An application of multiple behavior SIA for analyzing data from student exams

Comportements multiples à l'analyse de données d'examens d'étudiants. Une application de l'ASI avec

THOMAS DELACROIX[1]
AHCENE BOUBEKKI[2]

Abstract

In this paper, we use the generalized SIA distributions developed in Delacroix (2013) with the model described in Delacroix and Boubekki (2012). The aim is to develop an analysis based on SIA theory which allows a researcher in social sciences to suppress uninteresting pseudo-implications a priori in the analysis. More precisely, we look at relations between the success of students to the different questions in an exam, while taking into account a notion of student level in a multiple behavior analysis.
Keywords: *Statistical implicative analysis, Multiple behaviors, Probability matrix, Math education, Modelling.*

Résumé
On applique, dans cet article, les distributions généralisées pour l'ASI développées dans Delacroix (2013) au modèle décrit dans Delacroix et Boubekki (2012). Il s'agit de développer une analyse de type ASI qui permette au chercheur en sciences sociales d'éliminer des pseudo-implications sans intérêt en amont de l'analyse. Plus particulièrement, on s'intéresse aux relations liant les réussites d'étudiants aux différentes questions d'un examen, en intégrant la notion de niveau d'un étudiant dans une ASI avec comportements multiples.
Mots-clés : *Analyse statistique implicative, Comportements multiples, Matrice de probabilités, Didactique des mathématiques, Modélisation.*

The notations used in this paper follow those from Delacroix (2013) and the reader is encouraged to read through that paper first. The different aspects from Delacroix and Boubekki (2012) that are used here are mostly recalled and are further developed.

Using a SIA to analyze responses to an exam

In the process of our research in Math Education, we have been looking for a way to reveal links between the different mathematical capabilities that students can have. If a student can solve a problem of type *A*, does this means he will be able to solve a problem

[1]The University of Auckland, Department of Mathematics, City-104.135, Auckland, New Zealand, maths.delacroix@gmail.com
[2]TUM School of Education, Chair of Methods in Empirical Educational Research, Arcisstrasse 21, 80333 München, Germany, ahcene.boubekki@tum.de

of type *B*? If the answer is yes, as we are dealing with human beings, we know it will not be a definite, 100%, yes. Therefore, we want to characterize positive answers and hierarchize them. For anyone who knows a little about SIA theory, these questions sound very familiar. If you could gather statistical data on whether a certain number of students, *n*, are capable of solving a certain number of problems, *p*, say by reporting information on success or failure at various questions in an exam, then this would seem to be a very good data set for a SIA. However, a problem will most surely arise while reading the results from such a SIA. It will appear that the harder questions are strongly linked together, the same being true for the easier ones. This is perfectly normal. Good students tend to be able to answer all the easier questions. Similarly, bad students are generally not capable of answering a single hard question. Though an SIA may make this obvious fact apparent, it is of little interest to a researcher in Math Education. What we really care about, is whether the fundamental nature of a problem is linked to that for another problem. Or if the necessary set of knowledge to solve one problem is contained in that same concept for a different problem. The fact that some questions are easier than others should be irrelevant. One way to palliate this problem, is to take student level and question difficulty into account in the SIA model. If we can model the fact that bad students can't manage the harder questions and good students can't fail the easier ones and if we can add this information to the SIA, then the results will not be disrupted by this trivial observation. Though this may not be possible in classical SIA theory, it can be done in multiple behavior SIA theory such as described in Delacroix (2013). For this, a probability matrix *P* corresponding to a probabilistic model for success and failure by each student to each question, based on the difficulty of each question and the level of each student, must first be defined. This is what we have started to work on in Delacroix and Boubekki (2012) and develop here. Note that if the model is well defined, it can automatically take into account another similar issue : the amount of non responses to certain questions. Indeed, if certain questions are skipped by many students, the information related to these questions in a classical SIA will be similar to that of hard questions, whether they are hard or not. This can be the case in very long exams with the questions that are at the far end of the exam, or when certain questions are not worth a lot of points. For the reader who has a specific interest in Math Education, or any educational field as a matter of fact, the construction we develop here is part of a larger perspective.

Indeed, we hope to adapt SIA theory to the construction of Learning spaces, as defined in Falmagne (2011), through the analysis of data collected from student exams.

A probabilistic model for success to questions of known difficulty by students of known level

Type of data and notations

We consider a data set consisting of information on the success (1) or failure (0) of n students to p questions of an exam. We will denote by n_i the number of students having successfully managed question i and p_j the number of questions that have been successfully managed by student j (if all the questions are worth as many points this is proportional to the student's grade). For each student j and each question i, we will define the probability $P_{i,j}$ that the student j successfully manages question i. This probability can be associated to a random variable $X_{i,j}$ as in Delacroix (2013). Furthermore, $P_{i,j}$ and $X_{i,j}$ are both considered coordinates to $p \times n$ matrices P and X. However, as described in Delacroix (2013), it is technically impossible to make calculations on a large number of students without grouping these students together. Therefore, we will be considering behavior classes of students. These classes will each correspond to a different student level as defined in the next paragraph. Hence, we refer to these as level classes. We will define q to be the number of different level classes, t_k to be the number of students in level class k, and $t_k = \frac{t_k}{n}$ the fraction of students in level class k. The probability that a student in level class k successfully manages question i will be denoted $\tilde{P}_{i,j}$.

Student level and question difficult

Before we define the model, it is important to specify the notions of student level and question difficulty that we will be using as these do not necessarily correspond to a 'real' characterization of student level or question difficulty. A question's difficulty will be characterized by the number of people who have successfully answered this question and a student's level will be the number of questions a student has answered successfully. Though it is disputable that the number of students having successfully answered a question characterizes the real difficulty of a question, it is a good enough way to characterize difficulty and it serves the purpose of the model well. For example, if an easy

question is answered by only a small number of students because it is the last question of a long exam, then it will be considered a hard question nevertheless in this model. However, as the students answering this question will generally be good students who were capable of going that far, the fraction of the population having successfully answered the question is similar to that of a hard question. Thus the model will characterize a notion that is slightly different from real question difficulty but that actually serves the purpose of the model better than real question difficulty. On a similar notice, student level, as it is defined here, does not characterize fully the real level of a given student. It only tells us how many questions this student was effectively capable of successfully answering at the time of the given exam. But this is good enough for the purpose of the model, as it will help us define the probability that this student gives a successful answer to a question. In the following, we will refer to the fractions below as question difficulty and student level relative to question *i* and student *j* in level class *k* respectively::

$$d_i = \frac{n_i}{n} \quad \text{and} \quad L_k = l_j = \frac{p_j}{p}$$

i.e. the level L_k of a class *k* is equal to the level l_j of any student *j* in the level class *k*.

Probability matrices

We will consider the following probabilistic model described in Delacroix and Boubekki (2012). The probability distribution on random variables *X* is a uniform probability conditional to knowing that:

$$\sum_{i=0}^{p} P_{i,j} = p_j \quad \text{and} \quad \sum_{j=0}^{n} P_{i,j} = n_i$$

This means we consider a uniform probability where the expected value of a student's grade will be his actual grade (if all questions are worth the same number of points) and the expected value of the number of students having successfully answered a question will be equal to the actual value for this. In other words, expected student level and question difficulty are taken to be equal to empirical student level and question difficulty.

If we denote by $\boldsymbol{C}_{n_i}(n)$ the set of n_i-combinations in a set of cardinality *n*, this gives :

$$P_{i,j} = \frac{\sum_{(j_1=j, j_2, \ldots, j_{n_i}) \in C_{n_i}(n)} \prod_{m=1}^{n_i} l_{j_m}}{\sum_{(j_1, j_2, \ldots, j_{n_i}) \in C_{n_i}(n)} \prod_{m=1}^{n_i} l_{j_m}}$$

By grouping students into level classes, we get :

$$\tilde{P}_{i,j} = \frac{\sum_{\substack{s_1+\ldots+s_q=n_i-1 \\ 0 \le s_{k'} \le t_{k'} \text{ and } s_k \le t_k - 1}} \binom{t_k - 1}{s_k} l_k^{s_k+1} \prod_{k'=1, k' \neq k}^{q} \binom{t_{k'}}{s_{k'}} l_{k'}^{s_{k'}}}{\sum_{\substack{s_1+\ldots+s_q=n_i \\ 0 \le s_{k'} \le t_{k'}}} \prod_{k'=1}^{q} \binom{t_{k'}}{s_{k'}} l_{k'}^{s_{k'}}}$$

A necessary approximation of the model

One important issue is that it can take a very long time to calculate these probability matrices. This is exactly the same type of issue as it is for 1.1.6 in Delacroix (2013). The complexity of the problem is linear in :

$$q \prod_{i=1}^{q} (t_i + 1)$$

Which, if we only consider the number of level classes and total number of students, is at most:

$$q \left(\frac{n}{q} + 1 \right)^q$$

If the number of different levels and the number of students is very large, we will need to consider an approximated model to the theoretical model, which will not take so much time to calculate. For this two solutions have been envisaged. They are both detailed and applied in the next section.

The first consists of approximating the model via a simpler model, with a smaller number of different level classes. Students with close enough level are grouped into larger level classes and the associated level for these new level classes will be the weighted average of the levels of the smaller level classes they group. This technique was tested in Delacroix and Boubekki (2012) for a variety of such approximations on the same data set as the one considered in the next section.

The second solution consists of finding a simple analytical expression to determine an approximate value for $\tilde{P}$. For the moment, we have not managed to define a satisfactory such function. We have built one which does present a certain number of necessary properties, however the justification of its construction is still mainly heuristic and the results it yields are not as good as those for the approximations. Further work on this may be pursued and is encouraged.

An evaluation of the model's adequacy

It is not our intention in this article to make a full evaluation of the presented model's adequacy to a type of study. It must be tested first in various contexts before such a study can be made. However, we can allow ourselves a few remarks. Firstly, the model satisfies usual requirements for statistical models such as identifiability and testability (see, for example, Bamber et al. (2000) for definitions of these notions). Secondly, as the model is made to explain part (the part we want to get rid of), but not all, of the observations, we don't expect it to stick perfectly to the observations. However, we do expect a better approximation of the model to stick closer to the empirical data than a lesser approximation. This is the case as we have shown in Delacroix and Boubekki (2012). Indeed, we calculated in this previous paper the distance (regular Euclidian distance) between the contingency tables for success and failure at two questions given by the empirical data and the same tables given by various approximations of the model. For a given approximation, two quantities where considered: 1) the average distance on all contingency tables and 2) the distance between the average contingency table for the empirical data and that for the approximation of the model. Both quantities decreased systematically as we tested increasingly better approximations of the model, raising the number of different level classes steadily by 1 from 3 to 12, hence contributing to evidence of the model’s adequacy.

Application

Presenting the data

The data used here was initially collected by one of the authors of this article, Thomas Delacroix, for a study in Math Education, Delacroix (2012). It was collected from the examination sheets of 104 Math students in first year at the University Paris 7. This examination on linear algebra was taken on the 2nd of March 2012 (the date on the subject

is wrong). They were assessed by the researcher using the marking-scheme indicated in the following table. For each question, a minimum number of points for success was defined. The text of the examination (in French) is added as an appendix.

Table 1. *Marking-scheme*

Question	Points	Success	Contents
Q_1	1	.75	Exercise 1 (1)
Q_2	.5	.5	Exercise 1 (2) reverse implication
Q_3	.5	.5	Exercise 1 (2) direct implication
Q_4	3	2	Exercise 2
Q_5	.5	.5	Exercise 3 (1) linearity
Q_6	.5	.5	Exercise 3 (1) inclusion
Q_7	2	1.5	Exercise 3 (2)
Q_8	1	.75	Exercise 3 (3)
Q_9	1	.75	Exercise 4 (1)
Q_{10}	2	1.5	Exercise 4 (2)
Q_{11}	1	.75	Exercise 5 (1) basis
Q_{12}	1	.75	Exercise 5 (1) equation
Q_{13}	1	.75	Exercise 5 (2) basis
Q_{14}	1	.75	Exercise 5 (2) equation
Q_{15}	.5	.5	Exercise 5 (3) belonging
Q_{16}	.5	.5	Exercise 5 (3) linear combination
Q_{17}	1	.5	Exercise 6 (1) belonging
Q_{18}	1	.5	Exercise 6 (1) independence
Q_{19}	1	.5	Exercise 6 (2)
Q_{20}	1	.5	Exercise 6 (1) first inclusion in the indication
Q_{21}	1	.5	Exercise 6 (1) second inclusion in the indication

Approximating *P*

As stated previously, calculating *P* exactly would take much too long: in Delacroix and Boubekki (2012) this was estimated to take 20 years using current home computer power!

Grouping level classes

The first approximations of *P* used here are obtained by grouping different level classes together. It can be noted that the level classes of students having succeeded in all questions, or none at all, have no influence on the complexity of the calculations for *P*. Therefore, we do not need to consider them for grouping with other level classes.

Different ways of grouping students were presented in Delacroix and Boubekki (2012). We will only consider two here which are summarized in the following table of repartition of students by level class. These correspond to a grouping of different level classes into groups of equal range (excluding the two classes at the extremities as they do not add any computational complexity to the problem and therefore may be kept as such). Recall that the level of a given grouping of level classes is defined as the weighted average of the levels of the level classes it groups. The level of each class is given out of 21.

Table 2. *Grouping into level classes*

Initial data		7-class approximation		12-class approximation	
L_k (×21)	t_k	L_k (×21)	t_k	L_k (×21)	t_k
0	2	0	2	0	2
1	4	2.6	20	1.5	8
2	4				
3	8			3.33	12
4	4				
5	5	6.72	28	5.58	12
6	7				
7	7			7.56	16
8	9				
9	9	10.37	27	9.31	13
10	4				
11	9			11.36	14
12	5				
13	6	14	14	13.4	10
14	4				
15	2			15.5	4
16	2				
17	4	18.4	10	17	4
18	0				
19	4			19.33	6
20	2				
21	3	21	3	21	3

An approximated formula

The second type of approximation used here consists of the following formula:

$$\tilde{P}_{i,k} \approx x_{i,k}^{1-f\left(\frac{p_i}{n}\right)} y_{i,k}^{f\left(\frac{p_i}{n}\right)}$$

$$\text{where} \qquad x_{i,k} = \frac{n_i}{2}\frac{L_k}{1+L_k}\mathrm{e}^{\frac{1}{2}+\frac{L_k}{1+L_k}\frac{\left(\sum_{k'=1} t_{k'}\frac{L_{k'}}{1+L_{k'}} - \frac{L_k}{1+L_k}\right)^{n_i-1}}{\left(\sum_{k'=1} t_{k'}\frac{L_{k'}}{1+L_{k'}}\right)^{n_i}}}$$

$$y_{i,k} = 1 - \frac{n-n_i}{2}\frac{1-L_k}{2-L_k}\mathrm{e}^{\frac{1}{2}+\frac{1-L_k}{2-L_k}\frac{\left(\sum_{k'=1} t_{k'}\frac{1-L_{k'}}{2-L_{k'}} - \frac{1-L_k}{2-L_k}\right)^{n-n_i-1}}{\left(\sum_{k'=1} t_{k'}\frac{1-L_{k'}}{2-L_{k'}}\right)^{n-n_i}}}$$

$$\text{and} \qquad f(t) = \frac{1-\cos(\pi x)}{2}$$

As stated previously, the justification for this formula and the formula itself must still be improved if it is to be used in any further work. We present it here to illustrate a type of approach to the problem and briefly explain how it was constructed.

We noticed that the formula for $\tilde{P}_{i,k}$ is similar to the formula for the probability in 1.1.6 in Delacroix (2013). Working with the analogy between a generalized binomial multiple behavior SIA and a generalized Poisson multiple behavior SIA, we determined a first approximation for $\tilde{P}_{i,k}$. This first approximation is $x_{i,k}$.

However, this approximation was good for low values of n_i but bad for high values. It can be shown that these two functions are equivalent when n_i goes towards *0*. As there is a symmetry in the problem considered here (between looking at success at questions or looking at failure at questions), we built another approximation which would match $\tilde{P}_{i,k}$ when n_i was close to n. This is $y_{i,k}$.

By taking a weighted geometric average between these two approximations, the weight depending on $\frac{n_i}{n}$, we obtain an approximation that is closer to $x_{i,k}$ when n_i is small and closer to $y_{i,k}$ when n_i is large. We started by taking f simply to be the identity function, but the results were closer to correct values for tests when taking a function with a more flattened curve around 0 and 1, hence the cosine. Note that these formulas have been applied to the population excluding classes where students had managed all questions successfully and where students had managed no question at all. Indeed as we already know the values of $\tilde{P}$ for these classes, the approximation is better if we do not consider

them first for the calculations and complete the matrix with the information concerning them later.

The SIAs

For these SIAs, we use the generalized Poisson distribution analysis described in 1.2.3 of Delacroix (2013). This will be done for four different values of $\tilde{P}$, the first three corresponding to the approximations previously described and the last corresponding to the unique class of a standard SIA. Note that, as this method is a generalization of SIA theory, the last case will simply correspond to a classical Poisson SIA.

For each of these four cases, we give the value of $\tilde{P}$[3] with the corresponding classes, the implicative tree and the hierarchy tree. For the implicative trees, we chose to look at a threshold of *0.98*. We then make a few crossed observations.

The calculations for the implicative and the hierarchy trees have been gracefully provided by Pablo Gregori based on code by himself and Larisa Zamora. Their program can take as an input the implicative intensity matrix, rather than the initial data, allowing us to calculate the implicative intensity using the generalized Poisson distribution analysis.

To calculate the $p \times p$ (here, p =21) implicative intensity matrix $\left(j_{a \Rightarrow b}\right)$, we start by calculating a matrix $\left(I_{a \Rightarrow b}\right)$ corresponding to the different parameters for the Poisson distributions in 1.2.3 of Delacroix (2013). This matrix is given by:

$$\left(I_{a \Rightarrow b}\right) = \tilde{P}\,\mathrm{Diag}(t)\left(J(q,p) - \tilde{P}^{T}\right)$$

Where Diag*(t)* is the diagonal matrix with diagonal coordinates $(t_1, ..., t_q)$ and *J(q,p)* is the $q \times p$ matrix with all its coordinates equal to 1.

Now, considering the matrix $\left(n_{a \Rightarrow b}\right)$, we determine $\left(j_{a \Rightarrow b}\right)$ by:

$$j_{a \Rightarrow b} = \begin{cases} 0 & \text{if } a = b \\ 1 - Pcdf\left(n_{a \Rightarrow b}, I_{a \Rightarrow b}\right) & \text{if } a \neq b \end{cases}$$

Where Pcdf$(\cdot,\lambda)$ is the cumulative distribution function of a Poisson distribution with parameter λ.

Multiple behavior SIA 7 level approximation

[3] The value given is rounded to two digits, the value used in the calculations was, of course, much more precise.

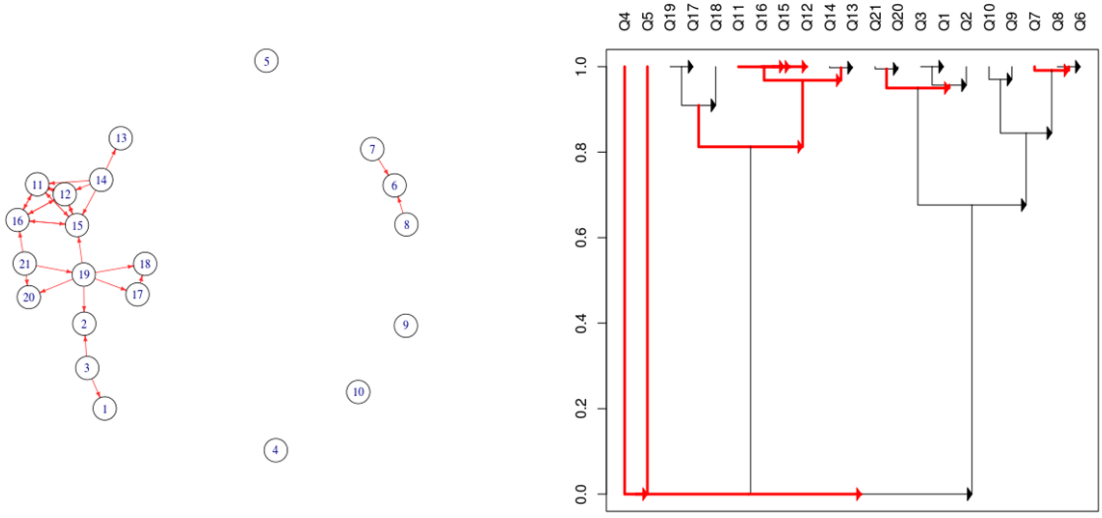
For this approximation, the matrix $\tilde{P}$ is a 21×7 matrix.

Table 3. *Probability matrix for 7 level approximation*

							Q_i	n_i
0	0.29	0.51	0.62	0.69	0.75	1	Q_1	56
0	0.27	0.49	0.6	0.67	0.73	1	Q_2	54
0	0.09	0.21	0.29	0.36	0.43	1	Q_3	27
0	0.8	0.91	0.94	0.96	0.97	1	Q_4	92
0	0.74	0.88	0.92	0.94	0.95	1	Q_5	89
0	0.23	0.44	0.55	0.62	0.68	1	Q_6	49
0	0.1	0.23	0.32	0.39	0.45	1	Q_7	29
0	0.14	0.29	0.39	0.46	0.53	1	Q_8	35
0	0.06	0.14	0.2	0.25	0.31	1	Q_9	19
0	0.04	0.09	0.14	0.18	0.23	1	Q_{10}	14
0	0.27	0.49	0.6	0.67	0.73	1	Q_{11}	54
0	0.27	0.49	0.6	0.67	0.73	1	Q_{12}	54
0	0.35	0.58	0.68	0.74	0.79	1	Q_{13}	62
0	0.2	0.39	0.5	0.58	0.64	1	Q_{14}	45
0	0.31	0.54	0.64	0.71	0.76	1	Q_{15}	58
0	0.19	0.37	0.48	0.55	0.62	1	Q_{16}	43
0	0.28	0.5	0.61	0.68	0.74	1	Q_{17}	55
0	0.31	0.54	0.64	0.71	0.76	1	Q_{18}	58
0	0.06	0.14	0.2	0.25	0.31	1	Q_{19}	19
0	0.17	0.35	0.46	0.53	0.6	1	Q_{20}	41
0	0.02	0.06	0.09	0.12	0.16	1	Q_{21}	10

0	2.6	6.71	10.37	14	18.4	21	L_k (×21)
2	20	28	27	14	10	3	t_k

Figure 1. Implicative graph and cohesion graph for 7 level approximation



Multiple behavior SIA 12 level approximation

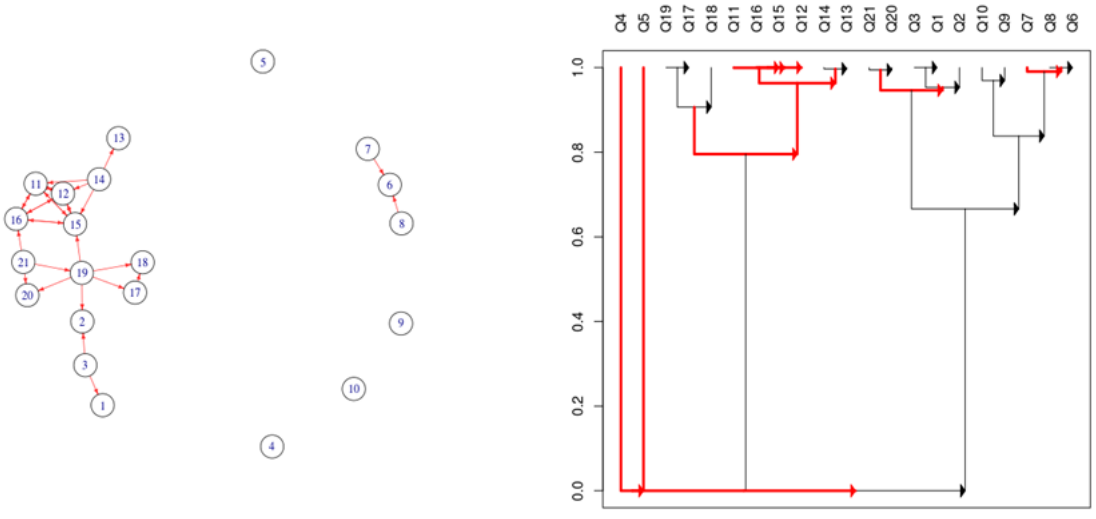
This time, the matrix $\tilde{P}$ is a 21×12 matrix.

Table 4. *Probability matrix for 12 level approximation*

											Q_i	n_i
0	0.19	0.34	0.47	0.55	0.6	0.65	0.68	0.71	0.73	0.76	1	Q_1	56
0	0.18	0.32	0.45	0.52	0.58	0.62	0.66	0.69	0.71	0.74	1	Q_2	54
0	0.06	0.12	0.18	0.23	0.27	0.31	0.35	0.39	0.41	0.44	1	Q_3	27
0	0.71	0.84	0.9	0.93	0.94	0.95	0.96	0.96	0.97	0.97	1	Q_4	92
0	0.63	0.79	0.87	0.9	0.92	0.93	0.94	0.95	0.95	0.96	1	Q_5	89
0	0.15	0.28	0.39	0.47	0.52	0.57	0.61	0.65	0.67	0.69	1	Q_6	49
0	0.06	0.13	0.2	0.25	0.29	0.34	0.38	0.41	0.43	0.47	1	Q_7	29
0	0.08	0.17	0.25	0.32	0.36	0.41	0.45	0.49	0.51	0.54	1	Q_8	35
0	0.03	0.07	0.12	0.15	0.18	0.21	0.24	0.27	0.29	0.32	1	Q_9	19
0	0.02	0.05	0.08	0.11	0.13	0.15	0.17	0.2	0.21	0.23	1	Q_{10}	14
0	0.18	0.32	0.45	0.52	0.58	0.62	0.66	0.69	0.71	0.74	1	Q_{11}	54
0	0.18	0.32	0.45	0.52	0.58	0.62	0.66	0.69	0.71	0.74	1	Q_{12}	54
0	0.24	0.41	0.54	0.61	0.66	0.71	0.74	0.77	0.78	0.8	1	Q_{13}	62
0	0.13	0.24	0.35	0.42	0.48	0.53	0.57	0.6	0.63	0.66	1	Q_{14}	45
0	0.2	0.37	0.49	0.57	0.62	0.67	0.7	0.73	0.75	0.77	1	Q_{15}	58
0	0.12	0.23	0.33	0.4	0.45	0.5	0.55	0.58	0.6	0.63	1	Q_{16}	43
0	0.18	0.33	0.46	0.54	0.59	0.63	0.67	0.7	0.72	0.75	1	Q_{17}	55
0	0.2	0.37	0.49	0.57	0.62	0.67	0.7	0.73	0.75	0.77	1	Q_{18}	58
0	0.03	0.07	0.12	0.15	0.18	0.21	0.24	0.27	0.29	0.32	1	Q_{19}	19
0	0.11	0.21	0.31	0.38	0.43	0.48	0.52	0.56	0.58	0.61	1	Q_{20}	41
0	0.01	0.03	0.05	0.07	0.08	0.1	0.12	0.13	0.14	0.16	1	Q_{21}	10

0	1.5	3.33	5.58	7.56	9.31	11.36	13.4	15.5	17	19.33	21	L_k (×21)
2	8	12	12	16	13	14	10	4	4	6	3	t_k

Figure 2. Implicative graph and cohesion graph for 12 level approximation



Multiple behavior SIA analytical approximation

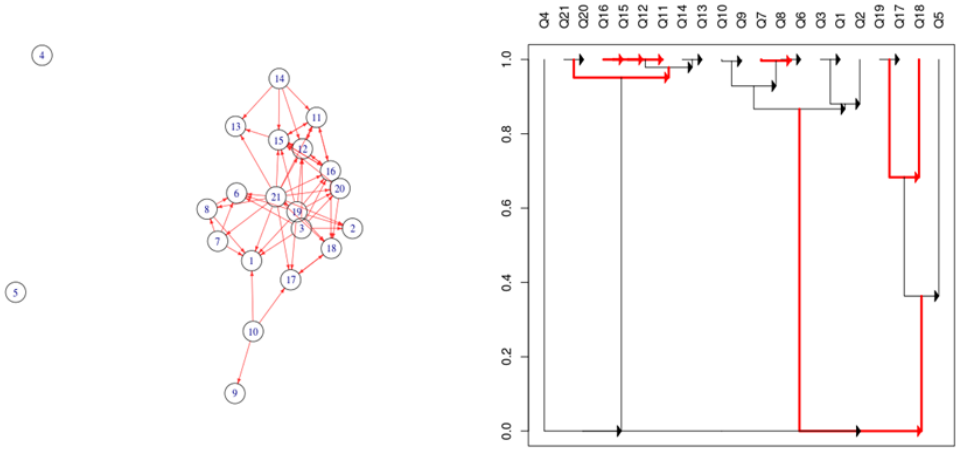
Here, $\tilde{P}$ is a 21×21 matrix.

TABLE 5. *Probability matrix for analytical approximation*

																				Q_i	n_i
0	.24	.31	.36	.4	.43	.46	.48	.5	.52	.54	.56	.58	.59	.61	.63	.65	.67	.71	.74	1	Q_1	56
0	.22	.29	.34	.38	.41	.44	.46	.48	.5	.52	.54	.56	.58	.6	.62	.63	.65	.7	.72	1	Q_2	54
0	.06	.1	.13	.16	.19	.22	.24	.26	.28	.3	.32	.34	.35	.37	.38	.4	.42	.45	.46	1	Q_3	27
0	.86	.87	.87	.88	.89	.89	.9	.9	.91	.91	.92	.93	.93	.94	.95	.95	.96	.97	.98	1	Q_4	92
0	.79	.81	.82	.83	.84	.85	.85	.86	.87	.88	.89	.89	.9	.91	.92	.93	.94	.96	.97	1	Q_5	89
0	.18	.25	.3	.34	.37	.4	.42	.44	.46	.48	.5	.52	.54	.56	.58	.6	.62	.66	.68	1	Q_6	49
0	.06	.11	.15	.18	.21	.23	.26	.28	.3	.32	.34	.36	.37	.39	.41	.42	.44	.47	.49	1	Q_7	29
0	.09	.14	.19	.22	.25	.28	.31	.33	.35	.37	.39	.41	.43	.45	.46	.48	.5	.53	.55	1	Q_8	35
0	.03	.06	.09	.11	.13	.15	.17	.19	.21	.22	.24	.25	.27	.28	.29	.31	.32	.34	.35	1	Q_9	19
0	.02	.04	.06	.08	.1	.11	.13	.14	.15	.17	.18	.19	.2	.22	.23	.24	.25	.27	.28	1	Q_{10}	14
0	.22	.29	.34	.38	.41	.44	.46	.48	.5	.52	.54	.56	.58	.6	.62	.63	.65	.7	.72	1	Q_{11}	54
0	.22	.29	.34	.38	.41	.44	.46	.48	.5	.52	.54	.56	.58	.6	.62	.63	.65	.7	.72	1	Q_{12}	54
0	.31	.38	.43	.46	.49	.51	.53	.55	.57	.59	.61	.62	.64	.66	.68	.7	.72	.76	.78	1	Q_{13}	62
0	.15	.21	.26	.3	.33	.36	.39	.41	.43	.45	.47	.49	.51	.53	.55	.56	.58	.62	.65	1	Q_{14}	45
0	.27	.34	.38	.42	.45	.47	.5	.52	.54	.56	.57	.59	.61	.63	.65	.67	.69	.73	.75	1	Q_{15}	58
0	.14	.2	.25	.29	.32	.35	.37	.4	.42	.44	.46	.48	.49	.51	.53	.55	.57	.61	.63	1	Q_{16}	43
0	.23	.3	.35	.39	.42	.45	.47	.49	.51	.53	.55	.57	.59	.6	.62	.64	.66	.71	.73	1	Q_{17}	55
0	.27	.34	.38	.42	.45	.47	.5	.52	.54	.56	.57	.59	.61	.63	.65	.67	.69	.73	.75	1	Q_{18}	58
0	.03	.06	.09	.11	.13	.15	.17	.19	.21	.22	.24	.25	.27	.28	.29	.31	.32	.34	.35	1	Q_{19}	19
0	.12	.18	.23	.27	.3	.33	.36	.38	.4	.42	.44	.46	.48	.5	.51	.53	.55	.59	.61	1	Q_{20}	41
0	.02	.03	.04	.06	.07	.08	.09	.1	.11	.12	.13	.14	.15	.16	.17	.18	.18	.2	.21	1	Q_{21}	10

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	20	21	L_k (×21)
2	4	4	8	4	5	7	7	9	9	4	9	5	6	4	2	2	4	4	2	3	t_k

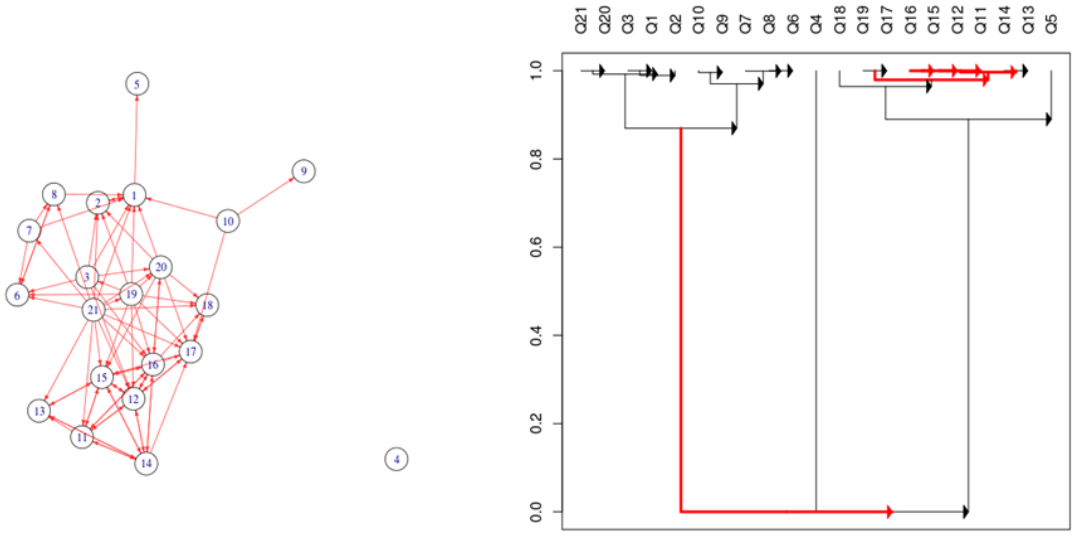
Figure 3. *Implicative graph and cohesion graph for analytical approximation*



Classical SIA

For a classical SIA, $\tilde{P}$ is a 21×1 matrix. It is quite simply given by $\tilde{P}_{i,1} = \frac{n_i}{n}$ for all i.

Figure 4. Implicative graph and cohesion graph for classical SIA



Comparing the SIAs

Without going into a detailed comparison between all these various SIAs, we can still make a certain number of observations. We base our remarks on the idea that, having removed relations of quasi-implications due to various difficulties in questions, we expect relations of quasi-implications to be concentrated between questions within a same exercise. Firstly, we consider the method to be successful in its purpose for the first two approximations of $\tilde{P}$, not so much for the analytical expression of $\tilde{P}$. This was to be expected as the analytical expression is not based on sufficient mathematical grounds. Secondly, lets us consider only the two first approximations compared to the classical SIA. The purpose of the method is to suppress parasitical pseudo-implications between variables. We see that the implicative graphs for the approximations contain much fewer arrows than the implicative graph for the classical SIA, which is basically unreadable given the number of arrows. In the cohesion graph, this translates to the fact that for the classical SIA, the arrows are crammed at the top of the graph, whereas they are more spread for the two first approximations. This shows we have indeed suppressed a certain number of pseudo-implications. The clusters are roughly the same between the cohesion

graphs for the first two approximations (for which they are exactly the same) and the cohesion graph for the classical SIA. However, there are some irregularities in the classical SIA cohesion graph that disappear in the first two cohesion graphs. Question 18, which naturally belongs with question 17 and question 19, is in a separate cluster with these two questions for the first two approximations but is only in a common cluster with these two simultaneously with all questions from exercise 5 for the classical SIA. Furthermore, in the classical SIA graph, question 5 is grouped in a larger cluster with questions from exercises 4 and 5 rather than questions from exercise 3 to which it belongs. This is due to the fact that nearly every student managed question 5. Therefore any information relating to this question should be considered highly irrelevant. This is corrected in the graphs from the first two approximations where question 5 is separated from all other questions. Thirdly, comparing the two first approximations together, the implicative graphs are identical. It is only when considering a threshold around *0.975* that a difference between the two implicative graphs appears. Furthermore, it is hard to see a difference between the two cohesion graphs. By plotting these graphs on a same grid, we notice that the graph for the 12-level approximation is slightly more spread than the graph for the 7-level approximation. This is coherent with the idea that, with a better approximation, we continue to suppress pseudo-implications, which are of no interest to us. However, this difference is very small and the calculation time for $\tilde{P}$ for the 7-level approximation was 3 seconds, whereas the calculation time for the 12-level approximation was 3 hours (see Delacroix and Boubekki (2012)). In such a case, it would therefore likely be more profitable to consider the 7-level approximation for practical reasons. Studying the convergence rate of the approximations towards $\tilde{P}$ when the number of levels increases could help us define a number of levels depending on the data for which we have a good enough approximation. To conclude, we can say that the model seems quite effective for removing parasitical pseudo-implications in certain SIA-based studies. And it is a well-known issue for researchers using SIAs that when the number of individuals considered is large (even for several hundreds) the graphs obtained are difficult to read as every variable seems to imply every other variable. Therefore, it is important for researchers to focus solely on the pseudo-implications that are interesting for their studies. The model that we have presented in this article can contribute to this and enhance SIAs in a certain number of studies.

References

DELACROIX T., (2012), É*tude d'un module "langage mathématique" en tant que module préparatoire à l'activité mathématique en algèbre linéaire de L1*, Masters thesis, Université Paris 7 Denis Diderot.

DELACROIX T., Boubekki A. (2012), *A regression analysis for taking students' levels into account in didactics studies*, preprint.

DELACROIX T. (2013), *A renewed approach to the foundations of SIA theory: Generalizing SIA to incorporate multiple behavior hypotheses; Thoughts on the implicative intensity*, submitted to the scientific committee of SIA7.

Falmagne J.C. (2011), *Learning spaces*, Springer-Verlag.

GRAS R., Bailleul M., et al. (2000), *Actes des journées sur : La fouille dans les données par la méthode d'analyse statistique implicative*, Eds Régis Gras et Marc Bailleul.

GRAS R., Kuntz P., Briand H. (2001), Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, **n° 154-155**, 9-29.

GRAS R., Suzuki E., Guillet F., Spagnolo F. (2008), *Statistical Implicative Analysis: theory and applications*, Vol. 127, Springer Verlag.

GRAS R., Régnier J.C., Marinica C., Guillet F. (2013), *L'analyse statistique implicative: Méthode exploratoire et confirmatoire à la recherche des causalités*, Ed. Cépaduès, Toulouse.

LERMAN I.C., Gras R., Rotsam H. (1981), Élaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines*, **n° 74**, 5-35, **n° 75**, 5-47.

REGNIER J.C., Bailleul M., Gras R., et al. (2012), *L'analyse statistique implicative: de l'exploratoire au confirmatoire*, Université de Caen.

BAMBER, D., van Santen, J. (2000), How to assess a model's testability and identifiability, *Journal of Mathematical Psychology*, **n° 44.1**, 20-40.

Appendix

Université Paris 7 • Licence 1, MM2 • Année 2011/2012

Partiel du 3 mars 2011.
Durée : 3 heures

Documents et matériels électroniques (dont traducteurs électroniques) interdits.
La clarté et la concision de la rédaction influencent l'appréciation de la copie.

1. EXERCICE 1 : QUESTION DE COURS *(2 points)*

Soient E et F deux espaces vectoriels, f une application linéaire de E dans F, et Ker $f = \{x \in E \,|\, f(x) = 0\}$ le noyau de f.

(1) Vérifier que Ker f est un sous-espace vectoriel de E.

(2) Vérifier que Ker $f = \{0\}$ si et seulement si f est injective.

2. EXERCICE 2 *(3 points)*

Calculer l'inverse de la matrice suivante :

$$A = \begin{pmatrix} 1 & 2 & -3 \\ 2 & -1 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

3. EXERCICE 3 *(4 points)*

On note $\mathbb{R}_2[X]$ l'espace vectoriel des polynômes à coefficients réels de degré inférieur ou égal à 2. On rappelle que les éléments de $\mathbb{R}_2[X]$ sont de la forme $a_0 + a_1X + a_2X^2$. Soit $\phi : \mathbb{R}_2[X] \to \mathbb{R}[X]$ définie par :

$$\phi(P) = 2XP - X^2P'$$

On note Im $\phi = \phi(\mathbb{R}_2[X])$ l'image de ϕ.

(1) Montrer que ϕ est linéaire, et que Im $\phi \subset \mathbb{R}_2[X]$.

(2) Déterminer Ker ϕ.

(3) Le polynôme $X^2 + X + 2$ appartient-il à Im ϕ ?

4. EXERCICE 4 *(3 points)*

Un vecteur colonne $X = (x_i) \in \mathbb{R}^n$ est dit *stochastique* si pour tous $1 \leq i \leq n$, on a $x_i \geq 0$, et $\sum_{i=1}^n x_i = 1$.

Une matrice carrée $A = (a_{i,j}) \in \mathcal{M}_n(\mathbb{R})$ est dite *stochastique* si pour tous $1 \leq j \leq n$ et $1 \leq i \leq n$, on a $a_{i,j} \geq 0$, et pour tout $j = 1..., n$, on a $\sum_{i=1}^n a_{i,j} = 1$.

(1) Montrer que si A est une matrice stochastique et X est un vecteur colonne stochastique alors AX aussi est un vecteur colonne stochastique.

(2) Trouver un vecteur $U \in \mathbb{R}^n$ non nul tel que pour toute matrice stochastique A on a ${}^tAU = U$ (U est indépendant du choix de la matrice A), où ${}^tA$ désigne la transposée de A. En déduire un vecteur stochastique V vérifiant la même propriété et donner les coefficients de V.

5. EXERCICE 5 *(5 points)*

Soit f l'application linéaire qui à tout vecteur $x \in \mathbb{R}^4$ associe $f(x) := Ax \in \mathbb{R}^3$, où la matrice A est donnée par :

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & -1 \\ 1 & -1 & 0 & 3 \end{pmatrix}$$

(1) Donner une base de l'image de f et un système d'équations cartésiennes de l'image de f.

(2) Donner une base du noyau de f et un système de deux équations cartésiennes pour le noyau de f.

(3) Applications : le vecteur colonne

$$\begin{pmatrix} -3 \\ 1 \\ -6 \end{pmatrix}$$

appartient-il à l'image de f ? Si oui, l'écrire comme combinaison linéaire des vecteurs de la base de l'image de f précédemment trouvée.

6. EXERCICE 6 *(5 points)*

Soit $l(\mathbb{R})$ l'espace vectoriel des suites à valeurs dans $\mathbb{R}$:

$$l(\mathbb{R}) := \{(u_0, u_1, u_2, \dots, u_n, \dots) \,|\, u_i \in \mathbb{R} \quad \text{pour tout } i = 0, 1, 2, \dots\}$$

On définit l'application $f : l(\mathbb{R}) \to l(\mathbb{R})$ de la manière suivante : pour toute suite $u \in l(\mathbb{R})$, $v := f(u)$ est donné par

$$v_n := n\,(u_{n+1} - u_n)$$

On admet que f est linéaire. Soit aussi

$$S := (1, 0, 0, \dots, 0, \dots,)$$
$$S' := (0, 1, 1, \dots, 1 \dots,)$$

deux suites dans $l(\mathbb{R})$.

(1) Vérifier que S et S' sont des vecteurs du noyau de f linéairement indépendants .

(2) En déduire que $Ker\, f$ a pour base $\{S, S'\}$.

(3) Soit $\mathcal{I}$ l'ensemble ds suites dans $l(\mathbb{R})$ dont la première composante est nulle :

$$\mathcal{I} = \{v \in l(\mathbb{R}) \,|\, v_0 = 0\}$$

Montrer que $\mathcal{I} = Im\, f$.
Indication : prouver d'abord que $Im\, f \subseteq \mathcal{I}$ et après que $\mathcal{I} \subseteq Im\, f$.