# ITERATIVE MULTIPLE COMPONENT ANALYSIS WITH AN ENTROPY-BASED DISSIMILARITY MEASURE

Vincent Vigneron
Equipe MATISSE-SAMOS CES CNRS-UMR 8173 , Université Paris 1
IBISC CNRS FRE 2494 , Université d'Evry, 91020 Courcouronnes, France
, .

**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 if 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.
.

**Key words.** Categorical data, Self Organized Map, clustering
.
**MSC**: 62Hl7
.

**RESUMEN**
En este trabajo estudiamos la noción de entropía para un conjunto de atributos de una tabla y proponemos un novedoso método para medir la disimilitud de datos categóricos. Experimentos muestran que nuestro método de estimación mejora la acuracidad si el popular Self Organized Map (SOM) no supervisado, en comparación al las distancias Euclidiana o de Mahalanobis. La comparación de las distancias es aplicado para el clustering de tablas multidimensionales de contingencia. Dos factores hacen de nuestra función de distancia atractiva: primero, el marco de trabajo general el que puede ser extendido a otras clases de problemas; segundo, puede normalizar esta medida para obtener un coeficiente similar por ejemplo para el coeficiente de Pearson de contingencia

## 1. MOTIVATIONS

Clustering is the problem of partioning 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 [Andersen, E. B. (1989).]. This is often the case in applications where the data are described by binary attributes [Gowda and .Diday, 1992, Gower and Legendre, 1986]. 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 [Huang.(1998), Guha, Rastogi and Shim (2000), .Vigneron, Maaref, Lelandais and Leitao ( 2003), Cottrell, Letremy, Roy (1993)]. 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 pro-cesses. 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 considers 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 functions admit efficient clustering algorithms? (ii) how to visualize the classes and the explanatory variables ? For answers to (ii), see e.g. Blayo [Blayo. F.; and P. Demartines (2000), ], Kohonen [1999] or Kruskal and Wish [1978]. The problem cor-responding 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 exploration 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 (dire[1]ctly) 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 $\chi^2$ divergence. In section 4, we investigate the proposed measure properties and its computational complexity. Experiments with artificial data are presented in section 4. 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 table of frequencies (Table 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 $f_J = \left\{ f_{j.} \right\}_{j \in J}$ and $f_I = \left\{ f_{i.} \right\}_{i \in I}$ 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:

$$p_{ij} = \frac{f_{ij}}{f_0}, p_{i.} = \frac{f_{i.}}{f_0}, p_{.j} = \frac{f_{.j}}{f_0}, \text{ where } f_0 = \sum\nolimits_{i=1}^{m} \sum\nolimits_{j=1}^{n} f_{ij} = \sum\nolimits_{j=1}^{n} f_{.j} = \sum\nolimits_{i=1}^{m} f_{i.} \,.$$

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

$$D_{12} = -\sum\nolimits_{i}^{m} \sum\nolimits_{j}^{n} \frac{f_{ij}}{f} \ln\left(\frac{f_{ij}}{f_0}\right) = \frac{1}{f_0}\left( f_0 \ln f_0 - \sum\nolimits_{i}^{m} \sum\nolimits_{j}^{n} f_{ij} \ln f_{ij} \right) \tag{1}$$

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

$$\max(D_{12}) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{i.} p_{.j} \ln\left( p_{i.} p_{.j} \right) = \sum_{i=1}^{m} p_{i.} \ln p_{i.} + \sum_{j=1}^{n} p_{.j} \ln p_{.j} = D_1 + D_2 \tag{2}$$

[1]

This shows that $D_{12} \le D_1 + D_2$. 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,

$$D_{12} - D_1 - D_2 = \sum\nolimits_i^m \sum\nolimits_j^n p_{ij} \ln \frac{p_{ij}}{p_{i.}p_{.j}} \tag{3}$$

| | | | | |
|---|---|---|---|---|
| $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1n}$ | $p_{1.}$ |
| $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2n}$ | $p_{2.}$ |
| $\vdots$ | | $\ddots$ | $\vdots$ | $\vdots$ |
| $p_{m1}$ | $p_{m2}$ | $\cdots$ | $p_{mn}$ | $p_{m.}$ |
| $p_{.1}$ | $p_{.2}$ | $\cdots$ | $p_{.n}$ | $1$ |

| | | | | |
|---|---|---|---|---|
| $f_{11}$ | $f_{12}$ | $\cdots$ | $f_{1n}$ | $f_{1.}$ |
| $f_{21}$ | $f_{22}$ | $\cdots$ | $f_{2n}$ | $f_{2.}$ |
| $\vdots$ | | $\ddots$ | $\vdots$ | $\vdots$ |
| $f_{m1}$ | $f_{m2}$ | $\cdots$ | $f_{mn}$ | $f_{m.}$ |
| $f_{.1}$ | $f_{.2}$ | $\cdots$ | $f_{.n}$ | $f_0$ |

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

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 $p_{ij} = p_{i.}p_{.j} + e_{ij}$, then from (3),

$$D_1 + D_2 - D_{12} = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{i.}p_{.j} \ln\left(1 + \frac{e_{ij}}{p_{i.}p_{.j}}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij} \ln\left(1 + \frac{e_{ij}}{p_{i.}p_{.j}}\right). \tag{4}$$

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

$$D_1 + D_2 - D_{12} = \sum_{j,i}^{n,m}\left(e_{ij} - \frac{e_{ij}^2}{2p_{i.}p_{.j}} + \frac{e_{ij}^3}{3\left(p_{i.}p_{.j}\right)^2}\right) + \sum_{j,i}^{n,m}\left(\frac{e_{ij}^2}{p_{i.}p_{.j}} - \frac{e_{ij}^3}{2\left(p_{i.}p_{.j}\right)^2}\right) + \cdots \tag{5}$$

where we have omitted $\sum\nolimits_{j,i} \frac{e_{ij}^4}{\left(p_{i.}p_{.j}\right)^3}, \sum\nolimits_{j,i} \frac{e_{ij}^5}{\left(p_{i.}p_{.j}\right)^4}, \cdots$

Now as $\sum\nolimits_{j,i} e_{ij} = \sum\nolimits_{j,i}\left(p_{ij} - p_{i.}p_{.j}\right) = 0$, so that up to this order of approximation, (5) becomes:

$$D_1 + D_2 - D_{12} \approx \sum_{j,i}^{n,m}\left(\frac{e_{ij}^2}{2p_{i.}p_{.j}} - \frac{e_{ij}^3}{6\left(p_{i.}p_{.j}\right)^2}\right) = \sum_{j,i}^{n,m}\left[\frac{\left(p_{ij} - p_{i.}p_{.j}\right)^2}{2p_{i.}p_{.j}} - \frac{\left(p_{ij} - p_{i.}p_{.j}\right)^3}{6\left(p_{i.}p_{.j}\right)^2}\right] \tag{6}$$

In (6), as such upto a first approximation,

$$D_1 + D_2 - D_{12} \approx \sum_{j,i}^{n,m} \frac{\left(p_{ij} - p_{i.}p_{.j}\right)^2}{2p_{i.}p_{.j}} = \frac{1}{2}\chi^2. \tag{7}$$

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 [1965] gave a comprehensive discussion of the use of maximum entropy principle in the case of multidimensional contingency tables. Tribus [1979] 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) [Stuart, Ord and Arnold (1999)] if we observe that:

$$\Delta D = D_1 + D_2 - D_{12} = \sum_{j,i}\sum_{k}^{\infty}\frac{(-1)^k}{k(k-1)}p_{i.}p_{.j}\left(\frac{p_{ij}-p_{i.}p_{.j}}{p_{i.}p_{.j}}\right)^k \qquad \textbf{(8)}$$

Let $\left(1-\frac{p_{ij}}{p_{i.}p_{.j}}\right)^k \approx 1-k\frac{p_{ij}}{p_{i.}p_{.j}}$. From infinite series, we know that $\sum_k^{\infty}\frac{1}{k(k-1)}=1.$ It can easily be seen that

$$\Delta D \approx \sum_{j,i}\left(p_{i.}p_{.j} - p_{ij}(C-\ln K)\right), K \gg 1, \qquad \textbf{(9)}$$

where $C$ is the Euler constant[2] and $K$ is the greatest value for $k$. (9) is an other measure of divergence to independence in a contingency table but with a *penalization* term $\mathrm{C}+\ln \mathrm{K}$ on the cells.

## 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

$$D(p\,||\,q) = \sum_{i=1}^{n} f\,(p_{ij,}q_{ij}) \qquad \textbf{(10)}$$

where $p=\left\{p_{ij}\right\}, q=\left\{q_{ij}\right\}$, are probabilities sets *of the same size*. An important class of such measures is given by

$$D(p\,||\,q) = \sum_{i=1}^{n} q_{ij}\,f\,(\frac{p_{ij}}{q_{ij}}), q_{ij} > 0 \qquad \textbf{(11)}$$

where $f$ is twice differentiable and a strictly convex function. When

$$f\,(x) = -x\ln x,\; f\,'(x) = 1+\ln x,\; f\,''(x) = \frac{1}{x} > 0 \text{ if } x > 0.$$

Accordingly, $D(p\,||\,q) = \sum_{i=1}^{n} q_{ij}\frac{p_{ij}}{q_{ij}}\ln\left(\frac{p_{ij}}{q_{ij}}\right) = \sum_{i=1}^{n} p_{ij}\ln\left(\frac{p_{ij}}{q_{ij}}\right)$. This is the so-called *Kullback-Leibler* measure of divergence we use in (1). This measure is non-negative and vanishes iff $q_{ij} = p_{ij}, \forall i, j.$ Table 3

[2] Euler constant $\mathrm{C}=0{,}577215$.

shows several common discrepancy measures, in which $f$ is twice differentiable and a strictly convex function. These functions also attain their global minimum when $p = q$ .

Some special cases can be derived from (11). For instance, when $f(x) = \frac{1}{x}$ , $f'(x) = -\frac{1}{x^2}$ , $f'''x) = \frac{2}{x^2} > 0$ when $x > 0$. Then $D(p \| q) = \sum_{j,i}^{n} \frac{q_{ij}^2}{p_{ij}} = \sum_{ij}^{m} \frac{\left(q_{ij} - p_{ij}\right)^2}{p_{ij}} + 1$, for which $D\left(p \| q\right)$ is minimum when $p_{ij} = q_{ij} \forall i, j$ , and its minimum value is unity. Let $f(x) = x^{\alpha}$ , $f'(x) = -\alpha x^{\alpha-1}$ , $f''(x) = \alpha(\alpha - 1)x^{\alpha-2} > 0$ , if $\alpha > 1, x > 0$ .

| Divergence measure | conditions |
|---|---|
| $D(p \| q) = \sum_{i=1}^{n} q_{ij} f(\frac{p_{ij}}{q_{ij}})$ | $q_i > 0 \forall i$ |
| $D(p \| q) = D(p \| q) = \sum_{i=1}^{n} (a_j + b_j q_i) f\left(\frac{a_j + b_j p_i}{a_j + b_j q_i}\right)$ | $a + bq_i > 0, \forall i$ |
| $D(p \| q) = \sum_{j=1}^{m} \sum_{i=1}^{n} (a_j + b_j q_i) f\left(\frac{a_j + b_j p_i}{a_j + b_j q_i}\right)$ | $a_j + b_j q_i > 0 \forall i, j$ |
| $D(p \| q) = \sum_{i=1}^{n} q_{ij} f\left(\frac{p_{ij}}{q_{ij}}\right) + \sum_{i=1}^{n} p_{ij} f\left(\frac{q_{ij}}{p_{ij}}\right)$<br>or $D(p \| q) = \sum_{i=1}^{n} q_{ij} \phi(\frac{p_{ij}}{q_{ij}})$ | $\phi(x) = f(x) + xf\left(\frac{1}{x}\right)$ |

**Table 3.** Some classes of measures.

Then $D\left(p \| q\right) = \sum_{j,i} q_{ij} \left(\frac{p_{ij}}{q_{ij}}\right)^{\alpha} = \sum_{j,i} p_{ij}^{\alpha} q_{ij}^{1-\alpha}$ . Its minimum value occurs when $q_{ij} = p_{ij}, \forall i, j.$ So that minimum value is unity. We can use

$$\sum_{j,i} p_{ij}^{\alpha} q_{ij}^{1-\alpha} - 1 \tag{12}$$

as a measure of discrepancy when $\alpha > 1$. Havrada and Charvat [1967 suggested the measure $\frac{\sum_{j,i} p_{ij}^{\alpha} q_{ij}^{1-\alpha} - 1}{e^{\alpha-1} - 1}, \alpha > 1$. If $0 < \alpha < 1$, it is again a convex function. When $\alpha \to 1$, this approaches the limit $\sum_{j,i}^{n,m} p_{ij} \ln \frac{p_{ij}}{q_{ij}}$ which is Kullback-Leibler's measure of divergence to independence. When $0 < \alpha < 1, \sum_{j,i} p_{ij}^{\alpha} q_{ij}^{1-\alpha}$ is a positive concave function and its logarithm is also a concave function. Rényi [1961] suggested using $\frac{1}{\alpha - 1} \sum_{j,i} p_{ij}^{\alpha} q_{ij}^{1-\alpha}$ as a divergence measure. When $\alpha = -1$, (12) reduces to Pearson's Chi-square.

## 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 $q(1 \le q \le Q)$ by choosing only one modality among the $m_q$ modalities. If we assume that $Q = 3$ and $m_1 = 3, m_2 = 2, m_3 = 3,$ 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 $n$ the total number of all the modalities: $n = \sum_{q=1}^{Q} m_q$ . To simplify, we can enumerate all the modalities from 1 to $n$ and denote by $Z_i(1 \le i \le n)$ the column vector constructed by the *m* 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 $K_{(m \times n)} = \left\{ k_{ij} \right\}$ the complete disjonctive table where $k_{ij} = 1$ if the individual $i$ chooses the modality $j$ and 0 otherwise (see Table 4).

The marginals of the rows of $K$ are constant and equal to the number $Q$ of questions, *i.e.* $k_{i.} = \sum_{j=1}^{n} k_{ij} = Q$ . $K$ 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 cross tabulations table, called *Burt matrix*, defined by $B = K^T K$ , where $K^T$ is the transposed matrix of $K$ (see Table 4).

$B$ is a $n \times n$ symmetrical matrix, composed of $Q \times Q$ blocks, such th t the $(m_q \times m_r)$ block $B_{qr}\left(1 < q, r < Q\right)$ contains the $m$ 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 $1, \cdots, m_q$ for the question $q$ . The Burt table $B_{(n \times n)}$ has to be seen as a *generalized contingency table*, when more than 2 kinds of variables are to be studied simultaneously (see [Lebart, Morineau and Piron (1995)]) 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 $b_{ij}$ the entries of the matrix $B$ , then the total sum of all the entries of $B$ is $b = \sum_{j,i} b_{ij} = Q^2 N$ . 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 row-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) [Saporta (1992)] is a *weighted principal component analysis* (PCA) performed on the row profiles or column-profiles of the matrix $R$ , each row being weighted by $p_{i.}$ . MCA would provide a simultaneous representation of the $M$ vectors on a low dimensional space which gives some information about the relations between the $M$ variables and minimize Chi-square. In Cottrell, . Letremy Roy (1993) consider the Euclidean distance between rows, each being weighted by $p_{i.}$ , to analyse multidimensional data, involving qualitative variables and feed a Kohonen map with these row vectors.

| $m_1$ | | | $m_2$ | | $m_3$ | | | |
|---|---|---|---|---|---|---|---|---|
| 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 |

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

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | 0 | 0 | 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 | 0 | 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.** Upper: disjunctive table $K_{(12\times 9)}$. Lower: Burt table $B_{(9\times 9)}$ from $K_{(12\times 9)}$.

But consider the case when the distance between two rows $r(i)=\left\{\frac{p_{ij}}{p_{i.}}\right\}$ and $r(i')=\left\{\frac{p_{i'j}}{p_{i'.}}\right\}$ of the table $R$ is given from (7) by (forgetting the sign):

$$d\left\{r(i),r(i')\right\}=\sum_k^{\infty}\sum_{j=1}^{n}\frac{p_{.j}}{k(k-1)}\left(\frac{p_{ij}}{p_{i.}^{1-1/k}\,p_{.j}}-\frac{p_{i'j}}{p_{i'.}^{1-1/k}\,p_{.j}}\right)^{k}$$

This gives a family of measures that may be called Hellinger's *generalized measures of divergence* [Hellinger (1909]. This distance may be chosen so that $k=2h, h=1,2,\cdots$. So it is equivalent to compute a profile matrix $C$ whose entry is $c(i)=\left(\frac{p_{i1}}{p_{i.}^{1-1/k}\,p_{.1}},\cdots,\frac{p_{im}}{p_{i.}^{1-1/k}\,p_{.m}}\right)$ and to consider the "distance" $d\left\{c(i),c(i')\right\}$ between its rows. A remark has to be made at this stage: two modalities will be close if there are a large proportion of individuals that choose them simultaneously. We would like to get these individuals grouped in the same region. We choose the clusters centers that minimize (13). This can be done in maximizing the likelihood function. We choose the cluster vector so as to minimize (13) so that the minimum divergence estimators $\left(\theta_1,\cdots,\theta_n\right)$ are obtained by solving the equations:

$$\sum\nolimits_{j=1}^{n}\frac{1}{k-1}\left(\frac{p_{ij}}{p_{i.}^{1-1/k}\,p_{.j}}-\frac{p_{i'j}}{p_{i'.}^{1-1/k}\,p_{.j}}\right)^{k-1}\frac{\partial\,\dfrac{\theta_j}{\bar{\theta}^{1-1/k}}}{\partial\theta_j}=0 \qquad \textbf{(14)}$$

with $\bar{\theta}=\sum\nolimits_j \theta_{j.}$. These equations leads to Best Asymptotic Normal estimates for $\theta$'s. However even in simple cases, these are not easy to solve.

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 [Ibbou and Cottrell (1999)]. We propose to train a Kohonen network with {it *row-profiles*} from table $C$ as inputs and to study the resulting map to extract the relevant information about the relations between the *Q* rows.

## 5. KOHONEN ALGORITHM

Let us describe the delicate problem of learning the code vectors. The code vectors are initialized at random. At each learning step :

1 one row of the matrix $C$ is presented to the network, *i.e.*

$$c_i=\left(\frac{p_{i1}}{p_{i.}^{1-1/k}\,p_{.1}},\cdots,\frac{p_{im}}{p_{i.}^{1-1/k}\,p_{.m}}\right)^T$$ according to the probability distribution $\pi$,

2 we look then for the winner unit $\omega(u_0)$ among all the units of the lattice such that

$$u_0=\arg\min_u \|\,\omega(u)-c_i\,\| \qquad \textbf{(15)}$$

3 we update the code vectors of the unit $\omega(u_0)$ and its neighbours according to the formula

$$\omega^{(t-1)}(u)=\begin{cases}\omega^{(t)}(u)-\varepsilon(t)\left(\omega^{(t)}(u)-c_i\right), & \forall i\in\Lambda(u_0,t)\\ \omega^{(t)}(u), & \forall i\notin\Lambda(u_0,t)\end{cases} \qquad \textbf{(16)}$$

where $\Lambda(u_0,t)$ is the neighbourhood function, which depends of the number of iterations and the winning unit. More precisely, the neighbourhood function $\Lambda(u_0,t)$ and the adaptation parameter $\varepsilon(t)$ are decreasing-time functions: we begin with a large radius and we decrease it to zero. See [. Kohonen (1989)] or Cottrell, . Letremy Roy (1993) for the definitions and properties of this well-known and largely used algorithm. The adaptative parameter $\varepsilon(t)$ verifies the Robbins-Monroe conditions: $\sum\nolimits_t \varepsilon(t)=\infty$ and $\sum\nolimits_t \varepsilon^2(t)<\infty$.

These steps are repeated about 4 or 5 times over the total number of input samples. Generally, the number of modalities is not very large; the training of the network is consequently very fast. This method is very interesting from the computing time saving point of view. If we consider the case of very large data files (by examples in marketing or insurance companies), it happens that the complete disjunctive table has hundred of variables and hundred of thousands (sometimes more) individuals. The use of classical MCA can take several hours to classify the individuals into groups by a hierarchical classification. Using this method, it is sufficient to compute the Burt matrix and to train the Kohonen network with its rows.

To represent the individuals on the same Kohonen map, we proceed the following way: we use the rows of the matrix $K$. For example, the individual $j$ corresponding to the row-vector $k_j$ will be affected to the unit $u_p$ such that

$$u_p = \arg\min_u \left\| \omega^*(u) - k_j \right\| \qquad \textbf{(17)}$$

where $\omega^*(u)$ is the final value of the weight-vector $\omega(u)$ after the training step.

After training, each row profile $c_i$ can be represented by its corresponding winner unit. Because of the topology preserving property of the Kohonen algorithm, the representation of the $n$ inputs on the grid emphasizes the *proximity* between the modalities of the $Q$ questions (or variables).

**Example(Binary toy-problem)**

Let us give for instance a two colour image. This can help to illustrates a problem of binary variables. The image $I$ as depicted in Fig. 1, is the image of rice grains. $I$ can be seen as a $(100 \times 256)$ matrix containing only 0/1 (pixels) values.

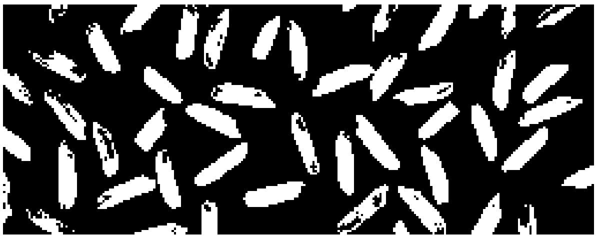

**Figure 1.** binarized image of rice grains

To give a representation of the columns of $I$, we train a Kohonen network with the 256-dimensional rows of the matrix $C$. After training, each row profile can be represented by its corresponding winner unit. The snapshot of Figure 2 confirms the effectiveness of the proposed method. Figure 2 reports the Kohonen grid state during the learning stage for $t = 1, 100, 500$ and $5000$ iterations: '+' are the pixels columns and '•' the units of the Kohonen grid, that appear linked in a 8 nodes neighborhood . We have a 256 variables problem.

A comparative study of the metric is reported to this particular data set for which an expected property could be defined. We compute the Euclidean distance between pairs of row-profiles of $n \times n$ data matrix $C$. There are different ways to compute such distance. Other distances are defined as follows: Euclidean distance $\left(c(i) - c(i')\right)^T \left(c(i) - c(i')\right)$, Mahalanobis distance $\left(c(i) - c(i')\right)^T V^{-1} \left(c(i) - c(i')\right)$, where $V$ is the sample covariance matrix. Figure 3.a plots the kernel-density estimation of the distributions of the distances between the variable row-profiles using Euclidean distance ('--'), Minkowski ('$-.-.$') in addition to our entropy-based metric with $k = 10$ ('_____'). Clearly, the latter is the most favourable metric because the support of the distribution is larger than the support of the other distribution, hence the metric is capable to take into account a larger spectrum of distances.

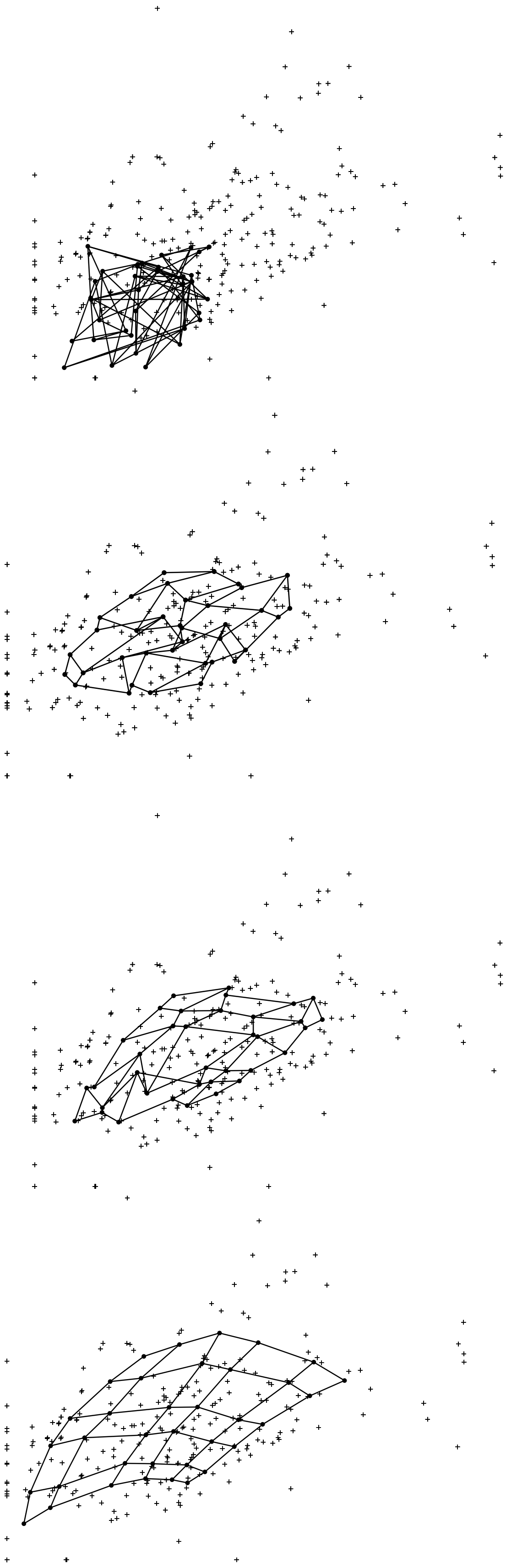

Figure **2.** Snapshots of the learning stage with Kohonen algorithm and our proposed entropy-based metric.

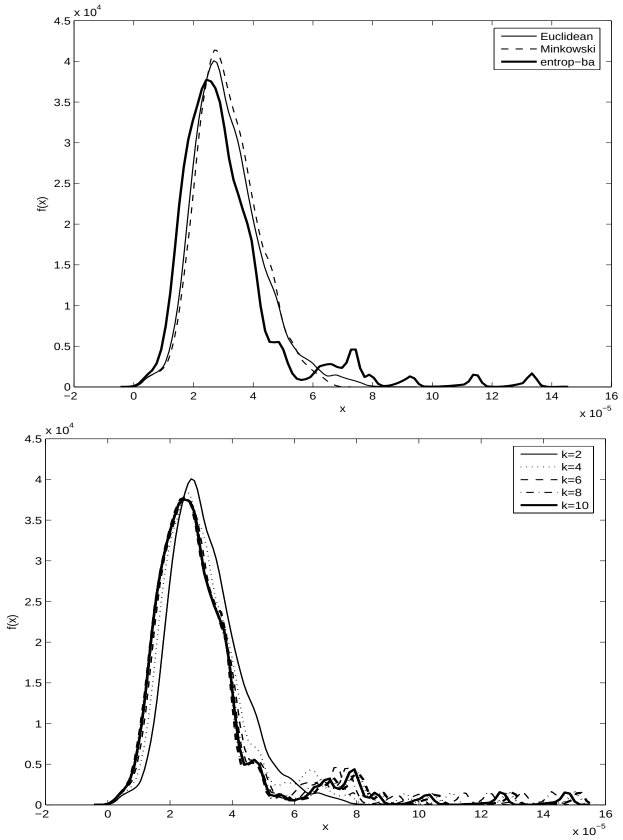

Figure . 3.. Distributions of the inter row-profiles distances. a) comparison of Euclidean, Mahalanobis and entropy-based metric ($k=10$). b) comparison of distance distributions for $k=2,4,6,8,10$ .

Note that for the special case of k=2, the metric gives the Euclidean metric.

## 6. CONCLUSION

In categorical data analysis, a key role is played by the computation of divergence measures. Many measures have been proposed it the literature, although a comparison that investigate their applicability to real data have never been reported. The main difficulty is due to the lack of a standard in the representation of categorical objects and the necessity of implementing many measures. In this work, a entropy-based discrepancy measure is used as a metric with a Kohonen algorithm to learn a set of binary variables. Interestingly, such a property has been observed only for some dissimilarity measures, which actually show very different behaviours. There are a number of possible directions for future research. One is to experiment whether other data sets with fully understandable and explainable properties are related to the proximity concept. Another direction is to extend the empirical evaluation to dissimilarity measure defined on probabilistic categorical objects. A third direction is to develop new dissimilarity measure for categorical objects that removes the two basic assumptions, namely independence and equal attribute relevance

Received March 2006
Revised February 2007

## REFERENCES

ANDERSEN, E. B. (1989). **Introduction to the statistical analysis of categorical data**. Springer.

BLAYO. F.; and P. DEMARTINES (1991).. Data analysis: How to compare Kohonen neural Networks to other techniques? In A.Prieto,editor, **Proceedings of IWANN'91, Lectures Notes in Computer Science,** 469–476.Springer.

COTTRELL, M; P. LETREMY, and E. ROY (1993). 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**, 305–311. Springer.

GOOD, I.J. (1965).. Maximum entropy for hypothesis formulation especially in multidimensional contingency tables. **Ann. Math. Stat**., 34:911–934.

GOWDA, K. and E.DIDAY (1992). Symbolic clustering using a new similarity measure. **IEEE Trans. Systems Man Cybernet.,** 22, 368-378.

GOWER, J. and P. LEGENDRE.. (1986). Metric and Euclidean properties of dissimilarity coefficients. **J. Classif**., 3:5–86,.

GUHA, S.; R. RASTOGI, and K. SHIM. (2000). ROCK: a robust clustering algorithm for categorical attributes. Information Systems, 23:345–366.

HAVRADA, F. and J.H.CHARVAt. (1967). Quantification methods of classificatory processes. Concept of structural entropy. **Kybernetica**, 3:30–35.

HELLINGER, E.. (1909). Neue begrundung der theorie quadratischer formen von unendichvicben veranderlichen. **J. für die Reine und Angew Math**., 36:210–271.

HUANG, Z. (1998). Extension to the k-means algorithm for clustering large datasets with categorical variables. **Data Mining and Knowledge Discovery**, 2:283–304.

IBBOU, S. and M.COTTRELL. (1999). Multiple correspondence analysis of a crosstabulations matrix using the Kohonen algorithm. In **Proceedings of ESANN'99.** Springer.

KOHONEN, T. (1989). **Self-organisation and Associative Memory.** Springer.

KRUSKAL. J. B. and M. WISH. (1978). **Multidimensional scaling**. Wiley, BeverlyHills, CA.

LEBART, L.; A. MORINEAU, and M. PIRON. (1995). **Statistique exploratoire multidimensionnelle**. Dunod, Paris, 1995.

RÉNYI, A.. On measures of entropy and information. **4th Berkeley Symp. Math. Stat. And Prob.,** 1:547–561, 1961.

SAPORTA, G. **Probabilités, analyse de données et statistiques.** Technip, Paris, 1992.

STUART, A; K. ORD, and S. ARNOLD. **Classical Inference and the Linear Model, volume2A of Kendalls advanced Theory of Statistics**. Arnold, sixth edition, 1999.

TRIBUS, M.. Rational descriptions, decisions, and designs. .**Pergamon Press**, New-York,1979.

VIGNERON, V; H. MAAREF, S. LELANDAIS, and A. P. LEITAO. "Poorman" 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 Authentification**, London, June 2003.