

Iterative multiple component analysis with a Rényi entropy-based dissimilarity measure

Vincent Vigneron^{1,2}

¹Équipe MATISSE-SAMOS CES CNRS-UMR 8173
Université Paris 1
75634 Paris cedex 13, France
vigneron@univ-paris1.fr

²IBISC CNRS FRE 2494
Université d'Evry
91020 Courcouronnes, France
vincent.vigneron@ibisc.univ-evry.fr

Abstract

In this paper, we study the notion of entropy for a set of attributes of a table and propose a novel method to measure the dissimilarity of categorical data. Experiments show that our estimation method improves the accuracy of the popular unsupervised Self Organized Map (SOM), in comparison to Euclidean or Mahalanobis distance. The distance comparison is applied for clustering of multidimensional contingency tables. Two factors make our distance function attractive: first, the general framework which can be extended to other class of problems; second, we may normalize this measure in order to obtain a coefficient similar for instance to the Pearson's coefficient of contingency.

1 Motivations

Clustering is the problem of partitioning a finite set of points in a multidimensional space into classes (called *clusters*) so that points belonging to the same class are *similar*. Measuring the (dis)similarity between data objects is one of the primary tasks for distance-based techniques in data mining and machine learning, in particular in the case of categorical data. If the data vectors contain *categorical variables*, geometric approaches are inappropriate and other strategies have to be found [10]. This is often the case in applications where the data are described by binary attributes [1, 2]. These methods transform each data object into a binary data vector, at which each bit (0 or 1) indicates the presence/absence of a positive attribute value.

Many algorithms have been designed for clustering analysis of categorical data [3, 4, 5, 6]. For instance, entropy-type metrics for *similarity* among objects have been developed from early on. SOM is a well known and quite widely used model that belongs to the unsupervised neural network category concerned with classification processes. In this paper, we focus on the metric choice for the prototype to observation distance estimation during the self-organization and exploration phases. The distance most widely used in SOM is the euclidean distance that consider each observation dimension with the same significance whatever the observation distribution inside classes. Obviously, if the data set variances are not uniformly shared out among the input dimensions, classification performances decrease. We address here the following questions: (*i*) what

class of discrepancy function admit efficient clustering algorithms ? (ii) how to visualize the classes and the explanatory variables ? For answers to (ii), see e.g. Blayo [7], Kohonen [8] or Krukal and Wish [9]. The problem corresponding to the question (i) becomes more challenging when the data is categorical, that is when there is no inherent distance measure between data objects. As a concrete example, consider a database that stores informations about physical characteristics. A sample is a tuple expressed over the attributes 'Age', 'Sex', 'Height' and 'Hair'. An instance of this database is shown in Table 1. In this setting it is not immediately obvious how to define a quality measure for the clustering. On the other hand, for humans, a good clustering is one where the clusters are *informative* about the tuples they contain, i.e. we require that the clusters be informative about the attribute values of the tuples they hold. In this case, the quality of measure of the clustering is the information that the clusters hold about the attributes. Our main contribution lies in the use of a non-euclidean metric in the learning or the exploring phase.

Age		Sex		Height		Hair		
Old	Young	Male	Female	Tall	Short	White	Brown	Blond
0	1	0	1	0	1	0	1	0
0	1	0	1	1	0	0	0	1
1	0	1	0	0	1	0	0	1
0	1	1	0	1	0	1	0	0

Table 1: An instance of the physical characteristics.

This paper is not (directly) concerned in numerical estimates of multidimensional entropy such as sample-spacings, kernel density plug-in estimates, splitting data estimates, etc.

The rest of the paper is organized as follows. Section 3 set down notations and shows the equivalence between the Rényi entropy-based dissimilarity measure and χ^2 divergence. In section 4, we investigate the proposed measure's properties and its computational complexity. Experiments with artificial data are presented in section 5. Conclusions, suggestions for drawbacks and further work are given lastly.

2 Entropy of a table of categorical data

Let J and I two finite sets indexing two categorical variables and let M be a $I \times J$ table of frequencies (Tab. 2). Let f_{ij} be the frequency (usually a integer) in the cell corresponding to the i th row and j th column of an $m \times n$ contingency table and let $f_J = \{f_{.j}\}_{j \in J}$ and $f_I = \{f_{i.}\}_{i \in I}$ be the vector of row and column marginals, i.e. the sums of elements in the i th row and j th column respectively. In the following $p_{ij} = \frac{f_{ij}}{f_0}$, $p_{i.} = \frac{f_{i.}}{f_0}$, $p_{.j} = \frac{f_{.j}}{f_0}$, where $f_0 = \sum_{i=1}^m \sum_{j=1}^n f_{ij} = \sum_{i=1}^m f_{i.} = \sum_{j=1}^n f_{.j}$.

f_{11}	f_{12}	\dots	f_{1n}	$f_{1.}$	\rightarrow	p_{11}	p_{12}	\dots	p_{1n}	$p_{1.}$
f_{21}	f_{22}	\dots	f_{2n}	$f_{2.}$		p_{21}	p_{22}	\dots	p_{2n}	$p_{2.}$
\vdots		\ddots				\vdots		\ddots		
f_{m1}	f_{m2}	\dots	f_{mn}	$f_{m.}$		p_{m1}	p_{m2}	\dots	p_{mn}	$p_{m.}$
$f_{.1}$	$f_{.2}$	\dots	$f_{.n}$	f_0		$p_{.1}$	$p_{.2}$	\dots	$p_{.n}$	1

Table 2: $m \times n$ contingency tables.

From elementary courses in statistics, we know that for any contingency table with given row and column sums, the maximum entropy value of

$$D_{12} = - \sum_i^m \sum_j^n \frac{f_{ij}}{f_0} \ln\left(\frac{f_{ij}}{f_0}\right) = \frac{1}{f_0} (f_0 \ln f_0 - \sum_i^m \sum_j^n f_{ij} \ln f_{ij}) \quad (1)$$

is obtained when $f_{ij} = \frac{f_{i \cdot} f_{\cdot j}}{f_0}$ or $p_{ij} = p_i p_j$, so that

$$\max(D_{12}) = - \sum_{i=1}^m \sum_{j=1}^n p_i \cdot p_j \ln p_i \cdot p_j = \sum_{i=1}^m p_i \ln p_i + \sum_{j=1}^n p_j \ln p_j = D_1 + D_2. \quad (2)$$

This shows that $D_{12} \leq D_1 + D_2$. The non-negative quantity $D_{12} - D_1 - D_2$ can therefore be considered as a measure of the *dependence* of the 2 attributes. Now,

$$D_{12} - D_1 - D_2 = \sum_{i=1}^m \sum_{j=1}^n p_{ij} \ln \frac{p_{ij}}{p_i \cdot p_j} \quad (3)$$

can also be interpreted in terms of Kullback-Leibler's measure of directed divergence (see section 3). Let us find its value for a small departure from independence e_{ij} . Let $p_{ij} = p_i \cdot p_j + e_{ij}$, then from (3),

$$D_1 + D_2 - D_{12} = \sum_{i=1}^m \sum_{j=1}^n p_i \cdot p_j \ln \left(1 + \frac{e_{ij}}{p_i \cdot p_j}\right) + \sum_{i=1}^m \sum_{j=1}^n e_{ij} \ln \left(1 + \frac{e_{ij}}{p_i \cdot p_j}\right) \quad (4)$$

Using Taylor's development of $\ln(1+x)$ in (4), we have:

$$D_1 + D_2 - D_{12} = \sum_{j,i} \left[e_{ij} - \frac{e_{ij}^2}{2p_i \cdot p_j} + \frac{e_{ij}^3}{3(p_i \cdot p_j)^2} \right] + \sum_{j,i} \left[\frac{e_{ij}^2}{p_i \cdot p_j} - \frac{e_{ij}^3}{2(p_i \cdot p_j)^2} \right] + \dots \quad (5)$$

where we have omitted $\sum_{j,i} \frac{e_{ij}^4}{(p_i \cdot p_j)^3}$, $\sum_{j,i} \frac{e_{ij}^5}{(p_i \cdot p_j)^4}$, \dots

Now, $\sum_{j,i} e_{ij} = \sum_{j,i} (p_{ij} - p_i \cdot p_j) = 0$, so that up to this order of approximation, (5) becomes:

$$D_1 + D_2 - D_{12} \approx \sum_{j,i} \left[\frac{e_{ij}^2}{2p_i \cdot p_j} - \frac{e_{ij}^3}{6(p_i \cdot p_j)^2} \right] = \sum_{j,i} \left[\frac{(p_{ij} - p_i \cdot p_j)^2}{2p_i \cdot p_j} - \frac{(p_{ij} - p_i \cdot p_j)^3}{6(p_i \cdot p_j)^2} \right] \quad (6)$$

In (6), as such up to a first approximation, $D_1 + D_2 - D_{12} = \sum_{j,i} \frac{(p_{ij} - p_i \cdot p_j)^2}{2p_i \cdot p_j} = \frac{1}{2} \chi^2$.

The above proof gives an interesting interpretation for the Chi-square which is now seen to represent twice the (approximated) difference between the observed and the maximum entropy. This shows that Chi-square is intimately connected with entropy maximization despite many lamentations of statisticians that Chi-square does not represent anything meaningful.

Good [11] gave a comprehensive discussion of the use of maximum entropy principle in the case of multidimensional contingency tables. Tribus [12] brought out the relationship between Chi-square test and maximization of entropy in contingency tables.

A measure of divergence (or deviation to independence) can be derived from (5) if we observe that

$$\Delta D = D_1 + D_2 - D_{12} = \sum_{i,j} \sum_k^{\infty} \frac{(-1)^k}{k(k-1)} p_i \cdot p_j \left(\frac{p_{ij} - p_i \cdot p_j}{p_i \cdot p_j} \right)^k, \quad k > 1. \quad (7)$$

Now, $d_{IJ} = \sum_{i,j} \sum_{k=1}^{\infty} p_i \cdot p_j \frac{(-1)^k x^k}{k(k-1)}$, where $\sum_{k=1}^{\infty} \frac{(-1)^k x^k}{k(k-1)}$ is the infinite series of the second derivative of the function $\phi(x) = \frac{1}{1+x}$. A primitive of ϕ is $\psi(x) = (x+1) \ln(x+1) - x$.

3 Maximum entropy and minimum Chi-square

As a whole information theory (IT) provides the necessary foundations for the statistical analysis of categorical variables. It may be used to characterize single variables (entropy) as well as group of variables (joint entropy, mutual information, conditional entropy). A major advantage of information theory is its nonparametric nature. Entropy does not require any assumptions about the distribution of variables. Consider the general class of measures of directed divergence

$$D(p||q) = \sum_{i=1}^n f(p_i, q_i) \quad (8)$$

where $p = \{p_i\}$, $q = \{q_i\}$ are probabilities sets of the same size. An important class of such measures is given by

$$D(p||q) = \sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right), q_i > 0 \quad (9)$$

where f is twice differentiable and a strictly convex function. When $f(x) = -x \ln x$, $f'(x) = 1 + \ln x$, $f''(x) = \frac{1}{x} > 0$ if $x > 0$. Accordingly, $D(p||q) = \sum_{i=1}^n q_i \frac{p_i}{q_i} \ln\left(\frac{p_i}{q_i}\right) = \sum_i p_i \ln\left(\frac{p_i}{q_i}\right)$. This is the so-called *Kullback-Leibler measure* of divergence. This measure is non-negative and vanishes iff $q_i = p_i, \forall i^1$. Table 3 shows several common discrepancy measures, in which f is twice differentiable and a strictly convex function. These functions also attain their global minimum when $p = q$.

Divergence measure	condition	Ref.
$D(p q) = \sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right)$	$q_i > 0 \forall i$	
$D(p q) = \sum_{i=1}^n (a + bq_i) f\left(\frac{a+bq_i}{a+bq_i}\right)$	$a + bq_i > 0 \forall i$	
$D(p q) = \sum_{j=1}^m \sum_{i=1}^n (a_j + bq_i) f\left(\frac{a_j+bq_i}{a_j+bq_i}\right)$	$a_j + bq_i > 0 \forall i, j$	
$D(p q) = \sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^n p_i f\left(\frac{q_i}{p_i}\right)$ or $\sum_{i=1}^n q_i \phi\left(\frac{p_i}{q_i}\right)$	$\phi(x) = f(x) + x f\left(\frac{1}{x}\right)$	

Table 3: Some classes of measures.

¹When $0 < \alpha < 1$, $\sum_{i=1}^n q_i^{1-\alpha} p_i^\alpha$ is a concave function and so its logarithm is also a concave function. We can use $\frac{1}{1-\alpha} \sum_{i=1}^n q_i^{1-\alpha} p_i^\alpha, 0 < \alpha < 1$ as a measure of discrepancy. This measure was suggested by Rényi in 1961.

4 Generalized contingency table

4.1 Notations

We consider the situation in which N individuals answer to Q questions (variables). Each question has m_q possible answers (or modalities). The individuals answer each question q ($1 \leq q \leq Q$) by choosing only one modality among the m_q modalities. If we assume that $Q = 3$ and $m_1 = 3$, $m_2 = 2$ and $m_3 = 3$, then an answer of an individual could be $(0, 1, 0|0, 1|1, 0, 0)$, where 1 corresponds to the chosen modality for each question. Let us denote by M the total number of all the modalities: $M = \sum_{q=1}^Q m_q$. To simplify, we can enumerate all the modalities from 1 to M and denote by Z_i , ($1 \leq i \leq M$) the column vector constructed by the N answers to the i -th modality. The k -th element of the vector Z_i is 1 or 0, according to the choice of the individual k . Let $K_{(N \times M)} = \{k_{ij}\}$ the complete disjunctive table where $k_{ij} = 1$ if the individual i chooses the modality j and 0 otherwise (see Tab.4). The marginals of the rows of K are constant and equal to the number Q of questions, i.e. $k_{i.} = \sum_{j=1}^M k_{ij} = Q$. K is essential if we want to remember who answered what, but if we only have to study the *relations between the Q variables* (or questions), we can sum up the data in a crosstabulations table, called *Burt matrix*, defined by $B = K^T K$, where K^T is the transposed matrix of K (see Tab.4).

m_1			m_2		m_3			
0	1	0	0	1	0	0	0	1
0	1	0	1	0	0	0	1	0
0	0	1	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1
1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	1	0	1	0	0	0
1	0	0	1	0	1	0	0	0
0	1	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	1	0	0
1	0	0	1	0	0	0	0	1

 $\rightarrow B_{(9 \times 9)} =$

4	0	1	2	2	1	0	1	2
0	5	0	2	3	0	1	3	1
0	0	3	2	1	1	2	0	0
2	2	2	6	0	1	2	1	1
2	3	1	0	6	0	1	3	2
1	0	1	2	0	2	0	1	0
0	1	2	2	1	0	3	0	0
1	3	0	1	3	0	0	4	0
2	1	0	1	2	0	0	0	3

Table 4: Left: disjunctive table $K_{(12 \times 3)}$. Right: Burt table $B_{(9 \times 9)}$ from $K_{(12 \times 3)}$.

B is a $(M \times M)$ symmetrical matrix, composed of $Q \times Q$ blocks, such that the $(q \times r)$ block B_{qr} ($1 \leq q, r \leq Q$) contains the N answers to the question r . The block B_{qq} is a diagonal matrix, whose diagonal entries are the numbers of individuals who have respectively chosen the modalities $1, \dots, m_q$ for the question q . The Burt table $B_{(M \times M)}$ has to be seen as a *generalized contingency table*, when more than 2 kinds of variables are to be studied simultaneously (see [13]). In this case, we loose a part of the information about the individuals answers, but we keep the information regarding the relations between the modalities of the qualitative variables. Each row of the matrix B characterizes a *modality of a question* (or variable). Let us denote by f_{ij} the entries of the matrix B , then the total sum of all the entries of B is $b = \sum_{i,j} b_{ij} = Q^2 N$. One defines successively (i) F the table of the relative frequencies, with entry $p_{ij} = \frac{b_{ij}}{b}$ with margins $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$, (ii) R the table of the profiles which sum to 1, with entry $R_{ij} = \frac{p_{ij}}{p_{i.}}$.

4.2 Clustering row profiles

The classical multiple correspondence analysis (MCA) ([14]) is a *weighted* principal component analysis (PCA) performed on the row profiles or column-profiles of the matrix R , each row being weighted by $p_{i\cdot}$. MCA would provide a simultaneous representation of the M vectors on a low dimensional space which gives some information about the relations between the Q variables and minimize χ^2 . In [6], Cottrell *et al.* consider the Euclidean distance between rows, each being weighten by $p_{i\cdot}$, to analyse multidimensional data, involving qualitative variables and feed a Kohonen map with these row vectors. We can do better: from (8), it comes that the distance between two rows $r(i)$ and $r(i')$ of the table R is “exactly” given by $d\{r(i), r(i')\} = \sum_k \sum_{j=1}^M \frac{(-1)^k}{(p_{\cdot j})^{k-1} k(k-1)} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}} \right)^k$. Let $x = \left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} - \frac{p_{i'j}}{p_{i'\cdot} p_{\cdot j}} \right)$.

Now, $d\{r(i), r(i')\} = \sum_j p_{\cdot j} \sum_{k=1}^{\infty} \frac{(-1)^k x^k}{k(k-1)}$, which is the infinite series of the second derivative of the function $\phi(x) = \frac{1}{1+x}$. A primitive of ϕ is $\psi(x) = (x+1) \ln(x+1) - x$. Hence, the *total deviation rate* to independence of Q categorical variables comes as above from Pearson’s approximation of independence:

$$d_Q = \sum_{i,i'} \sum_j p_{\cdot j} \{(\alpha_{ij} + 1) \ln(\alpha_{ij} + 1) - \alpha_{ij}\}, \quad (10)$$

with $\alpha_{ij} = \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} - \frac{p_{i'j}}{p_{i'\cdot} p_{\cdot j}}$. So it is equivalent to compute a profile matrix C whose entry is $c_{ij} = \frac{p_{ij}}{p_{\cdot j} p_{i\cdot}}$ and to consider the “distance” $d\{r(i), r(i')\}$ between its rows.

A remark has to be made at this stage: two modalities or more will be close if there is a large proportion of individuals that choose them simultaneously. We would like to get these individuals grouped in the same region.

5 Experiments

It is possible at this stage to use a Kohonen algorithm to get such a representation (for which there is no more constraint of linearity of the projection), as it has been already proposed by [15]. we propose to train a Kohonen network with these *row-profiles* as inputs and to study the resulting map to extract the relevant information about the relations between the Q . See [8] for further details on the Kohonen algorithm. The difference with the usual Kohonen algorithm sets in the search of the winner unit $\omega_0 = \arg \min_u \psi(\omega(u), c_i)$, where each unit u is represented in the R^M space by its *weight-vector* $\omega(u)$ and $c_i = \left(\frac{p_{1j}}{p_{\cdot j} p_{1\cdot}}, \dots, \frac{p_{Mj}}{p_{\cdot j} p_{M\cdot}} \right)$, among all the units of the lattice using the fonction ψ which rules now the metric space. ψ is now the Bregman measure to take advantage of the convexity of the criterion.

Using a black and white image of rice grains, one can illustrates a process on binary variables. The image I in Fig. 2 is a (100×256) -matrix containing only 0/1 ((pixels).

To represent the columns of I in \mathbb{R}^{256} , we train a Kohonen network with the rows of the Burt Matrix and using the Bregman divergence (see previous section). After training, each row profile can be represented by its corresponding winner unit : in Fig. 2, ‘+’ represent the pixel columns, ‘•’ the units of the Kohonen grid. To evaluate the effect of the Bregman divergence in the representation space, we plot in Fig. 3 the kernel-density estimation of the distributions of the distances between row-profiles of B , i.e. $\text{RowProfile}(i, :)$ and $\text{RowProfile}(j, :)$: Euclidean (‘-’), Citybloc (‘...’), Minkowski with $p = 4$ (‘---’) and our Bregman metric (‘·---’). Clearly, the most favourable case is the Bregman because (i) the spread of the distribution is bigger, (ii) the distribution is centered.

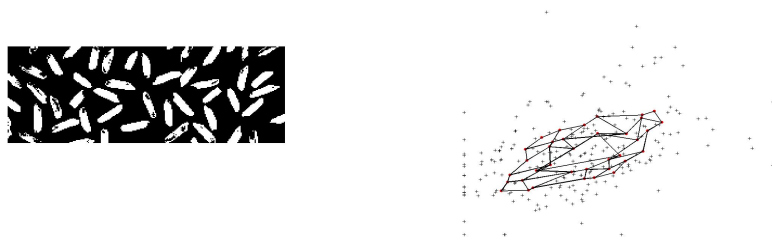


Figure 1: Left: image of rice grains. Right : Kohonen map of columns of pixel.

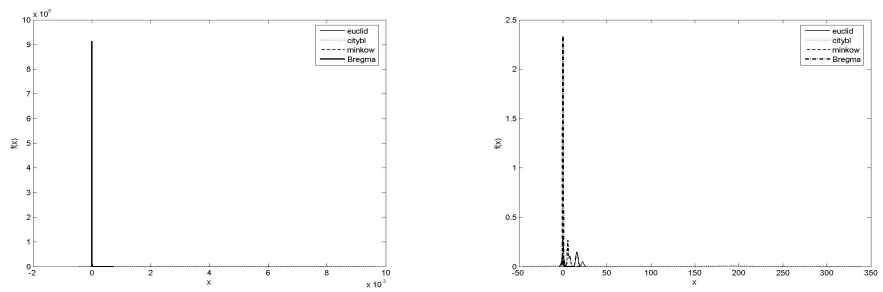


Figure 2: kernel-density estimation of the distributions of the inter row-profiles distances. Left: row-profiles of B , left : row-profiles of R .

6 Conclusion

In this paper, we derive from the entropy-based criterion for categorical data clustering a Bregman divergence measure and illustrate its relation with other criteria. The Bregman measure is used as a metric in a Kohonen algorithm to take advantage of the convexity of the criterion. The experimental results indicates the effectiveness of the proposed method. The above formulation is applicable when the data matrix directly corresponds to an empirical joint distribution. However, there are important situation in which the data matrix is more general and may contain for instance, negative entries and a distorsion measure such as the Euclidean distance might be inappropriate.

References

- [1] K. Gowda and E. Diday. Symbolic clustering using a new similarity measure. *IEEE Trans. Systems Man Cybernet.*, 22(368-378), 1992.
- [2] J. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *J. Classif.*, 3:5–86, 1986.
- [3] Z. Huang. Extension to th k -means algorithm for clustering large data sets with categorical variables. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.
- [4] S. Guha, R. Rastogi, and K. Shim. ROCK : a robust clustering algorithm for categorical attributes. *Information Systems*, 23:345–366, 2000.
- [5] V. Vigneron, H. Maaref, S. Lelandais, and A.P. Leitao. "Poor man" vote with m -ary non-parametric classifiers based on mutual information. application to iris recognition. In *4th AVBPA International Conference on Audio-Video Based Biometric Person Authentication*, London, june 2003.
- [6] M. Cottrell, P. Letremy, and E. Roy. Analysing a contingency table with kohonen maps: a factorial correspondence analysis. In J. Cabestany, J. Mary, and A. Prieto, editors, *Proceedings of IWANN'93*, Lectures Notes in Computer Science, pages 305–311. Springer, 1993.
- [7] F. Blayo and P. Demartines. Data analysis: How to compare kohonen neural networks to other technics ? In A. Prieto, editor, *Proceedings of IWANN'91*, Lectures Notes in Computer Science, pages 469–476. Springer, 1991.
- [8] T. Kohonen. *Self-organisation and Associative Memory*. Springer, 1989.
- [9] J.B. Kruskal and M. Wish. *Multidimensional scaling*. Wiley, Beverly Hills, CA, 1978.
- [10] E.B. Andersen. *Introduction to the statistical analysis of categorical data*. Springer, 1989.
- [11] I.J. Good. Maximum entropy for hypothesis formulation especially in multi-dimensional contingency tables. *Ann. Math. Stat.*, 34:911–934, 1965.
- [12] M. Tribus. *Rational descriptions, decisions, and designs*. Pergamon Press, New York, 1979.
- [13] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.
- [14] G. Saporta. *Probabilités, analyse de données et statistiques*. Technip, Paris, 1992.

- [15] S. Ibbou and M. Cottrell. Multiple correspondence analysis of a crosstabulations matrix using the kohonen algorithm. In *Proceedings of ESANN'99*. Springer, 1999.