for binary data in [11] and for categorical data in [5].

Using the conditional independence of the z's and the w's conditionally to π, ρ and α, the conjugate properties of the prior Dirichlet distributions and the conditional independence of the y_{ij} given the latent variables z and w, the ICL criterion can be written as

$$\begin{aligned} ICL_{(z,w)}(g,m) &= \log\Gamma(ga) + \log\Gamma(ma) - (m+g)\log\Gamma(a) \\ &\quad + mg(\log\Gamma(2b) - 2\log\Gamma(b)) - \log\Gamma(n+ga) - \log\Gamma(q+ma) \\ &\quad + \sum_{k}\log\Gamma(z_{.k}+a) + \sum_{l}\log\Gamma(w_{.l}+a) \\ &\quad + \sum_{k,l}\left[\log\Gamma(N_{kl}^{1}+b) + \log\Gamma(N_{kl}^{0}+b) - \log\Gamma(z_{.k}w_{.l}+2b)\right] \end{aligned}$$

where

$$z_{.k} = \sum_{i}\mathbb{1}_{\{z_i=k\}}, \qquad w_{.l} = \sum_{j}\mathbb{1}_{\{w_j=l\}},$$

$$N_{kl}^{1} = \sum_{i,j} y_{ij}\mathbb{1}_{\{z_i=k,w_j=l\}}, \qquad N_{kl}^{0} = \sum_{i,j}(1-y_{ij})\mathbb{1}_{\{z_i=k,w_j=l\}}.$$

Details can be found in [5].

Given (g, m), the maximization of $ICL_{(z,w)}(g, m)$ over all partitions (z, w) would require a too high computational cost. In their paper, Biernacki et al. biernacki2000 proposed to use $ICL_{(\widehat{z},\widehat{w})}(g, m)$ where $(\widehat{z}, \widehat{w})$ is the maximizing partition obtained with a maximum a posteriori (MAP) rule after the last V-Bayes step.

In order to select the numbers of students groups g and of items groups m, we calculate every criterion $ICL_{(\widehat{z},\widehat{w})}(g, m)$ where (g, m) runs on a given grid, and we finally define $(\widehat{g}, \widehat{m})$ as the pair that maximizes these criteria.

4. Effect of initialization strategies on the model selection procedure

The objective of this section is to tune the parameter T, namely the number of initializations needed to limit the initial values problem in the estimation algorithm. For that purpose, we investigate two computational studies. The first one is based on simulated data sets, and the second one is based on the two real data sets described in Section 2.

4.1. Simulation study

4.1.1. Simulation plan and indicators

In this simulation study, we simulate $L = 100$ data sets from the model described in Section 3.1. Mimicking the Japanese SELF placement test data set, we set $n = 137$ and $q = 33$. We also set $g = 3$, $m = 4$, $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$ and $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \frac{1}{4}$. Finally, we define parameter α_{kl} equal to ε if $k \geq l$ and $1 - \varepsilon$ if $k < l$ with $\varepsilon \in \{0.05, 0.15, 0.2, 0.25, 0.3\}$. Note that parameter ε can be considered as an indicator of the estimation difficulty.

In order to tune parameter T, we run the following algorithm:

Table 1 displays the distribution of parameter T for each value of ε.

4.1.2. Results

Results displayed in Table 1 show that for small values of ε ($\varepsilon \leq 0.25$), $T = 1$ is enough to select the simulated model ($g = 3$ and $m = 4$) most of the time (96% for $\varepsilon = 0.05$, 99% for $\varepsilon = 0.15$, 98% for $\varepsilon = 0.2$ and 79% for $\varepsilon = 0.25$). That means that for simple cases running the estimation algorithm once per pair (g, m) is usually sufficient.

for $\varepsilon \in \{0.05, 0.15, 0.2, 0.25, 0.3\}$ **do**
  for *L from 1 to 100* **do**
    simulate a data set
    $T = 0$
    $(\hat{g}, \hat{m}) = (1, 1)$
    while $(\hat{g}, \hat{m}) \neq (3, 4)$ **do**
      $T = T + 1$
      for *each pair* (g, m)*, with* g *and* m *varying from 1 to 7* **do**
        1. calculate the estimators $(\widehat{\pi}, \widehat{\rho}, \widehat{\alpha})$;
        2. calculate the maximizing a posteriori partition $(\widehat{z}, \widehat{w})$;
        3. calculate the associated ICL value;

      end
      select the pair $(\hat{g}, \hat{m})$ that maximizes the ICL criterion.
    end
  end
end

Table 1: Distribution of T (columns) in function of the value ε (rows).

	1	2	14	≥ 50
$\varepsilon = 0.05$	96	4	0	0
$\varepsilon = 0.15$	99	1	0	0
$\varepsilon = 0.2$	98	0	0	2
$\varepsilon = 0.25$	79	0	1	20
$\varepsilon = 0.3$	24	0	0	76

For $\varepsilon = 0.3$, we can see from Table 1 that the distribution of parameter T is shifted to higher values. That means that for more difficult cases, we need to run the estimation algorithm a high number of times per pair (g, m) to retrieve the simulated model.

4.2. Study based on real data sets

In this subsection, we evaluate the performance of our procedure from the two data sets described in Section 2. For that purpose, we proceed in two steps.

First, since there is no "true" model in a real dataset context, we run the

following algorithm a great number of times $K = 170\,000$ to get a reference model :

for $k = 1$ *to* K **do**
  for *each pair* (g, m)*, with* g *and* m *varying from 1 to 7* **do**
    1. calculate the estimators $(\widehat{\pi}, \widehat{\rho}, \widehat{\alpha})$;
    2. calculate the maximizing a posteriori partition $(\widehat{z}, \widehat{w})$;
    3. calculate the associated ICL value;
  end
  $(\hat{g}, \hat{m})_k$ is the pair $(\hat{g}, \hat{m})$ that maximizes the ICL criterion.
end
$(g, m)_{ref}$ is the pair $(\hat{g}, \hat{m})_k$ that maximizes the ICL criterion over the K pairs.

Secondly, we consider the number of times this algorithm finds the reference model over the K runs and we study the inter-arrival times of the reference model along the K runs.

4.2.1. The Japanese SELF placement test data set

For the Japanese SELF dataset, the obtained reference model is $(3, 4)$. Over the $K = 170\,000$ runs, the reference model has been obtained 170 000 times. That means that the inter-arrival times are all equal to 1. This dataset can be considered as an "easy" case, for which only one initialization is needed.

Table 2 gives one realization of the estimation of each parameter and Figure 2 displays a summarized representation of the parameters.

Table 2: A realization of the estimation of $\boldsymbol{\rho}$ (top), $\boldsymbol{\pi}$ (left) and $\boldsymbol{\alpha}$ (bottom right) for the Japanese data set.

	0.291	0.264	0.299	0.147

0.329	0.845	0.952	0.995	0.987
0.342	0.414	0.736	0.919	0.916
0.329	0.259	0.37	0.388	0.738

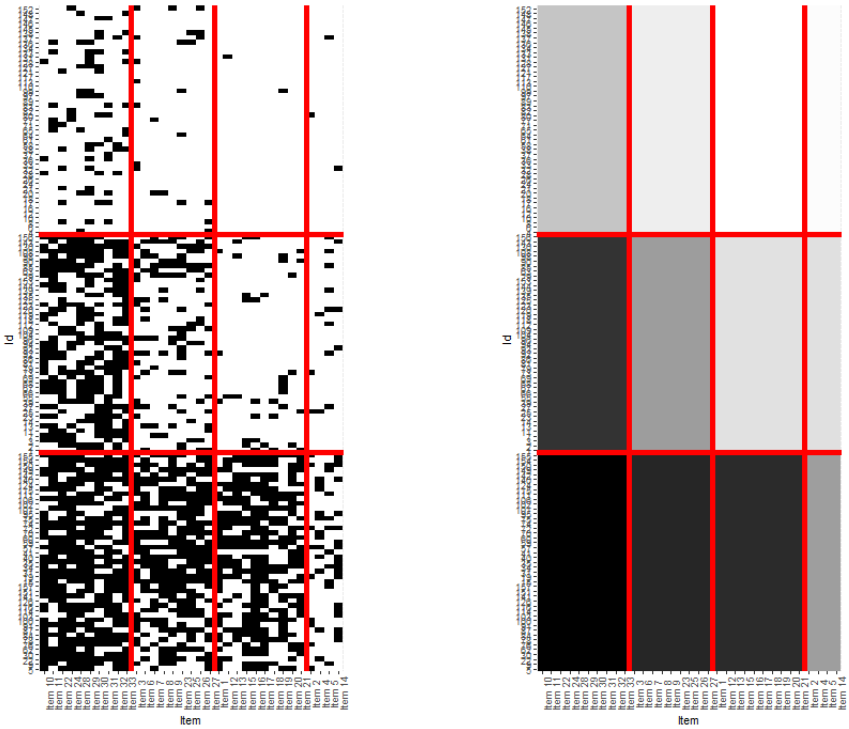

Figure 2: Data representation after the rows and columns have been ordered by classes (left) and estimated model representation (right) for the Japanese SELF test data set. The closer the $\hat{\alpha}_{k\ell}$ value is to 1, the whiter the block.

4.2.2. The English SELF placement test data set

For the English SELF test data set, the obtained reference model is $(4, 5)$. This reference model has been obtained only 16 times over the $K = 170\,000$ runs. Thus, we observe 16 inter-arrival times. Table 3 displays the distribution of these 16 values. They vary between 700 and 36 345, with a median value equal to 6595.5. This means that we should run at least 6600 initializations to obtain the reference model with probability 1/2.

Table 4 gives one realization of the estimation of each parameter and Figure 3 displays a summarized representation of the parameters.

4.3. Discussion

The simulation study carried out previously shows that the results can be unstable depending on the difficulty of the considered case. That is why, in practice, we encourage to first run the model selection procedure with

Table 3: Distribution of the inter-arrival times of the reference model $(4, 5)$ for the English SELF dataset.

Max.	36345
3rd Qu.	13398.5
Mean	10534.125
Median	6595.5
1st Qu.	4533.75
Min.	700

Table 4: A realization of the estimation of $\boldsymbol{\rho}$ (on top), $\boldsymbol{\pi}$ (on left) and $\boldsymbol{\alpha}$ (on bottom right) for the English SELF test data set.

	0.137	0.235	0.256	0.117	0.254
0.145	0.641	0.779	0.905	0.784	0.958
0.106	0.0303	0.664	0.931	0.766	0.922
0.317	0.257	0.476	0.637	0.724	0.816
0.432	0.0334	0.295	0.424	0.556	0.538

only one initialization in the estimation algorithm ($T = 1$) and then examine the estimated values $\widehat{\alpha}_{kl}$ of the parameters α_{kl} in the selected model. If the $\widehat{\alpha}_{kl}$ matrix lines and columns have different profiles, we can conclude that the case is quite simple and there is no need to increase the number of initializations in the estimation algorithm. This is for instance the case in the Japanese SELF placement test data set (see Figure 2). On the contrary, when the profiles are similar, it would be relevant to increase the number of initializations in order to stabilize the procedure. As an illustration, one can see in Figure 3 that columns 3 and 5 show quite similar profiles which may explain the selection model difficulties. To determine the relevant number of initializations, a possibility could be to run a simulation study mimicking the real data set from the estimated parameters in order to examine the results stability with respect to T.

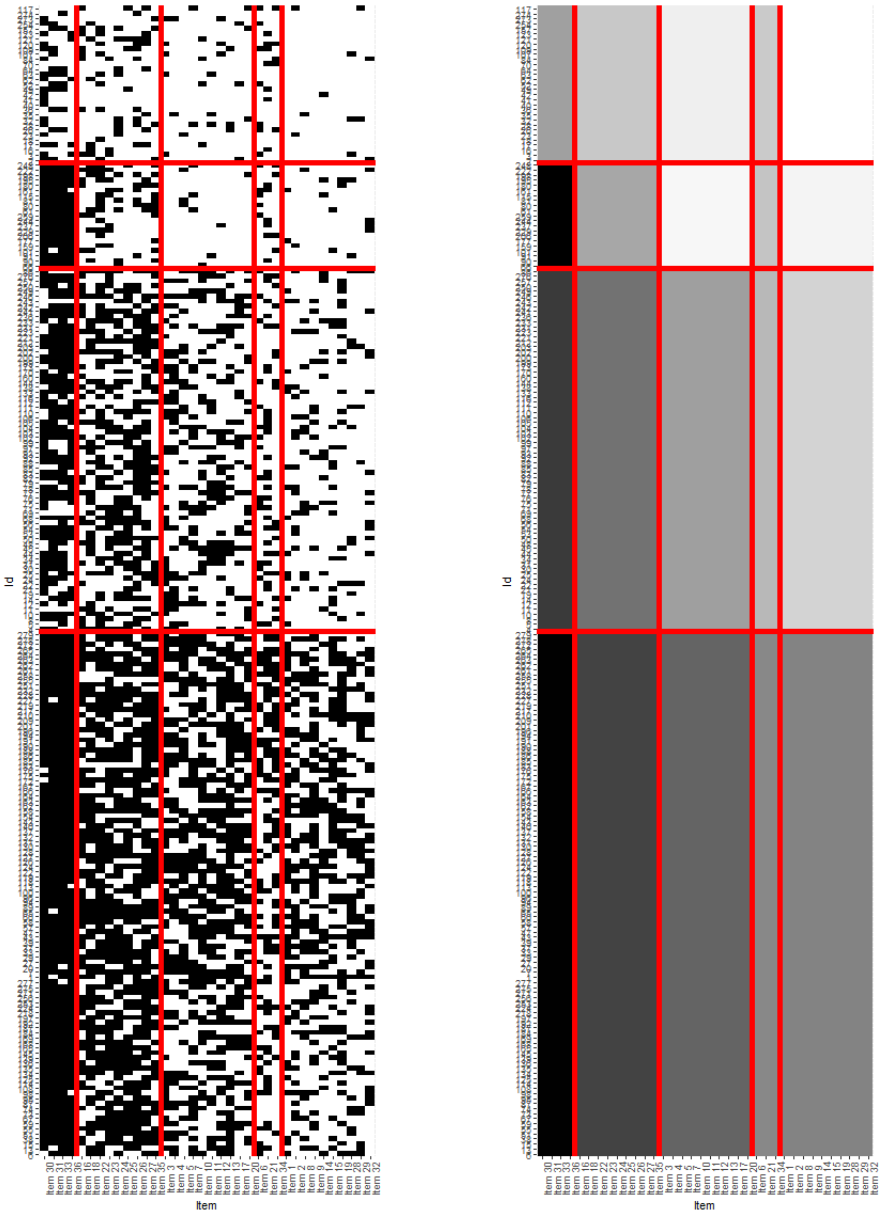

Figure 3: Data representation after the rows and columns have been ordered by classes (left) and estimated model representation (right) for the English SELF test data set. The closer the $\hat{\alpha}_{k\ell}$ value is to 1, the whiter the block.

5. Robustness

This section is devoted to a robustness study of the proposed model selection procedure in the following senses:

- the number of students groups with respect to the sample size,
- the belonging of two given students to a same group with respect to the sample size.

5.1. Sampling plan and indicators

To explore the robustness of the proposed model selection procedure, we simulate $L = 100$ data sets from the simulation plan described in Section 4.1.1 with $\varepsilon = 0.15$, 0.2 and 0.25. For $L = 1$ to 100, we run the algorithm once and we check that $(\hat{g}, \hat{m}) = (3, 4)$. If this is not the case, we simulate another data set. While running the algorithm, we get the estimated students groups proportions $\widehat{\pi}_1$, $\widehat{\pi}_2$ and $\widehat{\pi}_3$.

For a given students sample size ($n = 20, 40, 60, \ldots, 120$), we draw 10 students samples from the 137 students respecting the $\widehat{\pi}_1$, $\widehat{\pi}_2$ and $\widehat{\pi}_3$ proportions and we apply our procedure to these 10 samples. We display in Tables 6, 7 and 8, for each value of ε, the distribution of the $(\widehat{g}, \widehat{m})$ pairs selected by the proposed model selection procedure with respect to n over the 100×10 samples.

In a second step, we compare the n-students partition with the initial 137-students partition. For that purpose, following [14] extending [15],

- when the selected number of students groups is equal to $g = 3$, we draw the contingency table of the students belonging groups in the reference model and in the selected model. Students on the diagonal are defined as well classified whereas the students out of the diagonal are defined as misclassified. As illustrated in Table 5, we consider all possible labels switching and we keep the one that gives the smallest misclassified students number.
- when the selected number of students groups is greater than $g = 3$, we consider all possible groups unions of the selected model in order to get only 3 groups and we keep the group union that gives the smaller misclassified students number,
- when the selected number of students groups is smaller than $g = 3$, we consider all possible groups unions of the reference model and we proceed as previously.

Table 5: Example of label switching ([14]). In the right table, the number of misclassified students is equal to 13. By switching groups $\widehat{G2}$ and $\widehat{G3}$, we obtained a lower number of misclassified students.

Ref	$\widehat{G1}$	$\widehat{G2}$	$\widehat{G3}$	Total
G1	6	1	1	8
G2	0	1	6	7
G3	0	5	0	5
Total	6	7	7	20

Ref	$\widehat{G1}$	$\widehat{G3}$	$\widehat{G2}$	Total
G1	6	1	1	8
G2	0	6	1	7
G3	0	0	5	5
Total	6	7	7	20

Tables 6, 7 and 8 display the distribution of $(\widehat{g}, \widehat{m})$ pairs selected by the proposed procedure with respect to ε and n. We can observe that, for $\varepsilon = 0.15$, the distribution of $(\widehat{g}, \widehat{m})$ is very well concentrated on the reference pair $(3, 4)$, even when the number of students n is small. As ε increases, the distribution is more scattered, mainly for small n values. Nevertheless, for $\varepsilon = 0.25$, when n increases, we retrieve a concentrated distribution around the reference pair.

In Figures 4, 5 and 6, we focus on the numbers of students groups. The misclassified students rate distribution is displayed with respect to the number of students n and the selected number $\widehat{g}$ of students groups. Note that this rate cannot be greater than $(g-1)/g$ ([14]). When $\widehat{g} = 3$ (the reference number of students groups, blue boxplots), the misclassified students rate decreases when n increases. When $\widehat{g} = 2$ (green boxplots), the rate of misclassified students does not tend to decrease but this value of $\widehat{g}$ is less and less selected. Note that values 1 and 4 are very rarely selected, whatever the values of n and ε.

Table 6: Distribution of the $(\widehat{g}, \widehat{m})$ pairs selected by the proposed procedure with respect to n for $\varepsilon = 0.15$. The boxed data corresponds to the reference pair.

$\varepsilon = 0.15$

$n = 20$

g \ m	3	4	5	6	Total
2	190	13	0	0	203
3	45	**740**	2	0	787
4	9	1	0	0	10

$n = 80$

g \ m	3	4	5	6	Total
2	0	0	0	0	0
3	0	**990**	10	0	1000
4	0	0	0	0	0

$n = 40$

g \ m	3	4	5	6	Total
2	18	10	0	0	28
3	0	**953**	10	0	963
4	0	9	0	0	9

$n = 100$

g \ m	3	4	5	6	Total
2	0	0	0	0	0
3	0	**998**	1	0	999
4	0	1	0	0	1

$n = 60$

g \ m	3	4	5	6	Total
2	0	10	0	0	10
3	0	**979**	11	0	990
4	0	0	0	0	0

$n = 120$

g \ m	3	4	5	6	Total
2	0	0	0	0	0
3	0	**990**	1	9	1000
4	0	0	0	0	0

Table 7: Distribution of the $(\widehat{g}, \widehat{m})$ pairs selected by the proposed procedure with respect to n for $\varepsilon = 0.20$. The boxed data corresponds to the reference pair.

$\varepsilon = 0.20$

$n = 20$

g \ m	2	3	4	5	Total
1	28	1	0	0	29
2	52	485	67	0	604
3	1	108	**239**	0	348
4	0	19	0	0	19

$n = 80$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	0	30	0	30
3	0	0	**960**	10	970
4	0	0	0	0	0

$n = 40$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	1	119	79	0	199
3	0	9	**782**	9	800
4	0	0	1	0	1

$n = 100$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	0	1	0	1
3	0	0	**980**	19	999
4	0	0	0	0	0

$n = 60$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	11	49	9	69
3	0	0	**928**	3	931
4	0	0	0	0	0

$n = 120$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	**989**	10	999
4	0	0	1	0	1

Table 8: Distribution of the $(\widehat{g}, \widehat{m})$ pairs selected by the proposed procedure with respect to n for $\varepsilon = 0.25$. The boxed data corresponds to the reference pair.

$\varepsilon = 0.25$

$n = 20$

g \ m	2	3	4	5	Total
1	228	9	0	0	237
2	185	440	10	0	635
3	10	104	**12**	0	126
4	0	0	2	0	2

$n = 80$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	142	137	9	288
3	0	0	**674**	38	712
4	0	0	0	0	0

$n = 40$

g \ m	2	3	4	5	Total
1	19	11	0	0	30
2	10	579	97	0	686
3	0	21	**245**	18	284
4	0	0	0	0	0

$n = 100$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	9	80	0	89
3	0	0	**891**	20	911
4	0	0	0	0	0

$n = 60$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	217	109	0	326
3	0	0	**606**	57	663
4	0	0	11	0	11

$n = 120$

g \ m	2	3	4	5	Total
1	0	0	0	0	0
2	0	0	58	9	67
3	0	0	**924**	9	933
4	0	0	0	0	0

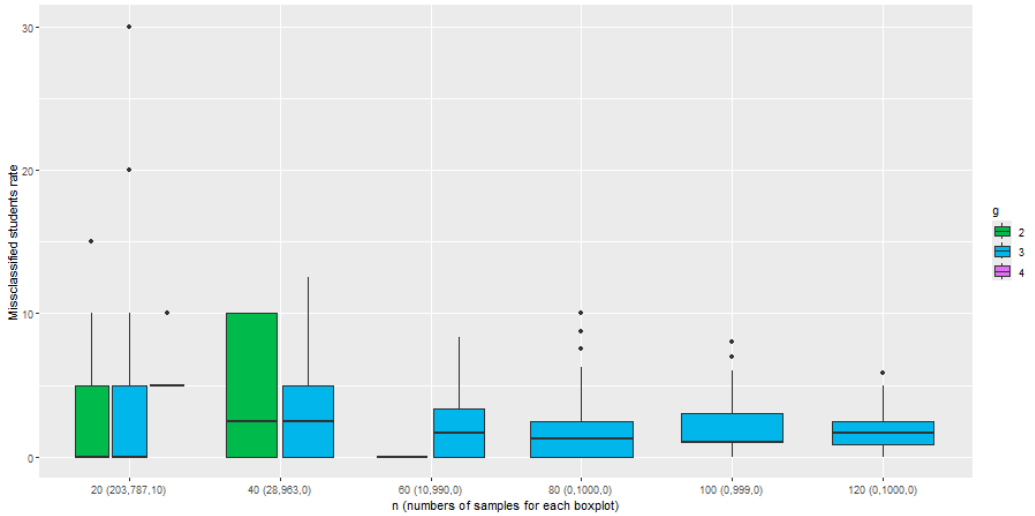

Figure 4: Misclassified students rate distribution with respect to the number n of students and the selected number $\widehat{g}$ of students groups for $\varepsilon = 0.15$. Each color corresponds to a given selected number of students groups ($\widehat{g} = 2, 3, 4$). The sample size n is followed by the numbers of samples for each boxplot, when these numbers are greater or equal to 10. For example, the first three boxplots correspond to $n = 20$. The left (resp. middle and right) one is drawn from the 203 (resp. 787 and 10) samples from which $\widehat{g} = 2$ (resp. 3 and 4) groups have been selected.

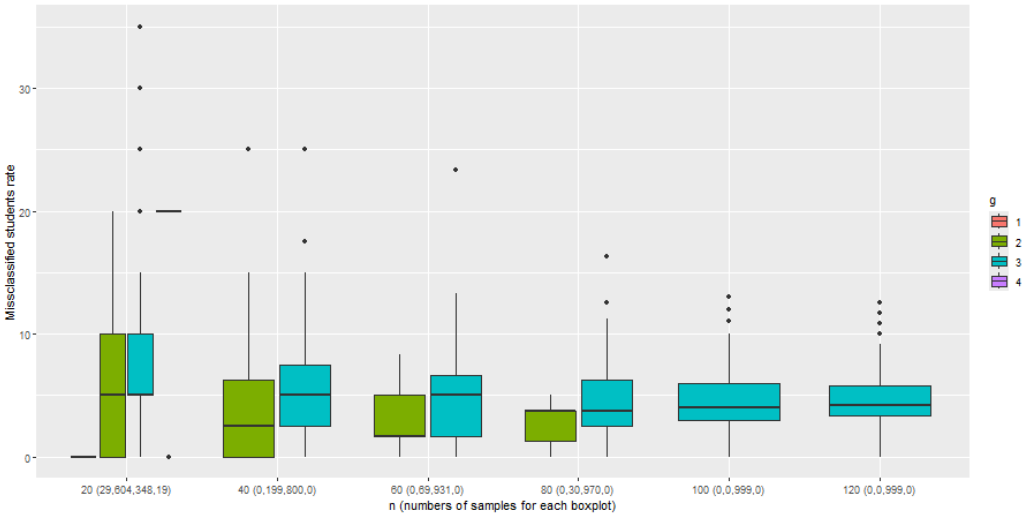

Figure 5: Misclassified students rate distribution with respect to the number n of students and the selected number $\widehat{g}$ of students groups for $\varepsilon = 0.2$. Each color corresponds to a given selected number of students groups ($\widehat{g} = 1, 2, 3, 4$). The sample size n is followed by the numbers of samples for each boxplot, when these numbers are greater or equal to 10. For example, the first four boxplots correspond to $n = 20$. The red (resp. green, blue and purple) one is drawn from the 29 (resp. $604, 348$ and 19) samples from which $\widehat{g} = 1$ (resp. $2, 3$ and 4) groups have been selected.

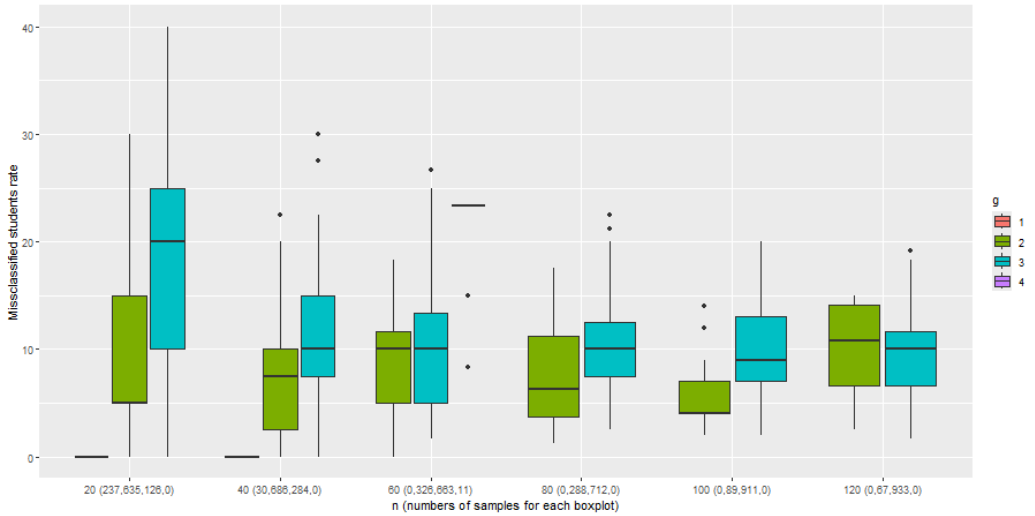

Figure 6: Misclassified students rate distribution with respect to the number n of students and the selected number $\widehat{g}$ of students groups for $\varepsilon = 0.25$. Each color corresponds to a given selected number of students groups ($\widehat{g} = 1, 2, 3, 4$). The sample size n is followed by the numbers of samples for each boxplot, when these numbers are greater or equal to 10. For example, the first four boxplots correspond to $n = 20$. The red (resp. green, blue and purple) one is drawn from the 237 (resp. 635, 126 and 2) samples from which $\widehat{g} = 1$ (resp. 2, 3 and 4) groups have been selected.

6. Conclusion

In this paper, we have proposed a model selection procedure for latent block models on binary data in the context of placement tests. We have examined the effect of initial values in the combined V-Bayes Gibbs algorithm on this procedure.

Our first results show that this effect is depending on the difficulty of the considered case. In simple cases, initial values have very small effects, which implies that only one initialization in the estimation algorithm is sufficient. On the contrary, in more difficult cases such as the English SELF test data set, initial values can have very important effects and lead to an inappropriate model selection. In such cases, we advice to strongly increase the number of estimation algorithm initializations. In the English SELF test data set, although the hyperparameter $a = 4$ is chosen in order to avoid empty classes, we have got two small probabilities values $\widehat{\rho_1} = 0.137$ and $\widehat{\rho_4} = 0.117$. With a smaller value of a, the number of initializations might have been reduced but with a higher risk of empty classes ([5]).

In a second part, we have studied the robustness of our model selection procedure by investigating the selected number of students groups and the stability of the obtained students partitions with respect to the sample size. As expected, the number of selected students groups tends to the reference one as the sample size increases. Moreover, the rate of misclassified students decreases.

In the English SELF test data set, we can observe in Table 4 one items group for which the probabilities of correct answer are similar whatever the students groups. In such a case, co-clustering methods may fail to provide a meaningful result due to the presence of noisy or irrelevant features. As a perspective, it could be relevant to consider another class of models namely the latent blocks models with noise class proposed by Laclau and Brault [16]. This class of models would enable to include a group of items for which all students would have the same probability of success regardless of the group they belong to. It would make it possible to isolate items which have small impact on the students classification and would speak in favour of the removal of these items in the placement tests.

Acknowledgment

This work has been carried out in the framework of the IRS project `COPOLangues` funded by IDEX Université Grenoble Alpes. It has also been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). All the computations presented in this paper were performed using the GRICAD infrastructure which is supported by Grenoble research communities. The authors thank Sylvain Coulange and Marie-Pierre Jouannaud for the data acquisition and Margaux Leroy for the preliminary robustness study in latent block models with application to an English SELF placement test.

References

References

[1] C. Cervini, M. Masperi, M.-P. Jouannaud, F. Scanu, Defining, modeling and piloting SELF, a new formative assessment test for foreign languages, in: J. Colpaert, M. Simons, A. Aerts, M. Oberhofer (Eds.), Language Testing in Europe: time for a new framework, University of Antwerp, 2013.

[2] V. Brault, S. Coulange, F. Letué, M.-P. Jouannaud, M.-J. Martinez, A.-C. Perret, Comment former des groupes d'étudiants homogènes à partir des résultats de SELF ? Présentation d'un outil d'aide à la décision pour la création de groupes, Mediazioni. Rivista online du studi interdisciplinari su lingue e culture 32 (2021) 185–203.

[3] G. Govaert, M. Nadif, Clustering with block mixture models, Pattern Recognition 36 (2) (2003) 463–473.

[4] G. Govaert, M. Nadif, Block clustering with Bernoulli mixture models: Comparison of different approaches, Computational Statistics and Data Analysis 52 (6) (2008) 3233–3245.

[5] C. Keribin, V. Brault, G. Celeux, G. Govaert, Estimation and selection for the latent block model on categorical data, Statistics and Computing 25 (6) (2015) 1201–1216.

[6] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em-algorithm, Journal of the Royal Statistical Society: Series B 39 (1) (1977) 1–22.

[7] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Proceedings, 2nd Internat. Symp. on Information Theory, 1973, pp. 267–281.

[8] H. Akaike, A new look at the statistical model identification, IEEE transactions on automatic control 19 (6) (1974) 716–723.

[9] G. Schwarz, et al., Estimating the dimension of a model, The annals of statistics 6 (2) (1978) 461–464.

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

[11] C. Keribin, V. Brault, G. Celeux, G. Govaert, Model selection for the binary latent block model, in: 20th International Conference on Computational Statistics (COMPSTAT 2012), Limassol, Cyprus, 2012, pp. 379–390.

[12] C. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics (1951) 79–86.

[13] V. Brault, C. Keribin, M. Mariadassou, Consistency and asymptotic normality of latent block model estimators, Electronic Journal of Statistics 14 (1) (2020) 1234–1268.

[14] V. Robert, Y. Vasseur, V. Brault, Comparing high-dimensional partitions with the Co-clustering Adjusted Rand Index, Journal of Classification 38 (2021) 158–186.

[15] A. Lomet, G. Govaert, Y. Grandvalet, Un protocole de simulation de données pour la classification croisée, in: 44e Journées de Statistique, SFdS, Bruxelles, Belgium, 2012, pp. 1–6.

[16] C. Laclau, V. Brault, Noise-free latent block model for high dimensional data, Data Mining and Knowledge Discovery 2 (2019) 446–473.