Optimal Quantization of the Support of a Discrete and Mixed Multivariate Distribution based on Mutual Information

Bernard Colin
Département de Mathématiques
Université de Sherbrooke
Sherbrooke J1K-2R1 (Québec)
Canada
bernard.colin@usherbrooke.ca

Abstract

Based on the notion of mutual information between the components of a discrete or mixed random vector, we construct, for data reduction reasons, an optimal quantization of the support of its probability measure. More precisely, we propose a simultaneous discretization of the whole set of the components of the random vector which takes into account, as much as possible, the stochastic dependence between them. Examples are presented.

Key words: Divergence, mutual information, correspondence analysis, optimal quantization, discrete and mixed multivariate random vectors.

1 Introduction

In statistics and data analysis, it is usual to take into account discrete random vectors. This is particularly the case in surveys and censuses, but also when some multidimensional discrete probabilistic models seem particularly well suited to the phenomenon under study. For example, using a descriptive and exploratory approach, some data analysis models, as correspondence analysis among others, are dedicated to highlight the stochastic links between the components of a random vector, by means of the associations between their respective categories. Similarly, the parametric framework of some discrete multidimensional distributions, leads to the estimation of the parameters of the joint distribution which will be used subsequently to quantify, for example, the stochastic dependencies between components. However in practice, one has to create for various reasons (easy use, clearness of results and graphical displays, confidentiality of the data, etc.), some classes for each component of a vector, by merging values or categories of it. For example, one often found, regarding the level of studies, some categories such as: "Primary", "Secondary", "Collegial" and "University", instead of the title of the diploma itself or the exact number of years of school. Similarly, regarding the monthly number of visits that a personne does at a given political information

website, one can find the following categories: "one visit or less", "2 to 5 visits", "6 to 10 visits", "11 to 25 visits", "26 visits and more", instead of the exact number of visits.

In practice, it is usual to create categories for each variable, regardless of the stochastic dependencies that exist between them. This process, although widespread, deprives nevertheless the statistician from information that could be crucial, for example in a predictive framework, since it degrades arbitrarily the stochastic dependence and could affect the quality of the forecast model. To alleviate this problem, we propose in what follows, to adapt to the discrete case the approach proposed in the continuous case, by Colin, Dubeau, Khreibani and de Tibeiro [10]. This one is based, in the process of data reduction, on the existence of an optimal quantization of the support of the probability measure of a continuous random vector, arising from the "minimal loss of mutual information" principle.

2 Theoretical framework

2.1 Generalities

We present briefly thereafter, the problem of finding an optimal quantization of a support of a probability measure in the context of discrete probability spaces, for which it will be assumed that reference measure will be systematically the counting measure. Let $X = (X_1, X_2, ..., X_k)$ be a random vector defined on a probability space $(\Omega, \mathcal{F}, \mu)$ and with values in a finite or countable set $\mathcal{X}$, where $\Omega = \{\omega\}$ is an arbitrary set, $\mathcal{F}$ a σ-field of subsets of Ω and μ a probability measure on $\mathcal{F}$. Generally one has: $\mathcal{X} = \times_{i=1}^{i=k} \mathcal{X}_i$ where $\mathcal{X}_i = \{X_i(\omega)\}$ is finite or countable and, in this last case, $\mathcal{X}_i$ is frequently choosen as $\mathbb{N}$, $\mathbb{N}^* = (\mathbb{N} \cup \{0\})$ or $\mathbb{Z}$. Let $\mathbb{P}$ be the probability measure on $\mathcal{P}(\mathcal{X})$ image of μ under the mapping X, given by:

$$\begin{aligned} \mathbb{P}(x_1, x_2, ..., x_k) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) \\ &= \mu\{\omega \in \Omega : X_1(\omega) = x_1, X_2(\omega) = x_2, ..., X_k(\omega) = x_k\} \end{aligned}$$

where $x = (x_1, x_2, ..., x_k)$ is a member of $\mathcal{X} = \times_{i=1}^{i=k} \mathcal{X}_i$. Finally, we note by $\mathbb{P}_{X_1}, \mathbb{P}_{X_2}, ..., \mathbb{P}_{X_k}$, the marginal probability measures of the components of X.

2.2 Mutual information

As defined in the continuous case, the mutual information $\mathcal{I}_\varphi(X_1, X_2, ..., X_k)$ between the random variables $X_1, X_2, ..., X_k$ is nothing else than the φ-divergence $I_\varphi\left(\mathbb{P}, \otimes_{i=1}^{i=k} \mathbb{P}_{X_i}\right)$ (see [13]) between the probabiliy measures $\mathbb{P}$ and $\otimes_{i=1}^{i=k} \mathbb{P}_{X_i}$ given by:

$$\begin{aligned} \mathcal{I}_\varphi(X_1, X_2, ..., X_k) &= I_\varphi\left(\mathbb{P}, \otimes_{i=1}^{i=k} \mathbb{P}_{X_i}\right) \\ &= \sum_{x \in \mathcal{X}} \left[\varphi\left(\frac{\mathbb{P}(x)}{\otimes_{i=1}^{i=k} \mathbb{P}_{X_i}(x)}\right) \otimes_{i=1}^{i=k} \mathbb{P}_{X_i}(x)\right] \\ &= \mathbb{E}^{\otimes_{i=1}^{i=k} \mathbb{P}_{X_i}}\left[\varphi\left(\frac{\mathbb{P}(X)}{\otimes_{i=1}^{i=k} \mathbb{P}_{X_i}(X)}\right)\right] \end{aligned}$$

where φ is a convex function from $\mathbb{R}_+ \backslash \{0\}$ to $\mathbb{R}$ (see Csiszár[13], Aczél and Daróczy[1], Rényi[22] for details). It is easy to check, using some elementary calculations, that all the

properties of divergence and mutual information, as set out in the continuous framework, are also valid in a discrete setting and, in particular, the one relating to the loss of information arising from a transformation of the random vector X, property known as: "*data-processing theorem*" (see [1], [12], [13], [26]).

2.3 Optimal partition

2.3.1 Mutual information explained by a partition

Without loss of generality, and for sake of brevity, it is assumed that each component X_i of the random vector X is, for $i = 1, 2, ..., k$, a random variable with values in $\mathbb{N}^*$. In addition, for all $i = 1, 2, ..., k$, we denote by $\eta_{il_i} \in \mathbb{N}^*$ any integer value of the random variable X_i, where $l_i \in \mathbb{N}^*$. Given k integers $n_1, n_2, ..., n_k$, we consider for every $i = 1, 2, ..., k$, a partition $\mathcal{P}_i$ of the support $S_{\mathbb{P}_{X_i}}$ of the random variable X_i, obtained by means of a set $\{\gamma_{ij_i}\}$ of n_i intervals of the following form:

$$\gamma_{ij_i} = [x_{i(j_i-1)}, x_{ij_i}[\text{ for } j_i = 1, 2, ..., n_i - 1 \text{ and } \gamma_{in_i} = [x_{i(n_i-1)}, \infty[$$

where the bounds of the intervals are given real numbers such that:

$$0 = x_{i0} < x_{i1} < x_{i2} < ... < x_{i(n_i-1)} < \infty$$

Remark: the choice for real bounds for the intervals of $\mathcal{P}_i$, is a simple consequence of the fact that it is not excluded that *a priori*, one of the elements of the optimum partition, may be a single point x of $\mathbb{N}^{*^k}$. Since the intervals are right closed and left open, such an event could possibly not occur if the bounds were some integers values of the components $X_1, X_2, ..., X_k$ of X.

The "cartesian product partition" or more simply the "product partition" $\mathcal{P}$ of the support $S_{\mathbb{P}}$ in $n = n_1 \times n_2 \times ... \times n_k$ elements or cells is then defined by:

$$\mathcal{P} = \otimes_{i=1}^{i=k} \mathcal{P}_i = \{ \times_{i=1}^{i=k} \gamma_{ij_i} \}$$

where $j_i = 1, 2, ..., n_i$, for every $i = 1, 2, ..., k$.

If $\sigma(\mathcal{P})$ is the σ-algebra generated by $\mathcal{P}$ (more precisely, the algebra generated by $\mathcal{P}$ in this case), the restriction of $\mathbb{P}$ to $\sigma(\mathcal{P})$, is given by:

$$\mathbb{P}\left(\times_{i=1}^{i=k} \gamma_{ij_i}\right) \quad \text{for every } j_1, j_2, ..., j_k$$

and for which it is easy to check that the marginal probability distributions $\mathbb{P}_{X_i}$, for $i = 1, 2, ..., k$, are given by:

$$\mathbb{P}_{X_i}\left(\gamma_{ij_i}\right) \quad \text{where } j_i = 1, 2, ..., n_i$$

The mutual information, noted by $\mathcal{I}_\varphi(\mathcal{P})$, explained by the partition $\mathcal{P}$ of $S_{\mathbb{P}}$ is then express as:

$$\mathcal{I}_\varphi(\mathcal{P}) = \sum\nolimits_{j_1, j_2, ..., j_k} \varphi\left(\frac{\mathbb{P}\left(\times_{i=1}^{i=k} \gamma_{ij_i}\right)}{\prod_{i=1}^{i=k} \mathbb{P}_{X_i}\left(\gamma_{ij_i}\right)}\right) \prod_{i=1}^{i=k} \mathbb{P}_{X_i}\left(\gamma_{ij_i}\right)$$

It follows that the mutual information loss, arising from the data reduction, is given by:

$$\mathcal{I}_\varphi\left(X_1, X_2, ..., X_k\right) - \mathcal{I}_\varphi(\mathcal{P})$$

which is, as a consequence of the "*data-processing theorem*", positive

2.3.2 Existence of an optimal partition

For every sequence of given integers $n_1, n_2, ..., n_k$ and for every $i = 1, 2, ..., k$, let $\mathcal{P}_{i,n_i}$ be the class of partitions of $S_{\mathbb{P}_{X_i}}$ in n_i disjoint intervals γ_{ij_i} as defined in the previous subsection, and let $\mathcal{P}_{\mathbf{n}}$ be the set of partitions of $S_{\mathbb{P}}$ given by:

$$\mathcal{P}_{\mathbf{n}} = \otimes_{i=1}^{i=k} \mathcal{P}_{i,n_i}$$

where $\mathbf{n}$ is the multi-index $(n_1, n_2, ..., n_k)$ of size $|\mathbf{n}| = k$.

The problem of finding a member $\mathcal{P}$ of $\mathcal{P}_{\mathbf{n}}$ for which the mutual information loss is minimal, is equivalent to find a solution of the following optimization problem:

$$\min_{\mathcal{P} \in \mathcal{P}_{\mathbf{n}}} \left(\mathcal{I}_{\varphi}(X_1, X_2, ..., X_k) - \mathcal{I}_{\varphi}(\mathcal{P}) \right)$$

which is equivalent to solve this one:

$$\max_{\mathcal{P} \in \mathcal{P}_{\mathbf{n}}} \mathcal{I}_{\varphi}(\mathcal{P}) = \max_{\mathcal{P} \in \mathcal{P}_{\mathbf{n}}} \sum\nolimits_{j_1, j_2, ..., j_k} \varphi \left(\frac{\mathbb{P}\left(\times_{i=1}^{i=k} \gamma_{ij_i} \right)}{\prod_{i=1}^{i=k} \mathbb{P}_{X_i}\left(\gamma_{ij_i} \right)} \right) \prod_{i=1}^{i=k} \mathbb{P}_{X_i}\left(\gamma_{ij_i} \right)$$

which simply consists in finding the real bounds of the intervals γ_{ij_i} for every $i = 1, 2, ..., k$ and for every $j_i = 1, 2, ..., n_i$. From this point of view, one can observe that since the probability distributions are discrete and since each component of X has values in $\mathbb{N}^* = (\mathbb{N} \cup \{0\})$, it is easy to check that the choice of a real bound x_{ij_i} in any interval of the form $\left]\eta_{il_i}, \eta_{i(l_i+1)}\right[$ is totally arbitrary for every $i = 1, 2, ..., k$ and $j_i = 1, 2, ..., n_i$. As a result, one can agree that the $\prod_{i=1}^{i=k}(n_i - 1)$ possible choices for real bounds, will be reduced to the choice of centers $\eta_{il_i} + \frac{1}{2}$ of intervals of the form $\left]\eta_{il_1}, \eta_{i(l_i+1)}\right[$.

If the support $S_{\mathbb{P}}$ is finite, one has a finite number of members of $\mathcal{P}_{\mathbf{n}}$ so an optimal partition automatically exists, while if the support $S_{\mathbb{P}}$ is countable, then the set $\mathcal{P}_{\mathbf{n}}$ is countable. It follows that the countable set of the real numbers $\mathcal{I}_{\varphi}(\mathcal{P}_{\eta})$, where $\mathcal{P}_{\eta} \in \mathcal{P}_{\mathbf{n}}$ for every $\eta \in \mathbb{N}^*$, may be ordered according to a non-decreasing sequence, bounded above by $\mathcal{I}_{\varphi}(X_1, X_2, ..., X_k)$. So in this case, the existence of an upper bound for the sequence $(\mathcal{P}_{\eta})_{\eta \geq 0}$, ensures the existence of an optimal partition or possibly the existence of a "quasi optimal" partition, if the upper bound is not reached by a member of $\mathcal{P}_{\mathbf{n}}$.

As an illustration of the previous comments, we consider thereafter, a bivariate discrete random vector $X = (X_1, X_2)$ with a finite support and whose components X_1 and X_2 take value respectively in $\{0, 1, 2, ..., p\}$ and $\{0, 1, 2, ..., q\}$. Let n_1 and n_2 be the numbers of the elements of a partition $\mathcal{P}_1$ of $\{0, 1, 2, ..., p\}$ and of a partition $\mathcal{P}_2$ of $\{0, 1, 2, ..., q\}$ with $n_1 \leq p + 1$ and $n_2 \leq q + 1$. A partition $\mathcal{P} = \mathcal{P}_1 \otimes \mathcal{P}_2$ has $n_1 \times n_2$ non-empty elements or cells and it is straightforward to check that $card(\mathcal{P}_{\mathbf{n}})$ where $\mathbf{n} = (n_1, n_2)$ is equal to:

$$card(\mathcal{P}_{\mathbf{n}}) = \binom{p}{n_1 - 1} \binom{q}{n_2 - 1}$$

It follows, as previously said, that since $card(\mathcal{P}_{\mathbf{n}})$ is finite, an optimal partition exists and this one can be found by simple inspection, at least from a theoretical point of view.

However, when the support $S_{\mathbb{P}}$ is countable and unbounded, as for example in the case of the multivariate *Poisson* distribution or in the case of the multivariate negative multinomial distribution, some difficulties to determining an optimal (or quasi-optimal) partition may occur for numerical reasons. Indeed, since there is a countable choice for the bounds of the intervals for each marginal partitions, then the set $\mathcal{P}_{\mathbf{n}}$ is countable and some convergence and stability problems may occur at the neighbourhood of the optimal solution, especially when the greatest bounds tend to infinity. As in the continuous case (see [10]), in order to bring back the optimization problem to a bounded support, it is possible to use an elementary transformation (the discrete counterpart of the continuous "*probability integral transform*") as illustrated below in the case of a bivariate discrete random vector $Z = (X, Y)$ with values in $\mathbb{N}^{*2}$.

To this end, let H, F and G, be respectively the joint and marginals cumulative distribution functions of the randon vector Z and $\mathbb{P}, \mathbb{P}_X$ and $\mathbb{P}_Y$ be the resulting joint and marginals probability density functions. The following notations are usual:

$$\mathbb{P}(X = i, Y = j) = p_{ij}, \text{ for } i, j = 0, 1, 2, ...$$

and:

$$\mathbb{P}_X(X = i) = p_{i\bullet} = \sum_{j=0}^{\infty} p_{ij}, \text{ for } i = 0, 1, 2, ...$$

$$\mathbb{P}_Y(Y = j) = p_{\bullet j} = \sum_{i=0}^{\infty} p_{ij}, \text{ for } j = 0, 1, 2, ...$$

Let x and y be two real non-negative numbers. One has:

$$H(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \sum_{i=0}^{\lfloor x \rfloor} \sum_{j=0}^{\lfloor y \rfloor} p_{ij}$$

where $\lfloor x \rfloor$ et $\lfloor y \rfloor$ are respectively the integer part of the reals numbers x and y. Similarly one has:

$$F(x) = \mathbb{P}_X(X \leq x) = \sum_{i=0}^{\lfloor x \rfloor} p_{i\bullet} = \sum_{i=0}^{\lfloor x \rfloor} \sum_{j=0}^{\infty} p_{ij}$$

and :

$$G(y) = \mathbb{P}_Y(Y \leq y) = \sum_{j=0}^{\lfloor y \rfloor} p_{\bullet j} = \sum_{j=0}^{\lfloor y \rfloor} \sum_{i=0}^{\infty} p_{ij}$$

One then considers the discrete random vector $C = (U, V)$ deduced from $Z = (X, Y)$ using the following transformation $\mathcal{T}$ from $\mathbb{N}^* \times \mathbb{N}^*$ to $\overline{\text{Im}(F)} \times \overline{\text{Im}(G)} \subset [0, 1] \times [0, 1]$ given by:

$$\mathcal{T} = \begin{cases} U = F(X) \\ V = G(Y) \end{cases}$$

where $\text{Im}(F)$ et $\text{Im}(G)$ are the sets of the values of the cumulative distribution functions F and G when X et Y vary in $\mathbb{N}^*$. Consequently for every $i, j \in \mathbb{N}^*$ one has:

$$\mathcal{T}(i, j) = \mathcal{T}(X = i, Y = j) = (U = u_i = F(i), V = v_j = G(j)) = (F(i), G(j))$$

Moreover, if $\mathbb{P}^{\mathcal{T}}$ is the probability measure image of $\mathbb{P}$ by the transformation $\mathcal{T}$, it follows that:

$$\mathbb{P}^{\mathcal{T}}(U = u_i, V = v_j) = \mathbb{P}^{\mathcal{T}}(U = F(i), V = G(j)) = \mathbb{P}(X = i, Y = j) = p_{ij}$$

Thus, the random vectors (X, Y) and (U, V) have the same probability measure (however defined on different σ-algebras) which implies that the mutual information between the components of the two vectors (X, Y) and (U, V) is invariant under the transformation $\mathcal{T}$. In other words, one has:

$$\mathcal{I}_\varphi(X, Y) = \mathcal{I}_\varphi(U, V)$$

for every admissible function φ. It follows from this result, that the previous optimization problem, as settled in $\mathbb{N}^* \times \mathbb{N}^*$, can be reduced to an equivalent optimization problem in $\overline{\mathrm{Im}(F)} \times \overline{\mathrm{Im}(G)} \subset [0, 1] \times [0, 1]$, which will allow to use possibly more efficient numericals methods since the research of the optimum partition can be done on a bounded support (included in $[0, 1]^2$, but however, non-compact) of the probability measure $\mathbb{P}^{\mathcal{T}}$.

Remark : It is worth noticing that the previous transformation, although conceptually similar to the one leading to the notion of copula, does not further pursue the analogy with the latter, since in the case where the variables $U = F(X)$ and $V = G(Y)$ take values on countable sets, the concept of uniform measure on such sets is a nonsense !

3 Examples

3.1 Some computational considerations

Without loss of generality and for sake of simplicity, we consider the case of a bivariate discrete random vector $X = (X_1, X_2)$, with a finite support:

$$S_{\mathbb{P}} = S_{\mathbb{P}_{X_1}} \times S_{\mathbb{P}_{X_2}} = [0, 1, 2, ..., r] \times [0, 1, 2, ..., s]$$

for which the probability density function is given by:

$$f(x_{1i}, x_{2j}) = \mathbb{P}(X_1 = x_{1i}, X_2 = x_{2j}) = p_{ij}$$

where $i = 0, 1, 2, ..., r$ and $j = 0, 1, 2, ..., s$.

For each component, we denote respectively by:

$$0 = \alpha_{10} < \alpha_{11} < \alpha_{12} < ... < \alpha_{1l} < ... < \alpha_{1(p-1)} < \alpha_{1p} = r$$

and by:

$$0 = \alpha_{20} < \alpha_{21} < \alpha_{22} < ... < \alpha_{2m} < ... < \alpha_{2(q-1)} < \alpha_{2q} = s$$

the bounds of the intervals of the partitions of $[0, 1, 2, ..., r]$ and of $[0, 1, 2, ..., s]$ in respectively, p and q elements. For every $l = 1, 2, ..., p$ and for every $m = 1, 2, ..., q$, the probability measure of the cartesian product:

$$\mathfrak{R}_{lm} = \left[\alpha_{1(l-1)}, \alpha_{1l}\right] \times \left[\alpha_{2(m-1)}, \alpha_{2m}\right]$$

is given by:

$$\sum_{(x_{1i}, x_{2j}) \in \mathfrak{R}_{lm}} p_{ij} = \pi_{lm}$$

while the joint product probability measure is expressed, using usual notations, as:

$$\left(\sum_{(x_{1i}, x_{2j}) \in \left[\alpha_{1(l-1)}, \alpha_{1l}\right] \times S_{\mathbb{P}_{X_2}}} p_{ij}\right) \times \left(\sum_{(x_{1i}, x_{2j}) \in S_{\mathbb{P}_{X_1}} \times \left[\alpha_{2(m-1)}, \alpha_{2m}\right]} p_{ij}\right) = \pi_{l\bullet}\pi_{\bullet m}$$

Hence it follows that the approximation of the initial mutual information $\mathcal{I}_{\varphi}(X_1, X_2)$, arising from the partition $\{\mathfrak{R}_{lm}\}_{l=1,m=1}^{p,q}$ of the support $[0, 1, 2, ..., r] \times [0, 1, 2, ..., s]$, has the following expression:

$$\sum_{l=1}^{p} \sum_{m=1}^{q} \varphi\left(\frac{\pi_{lm}}{\pi_{l\bullet}\pi_{\bullet m}}\right) \pi_{l\bullet}\pi_{\bullet m}$$

and for which, one has to find the maximum with respect to the unknown real numbers $\{\alpha_{1l}\}$ et $\{\alpha_{2m}\}$, under the previous sets of inequality constraints. The optimal partition being obtained, one can compare, on the basis of the percentage of the initial mutual information explained by the partition, the latter with others usual partitions as those with classes of equal size, or classes with "*equal probability*", adapted in this case to the discrete context. For multivariate random vectors with more than two components, the generalization of this procedure is straightforward.

3.2 Multinomial distribution

One will say that the discrete random vector $X = (X_1, X_2, ..., X_k)$ is distributed as a multivariate multinomial distribution, denoted $\mathcal{M}n(n; p_1, p_2, ..., p_k)$, with parameters $n, p_1, p_2, ..., p_k$, if its joint probability density function has the following form:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \binom{n}{x_1, x_2, ..., x_k} \prod_{i=1}^{k} p_i^{x_i} \mathbb{I}_{\left\{S_{\mathbb{P}} \subseteq [0,1,2,...,n]^k\right\}}(x_1, x_2, ..., x_k)$$

where:

$$\binom{n}{x_1, x_2, ..., x_k} = \frac{n!}{x_1! x_2! ..., x_k!}$$

where:

$$0 \leq p_i \leq 1 \text{ for every } i = 1, 2, ..., k \text{ with } \sum_{i=1}^{k} p_i = 1$$

and where:

$$S_{\mathbb{P}} \subseteq [0, 1, 2, ..., n]^k = \{(x_1, x_2, ..., x_k) : \sum_{i=1}^{k} x_i = n\}$$

For each components of the random vector X, the bounds of the intervals of a given partition of $[0, 1, 2, ..., n]$ will be as previously said of the form:

$$m + \frac{1}{2}$$

where $m = 0, 1, 2, ..., n-1$. If $n_1, n_2, ..., n_k$ denote, respectively, the number of elements of the partitions of the supports $\mathbb{P}_{X_1}, \mathbb{P}_{X_2}, ..., \mathbb{P}_{X_k}$, the number of possible choices for a partition of $S_{\mathbb{P}}$ in $n_1 \times n_2 \times ... \times n_k$ elements or cells, will be given by:

$$\prod_{i=1}^{k} \binom{n}{n_i - 1}$$

and by exhaustion, one can find an optimal partition of $S_{\mathbb{P}}$.

For example one considers the following multivariate mutinomial random vector:

$$X = (X_1, X_2, X_3) \sim \mathcal{M}n(20; 0.2, 0.5, 0.3)$$

if $n_1 = 5, n_2 = 7$ et $n_3 = 4$, then the number of partitions of $S_{\mathbb{P}}$ in 140 cells will be given by:

$$\binom{20}{4} \times \binom{20}{6} \times \binom{20}{3} = \frac{(20!)^3}{4!16!6!14!3!17!}$$

or approximately 2.1408×10^{11}...If the computer can evaluate 10^6 partitions per second, it will be necessary to run it during approximately 2.5 days !

3.3 Multivariate Poisson distribution

3.3.1 Multivariate Poisson distribution of type I

One considers two random variables distributed as two *Poisson* distributions $\mathcal{P}(\lambda)$ et $\mathcal{P}(\mu)$ with parameters λ and $\mu > 0$ and with probability density functions given, as usual, by:

$$\mathbb{P}_X(i) = \mathbb{P}_X(X = i) = e^{-\lambda}\frac{\lambda^i}{i!} \ \forall i \in \mathbb{N}^*$$

and by:

$$\mathbb{P}_Y(j) = \mathbb{P}_Y(Y = j) = e^{-\mu}\frac{\mu^j}{j!} \ \forall j \in \mathbb{N}^*$$

If F et G are the cumulative distribution functions of the random variables X and Y, one supposes that random vector $Z = (X, Y)$, named in this case by bivariate *Poisson* distribution of type I, has a cumulative distribution function given by:

$$H(x, y) = F(x)G(y)\left[1 + \gamma\left(1 - F(x)\right)\left(1 - G(y)\right)\right]$$

for every $(x, y) \in \mathbb{R}^2$ and where $-1 \leq \gamma \leq 1$. In particular for any $(i, j) \in \mathbb{N}^{*2}$, one has:

$$H(i, j) = F(i)G(j)\left[1 + \gamma\left(1 - F(i)\right)\left(1 - G(j)\right)\right]$$

One can easily check that the marginal cumulative distribution functions are F and G and that H is actually a cumulative distribution function. Indeed one has:

$$H(i, \infty) = F(i) \ \forall i \in \mathbb{N}^*; H(\infty, j) = G(j) \ \forall j \in \mathbb{N}^*$$

$$H(\infty, \infty) = 1; H(0^-, j) = H(i, 0^-) = 0 \ \forall (i, j) \in \mathbb{N}^{*2}$$

Moreover, H is a non decreasing function with respect to each of its components. Since for every $j \in \mathbb{N}^*$:

$$H(i + 1, j) = F(i + 1)G(j)\left[1 + \gamma\left(1 - F(i + 1)\right)\left(1 - G(j)\right)\right]$$

and since $F(i + 1) = F(i) + p_{(i+1)\bullet}$, one has:

$$H(i + 1, j) = [F(i) + p_{(i+1)\bullet}]G(j)\left[1 + \gamma\left(1 - (F(i) + p_{(i+1)\bullet})\right)\left(1 - G(j)\right)\right]$$

By expanding this last expression one obtains:

$$\begin{aligned} H(i + 1, j) = & \underbrace{F(i)G(j)\left[1 + \gamma\left(1 - F(i)\right)\left(1 - G(j)\right)\right]}_{H(i,j)} + F(i)G(j)\left[1 - \gamma p_{(i+1)\bullet}\left(1 - G(j)\right)\right] \\ & + p_{(i+1)\bullet}G(j)\left[1 + \gamma\left(1 - F(i + 1)\right)\left(1 - G(j)\right)\right] \end{aligned}$$

which allows to conclude that:

$$\begin{aligned} H(i+1,j)-H(i,j) &= F(i)G(j)\left[1-\gamma p_{(i+1)\bullet}\left(1-G(j)\right)\right] \\ &\quad +p_{(i+1)\bullet}G(j)\left[1+\gamma\left(1-F(i+1)\right)\left(1-G(j)\right)\right] \end{aligned}$$

and therefore one has: $H(i+1,j)-H(i,j)\geq 0$ for every i and j in $\mathbb{N}^*$. Thus $H(i,j)$ is a non decreasing function with respect to i for any given j. Of course, the same conclusion is true with respect to j for any given i. Furthermore, one can deduce easily from $H(i,j)$ the joint probabilities p_{ij} by the means of the usual equality:

$$\mathbb{P}(X=i,Y=j)=p_{ij}=H(i,j)-H(i-1,j)-H(i,j-1)+H(i-1,j-1)$$

for any $(i,j)\in\mathbb{N}^{*2}$. Finally, the mutual information between the components X and Y is given by:

$$\begin{aligned} \mathcal{I}_{\varphi}(X,Y) &= I_{\varphi}(\mathbb{P},\mathbb{P}_X\otimes\mathbb{P}_Y) \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\varphi\left[\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right]p_{i\bullet}p_{\bullet j} \end{aligned}$$

Using the transformation:

$$U=F(X)\text{ and }V=G(Y)$$

it follows that the support $S_{\mathbb{P}}$ of the discrete random vector $(U,V)=\{F(i),G(j)\}_{i,j=0,1,2,...}$, is a countable and bounded subset of $[0,1]^2$.

One considers a partition of $S_{\mathbb{P}}$ in $n\times m$ cells given by the cartesian product between the partitions of $S_{\mathbb{P}_U}$ and $S_{\mathbb{P}_V}$ in respectively n and m intervals. As in the case of finite supports, the bounds of the intervals may be choosen as the center the intervals $[F(i),F(i+1)[$ and $[G(j),G(j+1)[$. Let:

$$0<u_1<u_2<...<u_k<...<u_{n-1}<\infty$$

and:

$$0<v_1<v_2<...<v_l<...<v_{m-1}<\infty$$

be the real bounds of the intervals defining the partitions $\mathcal{P}_U$ and $\mathcal{P}_V$ of $S_{\mathbb{P}_U}$ and $S_{\mathbb{P}_V}$. For every $k=1,2,...,n$ and $l=1,2,...,m$, let γ_{u_k} and γ_{v_l} be the intervals of the form:

$$\gamma_{u_k} = [u_{k-1},u_k[\text{ where }u_0=0\text{ and }\gamma_{u_n}=[u_{n-1},1]$$

$$\gamma_{v_l} = [v_{l-1},v_l[\text{ where }v_0=0\text{ and }\gamma_{u_m}=[v_{m-1},1]$$

It follows that the mutual information explained by the partition $\mathcal{P}=\mathcal{P}_U\otimes\mathcal{P}_V$ of $S_{\mathbb{P}}$, is given by:

$$\mathcal{I}_{\varphi}(\mathcal{P})=\sum_{k=1}^{n}\sum_{l=1}^{m}\varphi\left(\frac{\mathbb{P}\left(\gamma_{u_k}\times\gamma_{v_l}\right)}{\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})}\right)\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})$$

where:

$$\mathbb{P}\left(\gamma_{u_k}\times\gamma_{v_l}\right)=\sum_{\{(i,j)\}\in\mathbb{N}^{*2}:(i,j)\in\gamma_{u_k}\times\gamma_{v_l}}p_{ij}$$

where:

$$\mathbb{P}_U\left(\gamma_{u_k}\right) = \sum_{\{(i,j)\in\mathbb{N}^{*2}:(i,j)\in\gamma_{u_k}\times\mathbb{N}^*\}} p_{ij}$$

and where:

$$\mathbb{P}_V(\gamma_{v_l}) = \sum_{\{(i,j)\in\mathbb{N}^{*2}:(i,j)\in\mathbb{N}^*\times\gamma_{v_l}\}} p_{ij}$$

To find an optimal partition $\mathcal{P}^*$ of $S_{\mathbb{P}}$ one has to choose, among a countable number of possible choices, $(n-1)$ and $(m-1)$ real numbers, shuch that the mutual information $\mathcal{I}_\varphi(\mathcal{P}^*)$ explained by the partition is maximum. This gives rise to the following formulation of the optimization problem:

$$\mathcal{I}_\varphi(\mathcal{P}^*) = \max_{\mathcal{P}\in\mathcal{P}_{\mathbf{n}}} \mathcal{I}_\varphi(\mathcal{P}) = \max_{\mathcal{P}\in\mathcal{P}_{\mathbf{n}}} \sum_{k=1}^{n}\sum_{l=1}^{m} \varphi\left(\frac{\mathbb{P}\left(\gamma_{u_k}\times\gamma_{v_l}\right)}{\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})}\right)\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})$$

or equivalently:

$$\mathcal{P}^* = \arg\left[\max_{\mathcal{P}\in\mathcal{P}_{\mathbf{n}}} \sum_{k=1}^{n}\sum_{l=1}^{m} \varphi\left(\frac{\mathbb{P}\left(\gamma_{u_k}\times\gamma_{v_l}\right)}{\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})}\right)\mathbb{P}_U\left(\gamma_{u_k}\right)\mathbb{P}_V(\gamma_{v_l})\right]$$

where $\mathbf{n} = (n, m)$.

In the case where some numericals methods fail to find easily an optimal solution for an unbounded support, it is possible to bring back this problem to the bounded case, using the following method. Let u_r and v_s be two real numbers shuch that:

$$1 - H(u_r, v_s) = (1 - F(u_r)) + (1 - G(v_s)) - \sum_{i\geq u_r}\sum_{j\geq v_s} p_{ij} \leq \epsilon$$

where ϵ is as small as possible and where:

$$H(u_r, v_s) = F(u_r)G(v_s)\left[1 + \gamma\left(1 - F(u_r)\right)\left(1 - G(v_s)\right)\right]$$

As an example, if $X \sim \mathcal{P}(2)$ and if $Y \sim \mathcal{P}(4)$ with $\gamma = .5$ and $\epsilon = 10^{-4}$, one has to find two real numbers u_r and v_s shuch that:

$$(1 - F(u_r)) + (1 - G(v_s)) - \sum_{i\geq u_r}\sum_{j\geq v_s} p_{ij} \leq 10^{-4}$$

where:

$$p_{ij} = H(i, j) - H(i-1, j) - H(i, j-1) + H(i-1, j-1)$$

with:

$$H(i, j) = F(i)G(j)\left[1 + .5(1 - F(i))(1 - G(j))\right]$$

Furthermore:

$$(1 - F(u_r)) + (1 - G(v_s)) \leq \epsilon \Longrightarrow 1 - H(u_r, v_s) \leq \epsilon$$

Using software (or tables) for *Poisson* distributions with parameters $\lambda = 2$ and $\mu = 4$, one easily finds:

$$1 - F(9) \approx 3 \times 10^{-5}$$

for$\lambda = 2$, and:

$$1 - G(14) \simeq 5 \times 10^{-5}$$

for$\mu = 4$.

It then follows that one can consider "new" discrete random variables U and V taking a finite number of values given by $0, F(1), F(2), ..., F(10)$ and by $0, G(1), G(2), ..., G(15)$ with "new" probabilities deduced from $\mathcal{P}(\lambda)$ and $\mathcal{P}(\mu)$ by a process of troncation. An another way to adjust the probability density functions is to define in the present case:

$$\mathbb{P}(F(10)) = 1 - F(9) \quad \text{and} \quad \mathbb{P}(G(15)) = 1 - G(14)$$

It is worth noticing that in general one can arbitrarly choose larger "terminal bounds", insofar as one retains the property of finite supports. An another way to reduce the problem to the bounded support case, is to choose two integers u and v in order that the relative residual mutual information, given by:

$$1 - \frac{\sum_{i=0}^{u}\sum_{j=0}^{v}\varphi\left[\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right]p_{i\bullet}p_{\bullet j}}{\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\varphi\left[\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right]p_{i\bullet}p_{\bullet j}}$$

is less than or equal to a given number ϵ, as small as possible. Nothing ensures however that the pair (u, v) is unique, the most important being that there at least one. Furthermore, it should be noted that more this quantity is close to 1, more the restriction of the mutual information to the domain:

$$\tilde{S}_{\mathbb{P}} = S_{\mathbb{P}} \backslash \{0 \leq i \leq u\} \times \{0 \leq j \leq v\}$$

is close to 0. Therefore, one can easily conclude that the random variables are almost independent on the set $\tilde{S}_{\mathbb{P}}$ and that this one has not a significant contribution in the expression of the stochastic dependence between the variables U and V.

Finally, having obtained an optimal partition for (U, V), it is straightforward to obtain, using the inverse transform, the correspondent optimal partition for (X, Y).

3.3.2 Multivariate Poisson distribution of type II

Following Johnson, Kotz et Balakrishnan [18] , one says that the discrete randon vector is distributed as a multivariate *Poisson* distribution $\mathcal{P}(\lambda_1, \lambda_2, ..., \lambda_k)$ with parameters $\lambda_1, \lambda_2, ..., \lambda_k$ and named thereafter multivatiate *Poisson* distribution of type II, if its probability density function is given by:

$$\begin{aligned}\mathbb{P}\left(\cap_{i=1}^{k}\{X_i = x_i\}\right) &= \left[\prod_{i=1}^{k}\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\right] \times \\ &\exp\left\{\begin{array}{l}\sum_i\sum_j \lambda_{ij}C(x_i)C(x_j) \\ +\sum_i\sum_j\sum_l \lambda_{ijk}C(x_i)C(x_j)C(x_l) \\ +... + \lambda_{12...k}C(x_1)C(x_2)...C(x_k)\end{array}\right\}\mathbb{I}_{\mathbb{N}^{*k}}(x_1, x_2, ..., x_k)\end{aligned}$$

where $\lambda_i > 0$ for every $i = 1, 2, ..., k$, where $C(\bullet)$ is a *Gram-Charlier* polynomial of type B (see [18]) and where:

$$\lambda_{ijl...} = \mathbb{E}\left[X_i X_j X_l ...\right]$$

for every $i, j, l, ...$belonging to $1, 2, ..., k$. For sake of simplicity, one only considers thereafter a bivariate Poisson distribution. Among the numerous methods for constructing families of bivariate *Poisson* distributions, the following was introduced by Holgate [16],[17] (see also Campbell [9], Aitken and Gonin [3], Aitken [2], Consael [11]).

Let Y_1, Y_2, Y_3 be three independent random variables, distributed respectively as *Poisson* distributions with parameters λ_1, λ_2 and λ_3. One considers two new variables defined by:

$$X_1 = Y_1 + Y_3 \text{ and } X_2 = Y_2 + Y_3$$

Then one can show (see [17]) that the probability density function of the random vector $X = (X_1, X_2)$ has the following form:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{i=0}^{\min(x_1,x_2)} \frac{\lambda_1^{x_1-i}\lambda_2^{x_2-i}\lambda_3^i}{(x_1-i)!(x_2-i)!i!} \mathbb{I}_{\mathbb{N}^{*2}}(x_1, x_2)$$

Obviously, the marginal distributions of the components X_1 and X_2 are univariate *Poisson* distributions with parameters $\lambda_1 + \lambda_3$ and $\lambda_2 + \lambda_3$. If necessary, one can also reduce this case to the one with a bounded support included in $[0, 1]^2$.

3.4 Multivariate hypergeometric distribution

This distribution appears naturally in the following context (also knowned under the name of "sampling in an urn having k categories"): one considers a population P divided in k classes or categories $C_1, C_2, ..., C_k$ and having m units whose m_i belong to C_i, for every $i = 1, 2, ..., k$ and where:

$$\sum_{i=1}^k m_i = m$$

A random sample S of n units is drawn from P without replacement and for every $i = 1, 2, ..., k$, let X_i be the random variable defined as the number of units of C_i in the sample S of size n. It is well known that the random vector $X = (X_1, X_2, ..., X_k)$ is distributed as a multivariate hypergeometric distribution, denoted by $\mathcal{HM}(n; m_1, m_2, ..., m_k)$ with parameters $n, m_1, m_2, ..., m_k$ and whose probability density function is given by:

$$P(X_1 = n_1, X_2 = n_2, ..., X_k = n_k) = \frac{\prod_{i=1}^k \binom{m_i}{n_i}}{\binom{m}{n}}$$

where one has :

$$\sum_{i=1}^k n_i = n; 0 \leq n_i \leq \min(n, m_i) \ \forall i = 1, 2, ..., k$$

It is obvious that for $k = 2$ this distribution is nothing else than the usual hypergeometric distribution $\mathcal{H}(n, m)$ and it is straightforward to check moreover, that the marginals distributions are also hypergeometric $\mathcal{H}(n, m_i)$ for every $i = 1, 2, ..., k$, with probability density functions given by:

$$\mathbb{P}(X_i = n_i) = \frac{\binom{m_i}{n_i}\binom{m-m_i}{n-n_i}}{\binom{m}{n}}$$

In practice, one encounters this kind of distribution in surveys, censuses, and data mining, where for various reasons, one has to create classes or categories. In order to retain as much as possible the stochastic dependence between the variables, it is important to use an optimal quantization of the support of the joint probability distribution of the whole set of variables.

3.5 Negative multinomial distribution

One will say that the discrete random vector $X = (X_1, X_2, ..., X_k)$ is distributed as a multivariate negative multinomial distribution, denoted by $\mathcal{MN}(p_0, p_1, p_2, ..., p_k)$, with parameters $p_0, p_1, p_2, ..., p_k$ if its probability density function has the following form:

$$P(X_1 = n_1, X_2 = n_2, ..., X_k = n_k) = \frac{\Gamma(n + \sum_{i=1}^{k} n_i)}{n_1! n_2! ... n_k! \Gamma(n)} p_0^n \prod_{i=1}^{k} p_i^{n_i} \mathbb{I}_{\mathbb{N}^{*k}}(n_1, n_2, ..., n_k)$$

where $0 < p_i < 1$, where $i = 1, 2, ..., k$ and where $\sum_{i=0}^{k} p_i = 1$. This distribution has many applications as, for example, in insurance, medical sciences, physics, epidemiology, etc. (for more informations one can see: Bates and Neyman [5], Sibuya, Yoshimura and Shimizu [24], Patil [20], Neyman [21], Sinoquet and Bonhomme [25], Guo [15], Arbous and Sichel [4]).

4 Mixed random vector

In many applications, it is usual that one has to deal with mixed random vectors, or in other words, with random vectors having some components continuous while the others are discrete. One encounters this kind of situation in surveys and censuses, where continuous and discrete variables are present simultaneously. It is the same thing in "data mining" where a very large number of random variables, both continuous as discrete, are observed on a huge number of observations.

For practical reasons, one will note respectively by $\tilde{X} = (X_1, X_2, ..., X_k)$ and by $\tilde{Y} = (Y_1, Y_2, ..., Y_l)$, the sets of the continuous and discrete components of X. Assuming that all the conditions are fulfilled in order to legitimate the following relations, the joint distribution function $f_X(\tilde{x}, \tilde{y})$ of the whole set of the components, may be written as:

$$f_X(\tilde{x}, \tilde{y}) = f_{\tilde{X}|\tilde{y}}(\tilde{x}|\tilde{y}) m_{\tilde{Y}}(\tilde{y}) = f_{\tilde{X}}(x_1, x_2, ..., x_k | y_1, y_2, ..., y_l) m_{\tilde{Y}}(y_1, y_2, ..., y_l)$$

where $f_{\tilde{X}}(\tilde{x}|\tilde{y})$ is the conditional probability density function of the random vector $\tilde{X}$ given $\tilde{Y} = \tilde{y}$ and where $m_{\tilde{Y}}(\tilde{y})$ is the marginal probability density function of the discrete random vector $\tilde{Y}$. Noting by $m_{\tilde{X}}(\tilde{x})$ the marginal distribution of $\tilde{X}$, by $m_{X_i}(x_i)$ $i = 1, 2, ..., k$ and by $m_{Y_j}(y_j)$ $j = 1, 2, ..., l$, the marginal distributions of the components of $X = (\tilde{X}, \tilde{Y})$, the mutual information $\mathcal{I}_\varphi(X_1, X_2, ..., X_k, Y_1, Y_2, ..., Y_l)$ between the components of X can be expressed as:

$$\sum_{\tilde{y}} \left\{ \int_{\tilde{x}} \varphi \left(\frac{f_X(x_1, x_2, ..., x_k, y_1, y_2, ..., y_l)}{[\otimes_{i=1}^{k} m_{X_i}(x_i)] \times [\otimes_{j=1}^{l} m_{Y_j}(y_j)]} \right) (\otimes_{i=1}^{k} m_{X_i}(x_i)) d(\tilde{x}) \right\} \times \otimes_{j=1}^{l} m_{Y_j}(y_j)$$

where $d(\tilde{x})$ is equal, as usual, to $dx_1 dx_2 ... dx_k$. It frequently happens in practice that, for various reasons, one has to create classes for continuous or discrete variables, leading through this process of data reduction, to a loss of the initial mutual information. In order that this loss be the smallest as possible, one will use an optimal quantization of the support $S_{\mathbb{P}}$ according to the same principles as those presented in [10] for the continuous case and, in this paper, for the discrete case. Form this point of view, a remark is in order to the extent that, one has to understand the word “discrete”, as a term covering all types of qualitative

or categorical variables, that the latter are really discrete, ordinal or nominal. As a matter of fact, some models of data analysis, as correspondence analysis, has to deal with all these types of variables.

However, in many applications, it is common that the support $S_{\mathbb{P}}$ is equal to $\mathbb{R}^k \times \mathbb{N}^{*l}$ or equal to $\mathbb{R}_+^k \times \mathbb{N}^{*l}$. Having choosen arbitrary numbers $n_1, n_2, ..., n_k, m_1, m_2, ..., m_l$ of classes for each components, the resulting partition $\mathcal{P}$ of $S_{\mathbb{P}}$ will consist of:

$$r = \left(\prod_{i=1}^{k} n_i\right) \times \left(\prod_{j=1}^{l} m_j\right)$$

cells, each of them being given by the cartesian product:

$$\left(\times_{i=1}^{i=k} \gamma_{ij_i}\right) \times \left(\times_{j=1}^{j=l} \gamma_{jm_j}\right)$$

where $\{\gamma_{ij_i}\}$ and $\{\gamma_{jm_j}\}$ are, respectively, the set of the intervals of the partition $\mathcal{P}_i$ related to the continuous component X_i for $i = 1, 2, ..., k$, and the set of the intervals of the partition $\mathcal{Q}_j$ related to the discrete component Y_j for $j = 1, 2, ..., l$. If $\mathcal{P}_{(\mathbf{n},\mathbf{m})}$, where $(\mathbf{n}, \mathbf{m})$ is the multi-index $(\mathbf{n} = (n_1, n_2, ..., n_k), \mathbf{m} = (m_1, m_2, ..., m_l))$, is the set of partitions of $S_{\mathbb{P}}$ in r non empty cells, then for each member $\mathcal{P}_\eta$ of $\mathcal{P}_{(\mathbf{n},\mathbf{m})}$ one will evaluate the mutual information $\mathcal{I}_\varphi(\mathcal{P}_\eta)$ explained by $\mathcal{P}_\eta$, for $\eta = 1, 2, ..., r$, and one will choose the bounds of the intervals in order that this last be maximum. The following elementary example illustrates the above procedure.

Let $\tilde{X} = (X_1, Y_1)$ be a bivariate mixed random vector where the random variables X_1 and Y_1 are, respectively, continuous and discrete. For a more easier presentation, one supposes that the support $S_{\mathbb{P}}$ is given by $[0, a] \times \{0, 1, 2, ..., n\}$ and that the probability distribution function of X is obtained from the marginal distribution $m_{Y_1}(y_1)$ of Y_1 and from the set of conditional distributions $f_{X_1|i}(x_1|i)$ for $i = 0, 1, 2, ..., n$. It is worth noticing that the conditioning is done with respect to the values of the discrete variable, instead of the values of the continuous one. This choice is natural and is a consequence of the fact that the stochastic dependance between the continuous variables, is usually describe by the means of the categories of the discrete variables and, moreover, are more easy to understand and vizualise. As an example, if the discrete variable Y_1 is the sex of respondents to a given survey and if the continuous variables X_1 and X_2 are the height and the weight of these respondents, it is not difficult to convince themselves that the mutual information between X_1 and X_2 depends on the sex, which is of course the case ! In fact, this example shows that in the case of categorical variables, the concept of conditional optimal partitions makes perfect sense.

Finally, let $n_1 = n_2 = 3$, be the number of intervals of the partition of the sets $[0, a]$ and $\{0, 1, 2, ..., n\}$ and whose bounds are real numbers α_1, α_2 and β_1, β_2 such that:

$$0 < \alpha_1 < \alpha_2 < a \quad \text{and} \quad 0 < \beta_1 < \beta_2 < n$$

If γ_l and δ_m, for $l, m = 1, 2, 3$, are the corresponding intervals of such partitions, the product partition of the support $[0, a] \times \{0, 1, 2, ..., n\}$ will be given by the set of the rectangles or cells $\{\mathfrak{R}_{lm} =_{\gamma_l \times \delta_m}\}_{l,m=1,2,3}$. The probability p_{lm} of $\mathfrak{R}_{lm}$ is then express by:

$$p_{lm} = \int\int_{\mathfrak{R}_{lm}} f_{X_1}(x_1, y_1) dx_1 dy_1 = \sum_{i \in \delta_m} \left[\int_{x_1 \in \gamma_l} f_{X_1|i}(x_1|i) dx_1\right] p_i$$

where p_i is the probability $\mathbb{P}(Y_1 = i)$ for $i = 0, 1, 2, ..., n$. As to the marginal probability $p_{l\bullet}$ of γ_l, one has:

$$p_{l\bullet} = \sum_{i=0}^{n} \left[\int_{x_1 \in \gamma_l} f_{X_1|i}(x_1|i) dx_1 \right] p_i$$

while the marginal probability $p_{\bullet m}$ of δ_m will be given by:

$$p_{\bullet m} = \sum_{i \in \delta_m} p_i$$

It follows that the mutual information explained by $\mathcal{P} = \{r_{lm}\}_{l,m=1,2,3}$ has for expression:

$$\mathcal{I}_\varphi(\mathcal{P}) = \sum_{l=1}^{3} \sum_{m=1}^{3} \varphi \left(\frac{p_{lm}}{p_{l\bullet} p_{\bullet m}} \right) p_{l\bullet} p_{\bullet m}$$

This last quantity depend on the real values $\alpha_1, \alpha_2, \beta_1$ and β_2 and one will determine these one so that the optimum partition $\mathcal{P}^*$ of $[0, a] \times \{0, 1, 2, ..., n\}$ in 9 cells satisfies the following optimization problem:

$$\mathcal{I}_\varphi(\mathcal{P}^*) \underset{\alpha_1,\alpha_2,\beta_1,\beta_2}{=\max} \mathcal{I}_\varphi(\mathcal{P}) \underset{\alpha_1,\alpha_2,\beta_1,\beta_2}{=\max} \left[\sum_{l=1}^{3} \sum_{m=1}^{3} \varphi \left(\frac{p_{lm}}{p_{l\bullet} p_{\bullet m}} \right) p_{l\bullet} p_{\bullet m} \right]$$

The unknown quantities (α_1, α_2) and (β_1, β_2) appear in the above equality as bounds of integrals or bounds of finite sums. Thus, for each choice of β_1 and β_2, one has to solve the following conditional optimization problem:

$$\mathcal{I}_\varphi(\mathcal{P}^*)_{\beta_1,\beta_2} = \max_{\alpha_1,\alpha_2} \sum_{l=1}^{3} \varphi \left(\frac{p_{lm}}{p_{l\bullet} p_{\bullet m}} \right) p_{l\bullet} p_{\bullet m}$$

Subsequently using derivation technics, one will select the pair (β_1^*, β_2^*), and consequently the optimal partition $\mathcal{P}^*$, such that:

$$\mathcal{I}_\varphi(\mathcal{P}^*)_{\beta_1^*,\beta_2^*} = \sup_{\beta_1,\beta_2} \mathcal{I}_\varphi(\mathcal{P}^*)_{\beta_1,\beta_2} = \mathcal{I}_\varphi(\mathcal{P}^*)$$

5 A more general setting

In some practical situations, one has often to deal in a certain number of circumstances, as for example in data mining, with a huge amount of data (quite often 10^4 to 10^6 observations). In this case, one can certainly rely on this mass information, in order to avoid any parametric or semi-parametric approaches, and use instead some convergence properties to design an optimal quantization of the support of an empirical probability measure. More precisely, one supposes that in a data wharehouse, one has at disposal k finite random variates $X_1, X_2, ..., X_k$ for which one postulates the existence of an unique, but unknown, joint probability measure given by:

$$p_{i_1 i_2 ... i_k} = \mathbb{P}(X_1 = i_1, X_2 = i_2, ..., X_k = i_k) > 0$$

where, for the sake of an easy presentation, one can suppose that $(i_1, i_2, ..., i_k) \in S$ where S is a finite subset of $\mathbb{N}^{*^k}$ and where $i_l = 1, 2, ..., \eta_l$ for $l = 1, 2, ..., k$. A random sample of size n of this distribution gives raise to a k-way contingency table with $\eta_1 \eta_2 ... \eta_k$ cells,

which in turn may be thought, by the means of a "*vec*" operator (see Fang and Zhang [14] , Bilodeau and Brenner [7]), as a multinomial distribution having $\eta_1\eta_2...\eta_k$ categories C_j, with parameters n and $\{(p_{i_1i_2...i_k})_{i_1i_2...i_k\in S}\}$ and whose its support is included in the simplex $\sum_{j=1}^{\eta_1\eta_2\cdots\eta_k} y_j = n$ of $\mathbb{R}^{\times_{l=1}^{k}\eta_l}$ with y_j being the number of the occurences of the jth category C_j. Let $\mathbf{p} =(p_{i_1i_2...i_k} : i_1i_2...i_k \in S)$ be the vector of probabilities, which may be rewritten under the form $\mathbf{p} =(p_j : j = 1, 2, ..., \eta_1\eta_2...\eta_k)$, and let $\hat{\mathbf{p}} = \left(\hat{p}_y = \frac{y_j}{n} : j = 1, 2, ..., \eta_1\eta_2...\eta_k\right)$ be the vector of the relative frequencies of the categories C_j. It is well known (see for example Serfling [23], Billingsley [6]) that:

$$\sqrt{n}\,(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

where:

$$\Sigma = diag(\mathbf{p}) - \mathbf{p}^t\mathbf{p}$$

or says otherwise: $\hat{\mathbf{p}}$ is $A\mathcal{N}(\mathbf{p}, \frac{1}{n}\Sigma)$. Moreover one has:

$$\hat{\mathbf{p}} \xrightarrow{p} \mathbf{p}$$

It is worth noticing that using the k-way contingency table representation of the data, same results hold for the marginal probabilities. Let $\widehat{\mathcal{I}}_\varphi\,(X_1, X_2, ..., X_k)$ be an estimator of $\mathcal{I}_\varphi\,(X_1, X_2, ..., X_k)$ given by:

$$\widehat{\mathcal{I}}_\varphi\,(X_1, X_2, ..., X_k) = \sum\nolimits_{(i_1i_2...i_k)\in S} \varphi\left[\frac{\hat{p}_{i_1i_2...i_k}}{\hat{p}_{1i_1}\hat{p}_{2i_2}...\hat{p}_{ki_k}}\right]\hat{p}_{1i_1}\hat{p}_{2i_2}...\hat{p}_{ki_k} = \psi(\hat{\mathbf{p}})$$

where $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}_j$ for $j = 1, 2, ..., k$ are respectively consistent estimators of the joint and marginal probability distribution functions. Since φ is a continuous function from $\mathbb{R}_+\backslash\{0\}$ to $\mathbb{R}$, it is straightforward to check, using *Slutsky* theorem, that:

$$\varphi\left[\frac{\hat{p}_{i_1i_2...i_k}}{\hat{p}_{1i_1}\hat{p}_{2i_2}...\hat{p}_{ki_k}}\right] \xrightarrow{p} \varphi\left[\frac{p_{i_1i_2...i_k}}{p_{1i_1}p_{2i_2}...p_{ki_k}}\right]$$

for all $i_1, i_2, ..., i_k$ and that:

$$\widehat{\mathcal{I}}_\varphi\,(X_1, X_2, ..., X_k) \xrightarrow{p} \mathcal{I}_\varphi\,(X_1, X_2, ..., X_k)$$

Moreover, since:

$$\hat{\mathbf{p}} \text{ is } A\mathcal{N}\left(\mathbf{p}, \frac{1}{n}\Sigma\right)$$

one has, under very mild conditions (see [23]):

$$\sqrt{n}\left[\widehat{\mathcal{I}}_\varphi\,(X_1, X_2, ..., X_k) - \mathcal{I}_\varphi\,(X_1, X_2, ..., X_k)\right] = \sqrt{n}\,(\psi(\hat{\mathbf{p}}) - \psi(\mathbf{p})) \xrightarrow{\mathcal{L}} N(0, {}^t\nabla\psi\Sigma\nabla\psi)$$

where:

$$\nabla\psi = \operatorname{grad}\psi|_{\hat{\mathbf{p}}=\mathbf{p}} = \left(\frac{\partial\psi}{\partial\hat{\mathbf{p}}}\right)\Bigg|_{\hat{\mathbf{p}}=\mathbf{p}}$$

In other words:

$$\psi(\hat{\mathbf{p}}) \text{ is } A\mathcal{N}\left(\psi(\mathbf{p}), \frac{1}{n}{}^t\nabla\psi\Sigma\nabla\psi\right)$$

Finally, as shown by Bosq and Lecoutre [19], one has the following inequality:

$$\forall \epsilon > 0, \mathbb{P}\left[\sup_x |\hat{F}(x) - F(x)| > \epsilon\right] \leq Ce^{-2n\epsilon^2}$$

where $\hat{F}(x)$ and $F(x)$ are respectively the empirical an theoretical cumulative distribution functions arising from $\hat{\mathbf{p}}$ and $\mathbf{p}$, where C is a constant and where n is the number of observations. From this last inequality one can deduce that (see Bosq and Lecoutre [8] or Lecoutre and Tassi [19]):

$$\sup_x |\hat{F}(x) - F(x)| \xrightarrow{c.co} 0$$

in the *completely convergence* mode, as n tends to infinity, and that:

$$\sup_x |\hat{F}(x) - F(x)| \overset{a.s}{=} O\left(\frac{\log n}{n}\right)^{1/2}$$

All these asymptotic results, shown that when the number n of observations is very large, as it is often the case in data mining ($n \simeq 10^5, 10^6$ or may be 10^7), it is not necessary, from a practical point of view, to use a semi-parametric framework, nor to estimate the probability vector $\mathbf{p}$. So, in order to find an optimal quantization of the support of the probability measure $\mathbf{p}$, one has just to proceed with $\hat{\mathbf{p}}$, as it was the known probability distribution.

6 Conclusions

Let $X = (X_1, X_2, ..., X_k)$ be a random vector, with values in a countable set, defined on a probability space $(\Omega, \mathcal{F}, \mu)$, with a joint probability measure $\mathbb{P}$ absolutely continuous with respect to a counting measure. For a given measure of mutual information $\mathcal{I}_\varphi(X_1, X_2, ..., X_k)$ between the components of X, we have shown, using a criterion based on minimization of the mutual information loss, that there exists for given integers $n_1, n_2, ..., n_k$, an optimal partition of the support $\mathcal{S}_\mathbb{P}$ of $\mathbb{P}$ in $n = \prod_{i=1}^k n_i$ elements, given by the cartesian product of the elements of the partitions of the support of each components $X_1, X_2, ..., X_k$ in, respectively, $n_1, n_2, ..., n_k$ subsets. This procedure allows to retain the stochastic dependence between the random variables $(X_1, X_2, ..., X_k)$ as much as possible and this may be significantly important for some data analysis as well as for statistical inference, as tests of independence. As illustrated by some examples, this optimal partition performs, from this point of view, better than any others having the same number of classes. Although this way of carrying out a quantization of the support of a probability measure is less usual than those associated with marginal classes of equal width or of equal probabilities, we think that practitioners could seriously consider it, at least, in the case where the conservation of the stochastic dependence between the random variables seems to be important. Finally, with a pratical point of view in mind, one has paid attention to the semiparametric case for which one can assume that the probability measure $\mathbb{P}$ is a member of a given family depending on a parameter θ.

As future researches, we will explore the following subjects. In data mining, the choice of a parametric statistical model is not quite realistic due to a huge number of variables and data and, in this case, a nonparametric framework is often more appropriate. To estimate

the probability density function of a random vector, we will use a kernel density estimator in order to evaluate the mutual information between its components and we will study the effects of the choice of the kernel on the robustness of the optimal partition. We will also look at the empirical mutual information as an estimator of the true one and we will deduce its main properties. In MCA and in classification, one has often to deal simultaneously with continuous and categorical variables, and it may be of interest to use an optimal partition in order to retain, as much as possible, the stochastic dependence between the random variables and we will explore the consequences of the choices of φ and of an optimal partition $\mathcal{P}^*$ on these models. In the multi-way stratified sampling framework, we will also pay attention to the problem of the conditional optimal partitioning depending on the values of the categorical variable.

References

[1] J. Aczél and Z. Daróczy: *On Measures of Information and Their Charaterizations*, New York: Academic Press, (1975).

[2] A.C. Aitken : *A further note on multivariate selection*, Proceedings of the Edinburgh Mathematical Society, **5**, 37-40, (1936).

[3] A.C. Aitken and H.T. Gonin : *On fourfold sampling with and without replacement,* Proceedings of the Royal Society of Edinburgh, **55**, I 14-1 25, (1935).

[4] A.G. Arbous and H.S. Sichel : *New techniques for the analysis of absenteeism data*, Biometrika, **41**, 77-90, (1954).

[5] G.E. Bates and J. Neyman : *Contributions to the theory of accident proneness*, University of California, Publications in Statistics, **1**, 215-253, (1952).

[6] P. Billingsley : *Probability and Measure*, 3rd ed. John Wiley & Sons, Inc. (1995).

[7] M. Bilodeau and D. Brenner: *Theory of Multivariate Statistics*, Springer-Verlag New York, Inc. (1999).

[8] D. Bosq et J-P. Lecoutre : *Théorie de l'Estimation Fonctionnelle*, Economica, (1987).

[9] J.T. Campbell : *The Poisson correlation function*, Proceedings of the Edinburgh Mathematical Society (Series 2), 4, 18-26, (1938).

[10] B. Colin, F. Dubeau, H. Khreibani et J. de Tibeiro : *Optimal Quantization of the Support of a Continuous Multivariate Distribution based on Mutual Information*, Journal of Classification, 30: 453-473, (2013).

[11] R. Consael : *Sur les processus composés de Poisson a deux variables aléatoires,* Académie Royale de Belgique, Classe des Sciences, Mémoires, **7**, 4-43, (1952).

[12] I. Csiszár: *A class of measures of informativity of observations channels*, Periodica Mathematica Hungarica, **Vol 2 (1-4)**, 191-213, (1972).

[13] I. Csiszár: *Information Measures: A Critical Survey*, Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions and Random Proccesses, Vol A, Prague: Publishing House of the Czechoslovak Academy of Sciences, 73-86, (1997).

[14] K.T. Fang and Y.T. Zhang: *Generalized Multivariate Analysis*, Springer-Verlag, (1990).

[15] G. Guo : *Negative multinomial regression models for clustered event counts*, Technical Report, Department of Sociology, University of North Carolina, Chapel Hill, NC, (1995).

[16] P. Holgate : *Estimation for the bivariate Poisson distribution*, Biometrika, 51, 241-245, (1964).

[17] P. Holgate : *Bivariate generalizations of Neyman's type A distribution*, Biometrika, 53, 241-245, (1966).

[18] N.L. Johnson, S. Kotz and N. Balakrishnan : *Discrete Multivariate Distribution*, John Wiley & Sons, Inc., (1997).

[19] J-P. Lecoutre et P. Tassi : *Statistique non paramétrique et robustesse*, Economica (1987).

[20] G.P. Patil : *On sampling with replacement from populations with multiple characters*, Sankhya, Series B, **30**, 355-364 (1968).

[21] J. Neyman : *Certain chance mechanisms involving discrete distributions* (Inaugural Address), Proceedings of the International Symposium on Discrete Distributions. Montréal pp. 4-14, (1963).

[22] A. Rényi: *On Measures of Entropy and Information*, Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, (I), Berkeley: University of California Press, 547-561, (1961)

[23] R.J. Serfling : *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Inc. (1980).

[24] M. Sibuya, I. Yoshimura, and R. Shimizu : *Negative multinomial distribution*, Annals of the Institute of Statistical Mathematics, **16**, 409-426, (1964).

[25] H. Sinoquet and R. Bonhomme : *A theoretical analysis of radiation interception in a two-species plant canopy*, Mathematical Biosciences, 105, 23-45, (1991).

[26] J. Zakai and M. Ziv: *On functionals satisfying a data-processing theorem*, IEEE Transactions, **IT-19**, 275-282, (1973).