Journal of Classification 30:453-473 (2013)
DOI: 10.1007/s00357-013-9127-6

Optimal Quantization of the Support of a Continuous Multivariate Distribution based on Mutual Information

Bernard Colin

Université de Sherbrooke, Québec, Canada

François Dubeau

Université de Sherbrooke, Québec, Canada

Hussein Khreibani

Université de Sherbrooke, Québec, Canada

Jules de Tibeiro

Université de Moncton à Shippagan, New Brunswick, Canada

Abstract: Based on the notion of mutual information between the components of a 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; Copula; Optimal quantization.

The authors are grateful to anonymous referees for a careful reading and detailed revision of the original manuscript. Their valuable comments contributed to improve the presentation of our work.

Corresponding Author's Address: B. Colin, Département de Mathématiques, Faculté des Sciences, Université de Sherbrooke, Sherbrooke J1K-2R1 Québec, Canada, tel : (819) 821-8000 ext. 62012, email: bernard.colin@usherbrooke.ca.

Published online: 27 April 2013

1. Introduction and Motivation

In data analysis, we frequently encounter in practice as in social, behavioral and biomedical sciences, as well as in marketing, education, public health, some situations where we have to deal with a great number of variables which are continuous for some of them while the others are categorical. In this case, according to the selected data analysis model, these continuous variables are often discretized in a small number of categories or classes. In this paper we address the problem of finding an optimal discretization (or quantization or coding) of the support $\mathcal{S}_{\mathbb{P}_X}$ of a continuous multivariate distribution, to retain as much as possible the stochastic dependence between the variables. More precisely, let us suppose that $X = (X_1, X_2, ..., X_k)$ is a random vector with values in $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k}, \mathbb{P}_X)$ where $\mathbb{P}_X$ is the probability measure of X, and let $\mathcal{S}_{\mathbb{P}_X} = \mathbb{R}^k$ is the support of $\mathbb{P}_X$. If $n = n_1 n_2 ... n_k$ is the product of k given integers, we consider a partition $\mathcal{P}$ of $\mathcal{S}_{\mathbb{P}_X}$ in n elements or classes where we suppose, for practical reasons, that $\mathcal{P}$ is obtained as a "product-partition" deduced from partitions $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_k$ of the supports of the marginal probability measures in, respectively, $n_1, n_2, ..., n_k$ intervals. Then, using a mutual information criterion, we propose to choose the set of all intervals such that the quantization of the support $\mathcal{S}_{\mathbb{P}_X}$ which results from this choice, retains, as much as possible, the stochastic dependence between the components of the random vector X.

Here are some examples for which such an optimal discretization might be desirable.

i) Let us suppose that we have a sample of individuals on which we observe the following variables:

$$X = \text{age}, Y = \text{salary}, Z = \text{socioprofessional group} .$$

If we want to take into account the variables simultaneously as, for example, in multiple correspondence analysis, we have to put them on the same form by means of a discretization of the first two ones. Instead of the usual independent categorization of the variables X and Y in a given number of classes (p for X and q for Y), it would be more relevant, using their stochastic dependence, to categorize simultaneously X and Y in pq classes (referred sometimes as a (p, q)-partition), in order to preserve as much as possible the dependence between them. Moreover, and depending on the values taken by the categorical variable Z, the (conditional) discretization of the random vector (X, Y) must differ, from one class to the others, to take into account the stochastic dependence between the continuous random variables and the categorical one. Usually, we do not take care of this dependence in creating classes for continuous random variables. However, the dependence between

X=age and Y=salary are certainly quite different between the socioprofessional groups.

ii) Here is another situation for which this kind of discretization may be necessary: let us suppose that X and Y are two continuous random variables, for which a χ^2 test of independence must be conducted on a sample of (X, Y), by means of partitions of X and Y in, respectively, p and q classes of equal widths, as it is often the case. Although it is usual to conduct such a test by means of a $p{\times}q$ contingency table, one can show that this choice of classes, instead of those arising from an optimal coding, is in favour of the hypothesis of independence, which is certainly not, for objectivity reasons, to be desired. In other words, by taking a different partition than the optimal one, one may increase the possibilities of accepting the null hypothesis $\mathcal{H}_0$ of independence (see end of Example 1 in Section 4).

iii) In large surveys or in data warehouses, we often have to deal with a large number of variables, but for data privacy reasons, we also have to create classes for the observations. So, does there exist a way to create a given number of classes which preserves as much as possible the stochastic dependence between the random variables? Surely, this question may be of some interest when dealing with forecasting models.

iv) In sampling theory when the population under consideration is very heterogeneous, it may be found impossible to get a sufficiently accurate estimation of a given parameter by taking a simple random sample from the entire population. Usually, in order to improve the accuracy of the estimation, one can resort to some stratification of the population in more internally homogeneous strata. But if the purpose of stratification is to achieve higher accuracy, one question arises for which answers must be found: how should the strata be made and how many of them should be made? Often stratification is performed on the basis of auxiliary information provided by a whole set of continuous random variables. For example in a two-way stratification strategy involving two continuous random variables X and Y, one has to find by means of a given criterion, the “cut-off” points for each variable in order to create an optimal stratification. From this point of view, it is surely of interest to consider the one that minimizes the loss of information between the random variables X and Y, before and after stratification. Indeed, for total, mean or ratio estimators, the accuracy of the estimations may depends on the stochastic dependence between X and Y and it seems natural to retain it, as much as possible, in the selected stratification.

Problems related to the quantization of the support of a probability measure, have been considered, in the last two decades, by many authors.

Among them are Österreicher (2003), Pötzelberger (2003), Liese and Varda (2006), Moddemeijer (1989; 1999), Varda (2002), Darbellay (1999), Darbellay and Vajda (1999), Liese, Morales and Vajda (2006), Haussler (1997), Morales, Pardo and Vajda (1995) and Beirlant, Dudewicz, Györfi and van der Meulen (1997). Among all these authors, some of them used quantizations of the support of a given probability measure, in order to define non-parametric estimators of the mutual information, mainly in the *Kullback-Leibler* sense. Then they consider convergence and asymptotic properties of these estimators (unbiasedness, consistency, or order of consistency of entropy estimators). Others have considered the problem of quantization for coding needs in a classical information-theoretical framework and have also studied some asymptotical properties as convergences and sufficiency. Finally, some of them used quantization techniques for clustering and classification purposes. As one can notice it, these approaches differ from ours.

After recalling in Section 2 the concepts of generalized divergence between probability measures and of mutual information between random variables, we propose in Section 3, which is the original part of this paper, a criterion in order to partition the support $\mathcal{S}_{\mathbb{P}_X}$, which consists in minimizing the loss of mutual information rising from the data reduction process. Furthermore, we prove the existence of an optimal partition. We illustrate in Section 4 this procedure on some examples and we compare, on the basis of this criterion, the optimal partition to some others having the same marginal number of classes. Finally, some conclusions will follow in Section 5.

2. Theoretical Framework

2.1 φ-Divergence

Since results presented in this paper are based on the existence of the mutual information between random variables, some well-known theoretical results are recalled for convenience. We will assume from now, that equalities and inequalities will be taken in the "almost surely" $(a.s.)$ sense.

Let $(\Omega, \mathcal{F}, \mu)$ be a measured space and let μ_1 and μ_2 be two probability measures defined on $\mathcal{F}$, such that $\mu_i \ll \mu$ for $i = 1, 2$. One defines the φ-divergence or the generalized divergence (Csiszár 1967) between μ_1 and μ_2, by:

$$I_\varphi\left(\mu_1, \mu_2\right) = \int \varphi\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_2 = \int \varphi\left(\frac{f_1}{f_2}\right) f_2 d\mu \, ,$$

where $\varphi(t)$ is a convex function from $\mathbb{R}_+\backslash\{0\}$ to $\mathbb{R}$ and where $f_i = \frac{d\mu_i}{d\mu}$ for $i = 1, 2$. Obviously, $I_\varphi\left(\mu_1, \mu_2\right)$ does not depend on the choice of μ. Moreover, for homogeneous models, one can write this expression as:

$$I_{\varphi}\left(\mu_1,\mu_2\right)=\int \frac{d\mu_2}{d\mu_1}\varphi\left(\frac{d\mu_1}{d\mu_2}\right)d\mu_1=\int \frac{f_2}{f_1}\varphi\left(\frac{f_1}{f_2}\right)f_1 d\mu\,.$$

The next lemma, due to Csiszár (1967; 1972; 1977) (see also Ali and Silvey 1966; Zakai and Ziv 1973), is particularly important insofar as it ensures, in a general theoretical framework, the existence of $I_{\varphi}\left(\mu_1,\mu_2\right)$ and shows that $I_{\varphi}\left(\mu_1,\mu_2\right)\geq\varphi\left(1\right)$, where equality holds if and only if $\mu_1=\mu_2$ and φ is strictly convex at $t_0=1$.

Lemma. *Let $\varphi(t)$ be a convex function from $\mathbb{R}_+\setminus\{0\}$ to $\mathbb{R}$, with the following usual conventions:*

$$\varphi(0)=\lim_{t\to0^+}\varphi(t)\ ;\ 0\varphi(\tfrac{0}{0})=0\ ,$$

$$0\varphi(\tfrac{a}{0})=\lim_{\delta\to0^+}\delta\varphi(\tfrac{a}{\delta})=a\lim_{\delta\to0^+}\delta\varphi(\tfrac{1}{\delta})\quad\text{for any } a>0\ ,$$

and let $(\Omega,\mathcal{F},\mu)$ be any measured space (one will suppose however that the measure μ is finite or σ-finite).

i) If α and β are two non-negative measurable functions defined on $(\Omega,\mathcal{F},\mu)$, then:

$$\int \mathbb{I}_A\beta\varphi\left(\frac{\alpha}{\beta}\right)d\mu\ ,$$

is defined for each $A\in\mathcal{F}$ on which α and β are integrable.

ii) Moreover, if for such a set A, $\int\mathbb{I}_A\beta d\mu$ is strictly positive and if $\varphi(t)$ is strictly convex at:

$$t_0=\frac{\int\mathbb{I}_A\alpha d\mu}{\int\mathbb{I}_A\beta d\mu}\ ,$$

then:

$$\int \mathbb{I}_A\beta\varphi\left(\frac{\alpha}{\beta}\right)d\mu\geq\left(\int\mathbb{I}_A\beta d\mu\right)\varphi\left(\frac{\int\mathbb{I}_A\alpha d\mu}{\int\mathbb{I}_A\beta d\mu}\right)>-\infty\ ,$$

with equality if and only if $\alpha=t_0\beta$, μ-a.s. on A.

The following table shows the usual measures of φ-divergence:

$\varphi(x)$	Name
$(x-1)^2$	χ^2
$x\ln x;\,(x-1)\ln x$	*Kullback* and *Leibler*
$\lvert x-1\rvert$	Distance in variation
$(\sqrt{x}-1)^2$	*Hellinger*
$1-x^{\alpha}\ \ 0<\alpha<1$	*Chernoff*
$\lvert 1-x^{1/m}\rvert^{m}\ \ m>0$	*Jeffreys*

The two most popular functions in practice are: $\varphi(x) = (x-1)^2$ in probability and statistics for its χ^2 similarity and: $\varphi(x) = x \ln x$ or $(x-1)\ln x$ in information theory in order to define the concept of relative entropy. Otherwise, distance in variation can be used in statistical models based on the L^1-norm, whereas the *Hellinger* distance appears in some data analysis models as "spherical correspondence analysis". The others are sometimes used in discrimination and classification models. For more details and for other class of divergences see, for instance, Goël (1981), Adhikari and Joshi (1956), Aczél and Daróczy (1975), Rényi (1959; 1961) and Österreicher (2003). Note that for all φ as they appear just above, one has $\varphi(1) = 0$.

2.2 Mutual Information

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X_1, X_2, ..., X_k$ be k random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in measured spaces $(\mathcal{X}_i, \mathcal{F}_i, \lambda_i), i = 1, 2, ..., k$. Let us denote respectively by $\mathbb{P}_X = \mathbb{P}_{X_1, X_2, ..., X_k}$ and by $\otimes_{i=1}^k \mathbb{P}_{X_i}$ the probability measures defined on the product space $\left(\times_{i=1}^k \mathcal{X}_i, \otimes_{i=1}^k \mathcal{F}_i, \otimes_{i=1}^k \lambda_i\right)$, equal to the joint probability measure and to the product of the marginal ones, which are supposed to be absolutely continuous with respect to the product measure $\lambda = \otimes_{i=1}^k \lambda_i$.

Definition. The φ-mutual information or the mutual information between the random variables $X_1, X_2, ..., X_k$, is given by:

$$\begin{aligned} \mathcal{I}_\varphi\left(X_1, X_2, ..., X_k\right) &= I_\varphi\left(\mathbb{P}_X, \otimes_{i=1}^k \mathbb{P}_{X_i}\right), \\ &= \int \varphi\left(\frac{d\mathbb{P}_X}{d\left(\otimes_{i=1}^k \mathbb{P}_{X_i}\right)}\right) d\left(\otimes_{i=1}^k \mathbb{P}_{X_i}\right), \\ &= \int \varphi\left(\frac{f_1}{f_2}\right) f_2 d\lambda, \end{aligned}$$

where f_1 and f_2 are respectively the probability density functions of the measures $\mathbb{P}_X$and $\otimes_{i=1}^k \mathbb{P}_{X_i}$ with respect to $\lambda = \otimes_{i=1}^k \lambda_i$.

In many cases, the space $\mathcal{X}_i$ is, for every i, either the real line $\mathbb{R}$ endowed with the *Lebesgue* measure, or a discrete space endowed with the counting measure. If $\varphi(1) = 0$, then $\mathcal{I}_\varphi\left(X_1, X_2, ..., X_k\right) \geq 0$ where the equality holds if and only if the random variables $X_1, X_2, ..., X_k$ are independent.

Numerous properties of the mutual information may be found among others, in: Pinsker (1961), McEliece (1977), Csiszár (1967; 1977), Rényi

(1961) and Gavurin (1968). One of them, known as the "data-processing theorem" is of major interest for approximation purposes and one of its versions may be expressed as follows:

Theorem 1. *If for every $j = 1, 2, ..., k$, the functions g_j from $(\mathcal{X}_j, \mathcal{F}_j)$ to $(\mathcal{Y}_j, \mathcal{G}_j)$ are measurable, then one has*:

$$\mathcal{I}_\varphi (Y_1, Y_2, ..., Y_k) \leq \mathcal{I}_\varphi (X_1, X_2, ..., X_k) \ ,$$

where $Y_j = g_j (X_j)$ with equality if for every $j = 1, 2, ..., k$, the functions g_j are one-to-one.

This result shows that any transformation of the initial variables never leads to a gain of mutual information.

3. Optimal Partition

3.1 Mutual Information Explained by a Partition

Let us suppose that the random vector X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ has values in $\left(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k}\right)$ and let $\mathbb{P}_X$ be its probability measure ($\mathbb{P}_X \ll \lambda$ where λ is the *Lebesgue* measure on $\mathbb{R}^k$), whose support $\mathcal{S}_{\mathbb{P}_X}$ may be assumed of the form $\times_{i=1}^k [a_i, b_i]$ where $-\infty < a_i < b_i < \infty$ for every $i = 1, 2, ..., k$ (as shown by Remarks 2 and 3 thereafter, this is not of real importance).

Given k integers $n_1, n_2, ..., n_k$, let $\mathcal{P}_i$, for $i = 1, 2, ..., k$, be a partition of $[a_i, b_i]$ in n_i intervals $\{\gamma_{ij_i}\}$ such that:

$$a_i = x_{i0} < x_{i1} < ... < x_{i(n_i-1)} < x_{in_i} = b_i \ ,$$

with:

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

The "product-partition" $\mathcal{P} = \otimes_{i=1}^k \mathcal{P}_i$ of $\mathcal{S}_{\mathbb{P}_X}$ in $n = n_1 n_2 ... n_k$ rectangles of $\mathbb{R}^k$ is then defined by:

$$\begin{aligned} \mathcal{P} \ &= \ \{\gamma_{1j_1} \times \gamma_{2j_2} \times ... \times \gamma_{kj_k}\} = \{\times_{i=1}^k \gamma_{ij_i}\} \\ & \qquad \text{where for every } i\text{: } j_i = 1, 2, ..., n_i \ . \end{aligned}$$

If $\sigma(\mathcal{P})$ denotes the σ-algebra generated by $\mathcal{P}$, the restriction of $\mathbb{P}_X$ to $\sigma(\mathcal{P})$ is given by:

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

whose marginal are, for every $i = 1, 2, ..., k$:

$$\mathbb{P}_X(\times_{r=1}^{i-1} [a_r, b_r] \times \gamma_{ij_i} \times_{r=i+1}^k [a_r, b_r]) = \mathbb{P}_{X_i}(\gamma_{ij_i}) \ .$$

The mutual information, denoted by $\mathcal{I}_\varphi(\mathcal{P})$, explained by the partition $\mathcal{P}$ of the support $\mathcal{S}_{\mathbb{P}_X}$ is then given by:

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

From a practical point of view, we may consider for each $i = 1, 2, \ldots, k$ a simple random variable defined on $[a_i, b_i]$ and given by:

$$\xi_i^{\mathcal{P}} = \sum\nolimits_{j_i=1}^{n_i} \alpha_{ij_i}\mathbb{I}_{\gamma_{ij_i}} \, , \alpha_{ij_i} \in \mathbb{R} \, , j_i = 1, 2, \ldots, n_i \, ,$$

for which we suppose that all the α_{ij_i} are different (the α_{ij_i} may be regarded as the label of each element of $\mathcal{P}_i$). Thus, the random vector $\xi^{\mathcal{P}} = (\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}, \ldots, \xi_k^{\mathcal{P}})$ is defined on the measurable partition $\mathcal{P}$ of $\mathcal{S}_{\mathbb{P}_X}$ and its joint and marginal probability measures $\mathbb{P}_{\xi^{\mathcal{P}}} = \mathbb{P}_{\xi_1^{\mathcal{P}},\xi_2^{\mathcal{P}},\ldots,\xi_k^{\mathcal{P}}}$ and $\mathbb{P}_{\xi_i^{\mathcal{P}}}$, $i = 1, 2, \ldots, k$, are nothing else than those obtained just above from the restriction of $\mathbb{P}_X$ to $\sigma(\mathcal{P})$. Therefore:

$$\mathcal{I}_\varphi\left(\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}, \ldots, \xi_k^{\mathcal{P}}\right) = \int \varphi\left(\frac{d\mathbb{P}_{\xi^{\mathcal{P}}}}{d\left(\otimes_{i=1}^k \mathbb{P}_{\xi_i^{\mathcal{P}}}\right)}\right) d\left(\otimes_{i=1}^k \mathbb{P}_{\xi_i^{\mathcal{P}}}\right) = \mathcal{I}_\varphi(\mathcal{P}) \, .$$

Since by the mean of the "data-processing theorem" (Csiszár 1967) one has, for any partition $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ of $[a_1, b_1], [a_2, b_2], \ldots, [a_k, b_k]$:

$$\mathcal{I}_\varphi\left(X_1, X_2, \ldots, X_k\right) \geq \mathcal{I}_\varphi\left(\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}, \ldots, \xi_k^{\mathcal{P}}\right) \, ,$$

one can deduce, as a corollary, that the mutual information loss, $\mathcal{I}_\varphi\left(X_1, X_2, \ldots, X_k\right) - \mathcal{I}_\varphi(\mathcal{P})$, rising from discretization of the support $\mathcal{S}_{\mathbb{P}_X}$ of $\mathbb{P}_X$, is nonnegative.

Remark 1. Even if we do not consider in this paper asymptotical properties, it will be noticed that, under very mild conditions, $\mathcal{I}_\varphi\left(\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}, \ldots, \xi_k^{\mathcal{P}}\right)$ converges in probability to $\mathcal{I}_\varphi\left(X_1, X_2, \ldots, X_k\right)$ as $n_i \to \infty$ for every $i = 1, 2, \ldots, k$, with $\sup_{j_i} l(\gamma_{ij_i}) \to 0$ where $l(\gamma_{ij_i})$ is the length of the interval γ_{ij_i} (Serfling 1980).

Remark 2. If, in the absolutely continuous case, the support $\mathcal{S}_{\mathbb{P}_X}$ of $\mathbb{P}_X$ is not compact, as for example in the case of a k-multidimensional normal distribution ($\mathcal{S}_{\mathbb{P}_X} = \mathbb{R}^k$) or a k-multidimensional exponential distribution ($\mathcal{S}_{\mathbb{P}_X} = \mathbb{R}_+^k = [0, \infty)^k$), one can use the representation of $\mathbb{P}_X$ under its copula whose support is the unit cube $[0, 1]^k$ of $\mathbb{R}^k$ (for more details see

Fréchet 1951; Sklar 1959). More precisely, using the transformation from $\mathbb{R}^k$ to $[0,1]^k$ given by:

$$U_i = F_i(X_i) \text{ for } i = 1, 2, ..., k \ ,$$

the cumulative distribution function $C(u_1, u_2, ..., u_k)$ of the copula C associated to $\mathbb{P}_X$, is given, for every $u = (u_1, u_2, ..., u_k) \in [0,1]^k$, by the following expression:

$$C(u_1, u_2, ..., u_k) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_k^{-1}(u_k)) \ ,$$

where F and F_i, $i = 1, 2, ..., k$ are respectively, the cumulative distribution functions of the random vector X and of its components. Therefore, one can reduce this case to the previous one, since under this transformation, the mutual information remains invariant (case of equality in the "data processing theorem"). Furthermore, for numerical methods, this setting may be more appropriate than the original one for finding an optimal partition.

3.2 Existence of an Optimal Partition

For 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 $[a_i, b_i]$ in n_i disjoint intervals and $\mathcal{P}_{\mathbf{n}}$ be the class of partitions of $\mathcal{S}_{\mathbb{P}_X}$ given by:

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

where $\mathbf{n}$ is the multi index $(n_1, n_2, ..., n_k)$. Each element $\mathcal{P}$ of $\mathcal{P}_{\mathbf{n}}$ may be considered as a vector of $\mathbb{R}^{\Sigma_{i=1}^{k}(n_i+1)}$ having components :

$$\big(a_1, x_{11}, ..., x_{1(n_1-1)}, b_1, a_2, x_{21}, ..., x_{2(n_2-1)}, \\ b_2, ..., a_k, x_{k1}, ..., x_{k(n_k-1)}, b_k\big) \ ,$$

under the constraints:

$$a_i < x_{i1} < ... < x_{i(n_i-1)} < b_i \text{ for every } i = 1, 2, ..., k.$$

In order to obtain a partition $\mathcal{P}$ of $\mathcal{S}_{\mathbb{P}_X}$, for which the mutual information loss is minimum, we have to solve the following optimization problem:

$$\underset{P \in P_{\boldsymbol{n}}}{min} \left(\mathcal{I}_\varphi \left(X_1, X_2, ..., X_k\right) - \mathcal{I}_\varphi \left(\mathcal{P}\right)\right) \ ,$$

which is equivalent to:

$$\underset{P \in P_{\boldsymbol{n}}}{max} \mathcal{I}_\varphi \left(\mathcal{P}\right) = \underset{P \in P_{\boldsymbol{n}}}{max} \sum\nolimits_{j_1, j_2, ..., j_k} \varphi \left(\frac{\mathbb{P}_X(\times_{i=1}^{k} \gamma_{i j_i})}{\prod_{i=1}^{k} \mathbb{P}_{X_i}(\gamma_{i j_i})} \right) \prod\nolimits_{i=1}^{k} \mathbb{P}_{X_i}(\gamma_{i j_i}) \ .$$

Theorem 2. *If* $\mathbf{n} = (n_1, n_2, ..., n_k)$ *is a multi index for which* $n_1, n_2, ..., n_k$ *are* k *given integers, then: i) there exists an element* $\mathcal{P}^*$ *of* $\mathcal{P}_{\mathbf{n}}$*, called an "optimal partition", such that:*

$$\mathcal{I}_\varphi(\mathcal{P}^*) = \underset{P \in P_{\boldsymbol{n}}}{max} \mathcal{I}_\varphi(\mathcal{P}) \ .$$

ii) $\mathcal{I}_\varphi(\mathcal{P}^*)$ *is a nondecreasing function of the* n_i'*s.*

Proof. i) Without loss of generality, we may assume that the function $\mathcal{I}_\varphi(\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}, ..., \xi_k^{\mathcal{P}})$ of the variables x_{ij_i} for $i = 1, 2, ..., k$ and $j_i = 1, 2, ...,$ $(n_i - 1)$, is defined on a compact subset of $\mathbb{R}^{\Sigma_{i=1}^k (n_i - 1)}$ of the form:

$$S = \times_{i=1}^k S_{i,n_i} \ ,$$

where S_{i,n_i} is the following subset of $\mathbb{R}^{n_i - 1}$:

$$a_i = x_{i0} \leq x_{i1} \leq ... \leq x_{i(n_i-1)} \leq x_{in_i} = b_i \ .$$

Since equalities, like $x_{ij_i} = x_{i(j_i+1)}$, correspond to the merging of two adjacent classes, this involves a loss of mutual information (except if the latter is 0) compare to the case $x_{ij_i} < x_{i(j_i+1)}$. However, this loss will be strict, if the probability measure $\mathbb{P}_X$ does not have the property of local (or conditional) independence (see Darbellay, 1999), which is required from now, since it is usually the case in practice. Consequently, the maximum is attained at a point in the interior $\overset{\circ}{S}$ of S. As $\mathcal{I}_\varphi(\mathcal{P})$ is a real continuous function of the variables x_{ij_i} on a compact subset of $\mathbb{R}^{\Sigma_{i=1}^k (n_i - 1)}$, the existence of an optimal partition

$$\mathcal{P}^* = \otimes_{i=1}^k \mathcal{P}_i^*,$$

with $\mathcal{P}_i^* \in \mathcal{P}_{i,n_i}$, is then ensured. If we note by $x_{ij_i}^*$ the solutions of the optimization problem, then for every $i = 1, 2, ..., k$ one has:

$$\mathcal{P}_i^* = \{\gamma_{ij_i}^*\} \ ,$$

where:

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

from which it follows that the optimal partition $\mathcal{P}^*$ is given by:

$$\mathcal{P}^* = \{\times_{i=1}^k \gamma_{ij_i}^*\} \text{ where for every } i\text{: } j_i = 1, 2, ..., n_i \ .$$

with:

$$\mathcal{I}_\varphi(\mathcal{P}^*) = \underset{P \in P_{\boldsymbol{n}}}{max} \mathcal{I}_\varphi(\mathcal{P}) \ .$$

ii) To prove the second part of the theorem, it is enough to consider one of the unspecified $n_i's$. Indeed, let us suppose that for choosen i and j_i, one splits the interval $\gamma^*_{ij_i} = [x^*_{i(j_i-1)}, x^*_{ij_i}[$ in two disjoint intervals $[x^*_{i(j_i-1)}, x^*[$ and $[x^*, x^*_{ij_i}[$ with $x^*_{i(j_i-1)} < x^* < x^*_{ij_i}$, and let us suppose moreover that all others intervals remaining the same. This gives rise to a new partition $\mathcal{P}'$ with $n' = n + \prod_{j=1;j\neq i}^{k} n_j$ elements. If $\sigma(\mathcal{P}')$ and $\sigma(\mathcal{P}^*)$ denote the σ-algebras generated respectively by $\mathcal{P}'$ and $\mathcal{P}^*$, then one has obviously: $\sigma(\mathcal{P}^*) \subseteq \sigma(\mathcal{P}')$. So

$$\mathcal{I}_\varphi(\mathcal{P}^*) \leq \mathcal{I}_\varphi(\mathcal{P}') \text{ (Csiszár 1967) .}$$

Therefore:

$$\mathcal{I}_\varphi(\mathcal{P}^*) = \underset{P\in P_{n}}{max} \mathcal{I}_\varphi(\mathcal{P}) \leq \mathcal{I}_\varphi(\mathcal{P}') \leq \underset{P'\in P_{n'}}{max} \mathcal{I}_\varphi(\mathcal{P}') = \mathcal{I}_\varphi\left(\mathcal{P}'^*\right)$$

which is the expected result.
■

Remark 3. As observed in Remark 2, in the case where we use a copula, if $\mathcal{P}^* = \otimes_{i=1}^{k}\mathcal{P}_i^*$ is an optimal partition of $[0,1]^k$, then $\mathcal{Q}^* = \otimes_{i=1}^{k}\mathcal{Q}_i^*$, where for every i:

$$\mathcal{Q}_i^* = \{\gamma^*_{ij_i}\} \text{ with } x^*_{ij_i} = F_i^{-1}(u^*_{ij_i}) \text{ for } j_i = 1, 2, ..., n_i \ ,$$

is also an optimal partition for the support $\mathcal{S}_{\mathbb{P}_X}$, with in addition, using obvious notations:

$$\mathcal{I}_\varphi(\mathcal{P}^*) = \mathcal{I}_\varphi(\mathcal{Q}^*).$$

3.3 Some Practical Aspects

In practice, we do not know the probability measure $\mathbb{P}_X$ of the random vector X but only a number m of independent observations X^r for $r = 1, 2, ..., m$. In order to obtain an optimal partition of $\mathcal{S}_{\mathbb{P}_X}$, we have to estimate the probability measure $\mathbb{P}_X$. Two main cases may arise.

1. If $(X^1, X^2, ..., X^m)$ is a sample of size m of the random vector X, we only know the empirical probability measure:

$$\mathbb{P}_X^m = \frac{1}{m}\sum_{r=1}^{m} \delta_{X^r} \ ,$$

or, which is equivalent, the empirical probability distribution function:

$$F_m(x) = \frac{1}{m}\sum\nolimits_{r=1}^{m} \mathbb{I}_{]-\infty,x]}(X^r) \ \text{ for all } x \in \mathbb{R}^k \ ,$$

where $]-\infty, x] = \times_{i=1}^{k}]-\infty, x_i]$. Based on the observations, one can consider as in Bosq and Lecoutre (1987), some kernel estimators of $\mathbb{P}_X$ in order to estimate $\mathcal{I}_\varphi(X_1, X_2, ..., X_k)$ and then, deduce from this estimation, the required optimal partition. Even if this approach seems natural in such a usual situation, the quality of the result may rely heavily on the choice of the kernel as on the width of the windows. Furthermore, some questions are in order : since we have an estimator of the optimal partition, what about its properties as, for example: unbiasedness, consistency, convergences? Also, how this estimator performs when compare to other ones as, for example, the estimator based on the histogram approach? For the moment, we will not consider these questions.

2. In addition to the observations, we may assume that the probability measure $\mathbb{P}_X$ is a member of a given family (as, for instance, a member of a multivariate exponential family) and that its probability density function $f(x, \theta)$ depends upon an unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$. This semiparametric case, that we retain now, is more often tractable than the first one and it can be reduce to the estimation of the parameter θ as illustrated below.

Under the assumptions on $\mathbb{P}_X$, it seems natural to choose as an estimator of the mutual information, the one based on $f(x, \widehat{\theta})$ where $\widehat{\theta}$ is an estimator of θ (for example the maximum likelihood estimator). In other words, since $\mathcal{I}_\varphi (X_1, X_2, ..., X_k) = \mathcal{I}_\varphi (X; \theta)$ may be generally regarded as a continuous and twice differentiable function $g(\theta) = g(\theta_1, \theta_2, ..., \theta_d)$ from $\mathbb{R}^d$ to $\mathbb{R}$, we can choose as an estimator of $\mathcal{I}_\varphi (X; \theta)$, the following random variable:

$$\begin{aligned} \widehat{\mathcal{I}}_\varphi (X; \theta) &= \int_{\mathbb{R}^k} \varphi \left(\frac{f(x, \widehat{\theta})}{\prod_{i=1}^{k} f_i(x_i, \widehat{\theta})} \right) \left(\prod_{i=1}^{k} f_i(x_i, \widehat{\theta}) \right) dx \,, \\ &= \mathcal{I}_\varphi(X; \widehat{\theta}) \,, \end{aligned}$$

or, if U is the copula associated to X:

$$\widehat{\mathcal{I}}_\varphi (U; \eta) = \mathcal{I}_\varphi (U; \widehat{\eta}) = \int_{[0,1]^k} \varphi (c(u, \widehat{\eta})) \, du \,,$$

where $\widehat{\eta} = \eta(\widehat{\theta})$ is the estimator of the parameter $\eta = \eta(\theta)$ of the probability density function $c(u_1, u_2, ..., u_k, \eta)$ of $\mathbb{P}_U$.

It is clear that the properties of the estimator $\widehat{\mathcal{I}}_\varphi (X; \theta)$ depend closely on those of the estimator $\widehat{\theta}_n$ of θ, where $\widehat{\theta}_n$ is obtained from a sample of size

n of X. For instance, if $\widehat{\theta}_n$ is a consistent estimator of θ, we know (Lehmann 1991; Serfling 1980) that, under mild conditions, one has:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{L} \mathcal{N}(0, \mathcal{I}_X^{-1}) \ ,$$

where $\mathcal{I}_X$ is the *Fisher* information matrix relative to Θ. Hence it follows that, $\widehat{\mathcal{I}}_\varphi\left(X;\theta\right)$ is, under usual conditions of regularity on $g\left(\theta\right)$ (Serfling 1980), a consistent estimator of $\mathcal{I}_\varphi\left(X;\theta\right)$ and that:

$$\sqrt{n}(\mathcal{I}_\varphi(X, \widehat{\theta}_n) - \mathcal{I}_\varphi\left(X, \theta\right)) \xrightarrow{L} \mathcal{N}(0, {}^t\nabla g_\theta \mathcal{I}_X^{-1} \nabla g_\theta) \ ,$$

if ∇g_θ is different from zero. Moreover one has:

$$\mathcal{I}_\varphi(X, \widehat{\theta}_n) \xrightarrow{p} \mathcal{I}_\varphi\left(X, \theta\right) \text{ as } \widehat{\theta}_n \xrightarrow{p} \theta \text{ when } n \to \infty$$

4. Computational Aspects and Examples

Without loss of generality, one considers, in order to facilitate the presentation, the case of a bivariate random vector $X = (X_1, X_2)$ with probability density function $f(x_1, x_2)$ whose support is $[0,1]^2$. For each component, let respectively:

$$0 = x_{10} < x_{11} < x_{12} ... < x_{1i} < ... < x_{1(p-1)} < x_{1p} = 1 \ ,$$

and

$$0 = x_{20} < x_{21} < x_{22} ... < x_{2j} < ... < x_{2(q-1)} < x_{2q} = 1 \ ,$$

be the ends of intervals of two partitions of $[0,1]$ in respectively p and q elements. For $i = 1, 2, ..., p$ and $j = 1, 2, ..., q$, the probability measure of a rectangle $[x_{1(i-1)}, x_{1i}[\times[x_{2(j-1)}, x_{2j}[$ is given by:

$$\int_{x_{1(i-1)}}^{x_{1i}} \int_{x_{2(j-1)}}^{x_{2j}} f(x_1, x_2) dx_1 dx_2 = p_{ij} \ ,$$

while its product probability measure is expressed as:

$$\int_{x_{1(i-1)}}^{x_{1i}} f_1(x_1) dx_1 \times \int_{x_{2(j-1)}}^{x_{2j}} f_2(x_2) dx_2 = p_{i\cdot} p_{\cdot j} \ ,$$

with $p_{i\cdot} = \sum_{j=1}^{q} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{p} p_{ij}$. Then, the approximation of the mutual information between the random variables X_1 and X_2, conveyed by the discrete probability measure $\{p_{ij}\}$ is given by:

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \varphi\left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}\right) p_{i\cdot} p_{\cdot j} \ ,$$

and for given p, q and $f(x_1, x_2)$, one has to maximize the following expression:

$$\max_{\{x_{1i}\},\{x_{2j}\}} \sum_{i=1}^{p} \sum_{j=1}^{q} \left[\varphi \left(\frac{\int_{x_{1(i-1)}}^{x_{1i}} \int_{x_{2(j-1)}}^{x_{2j}} f(x_1,x_2) dx_1 dx_2}{\int_{x_{1(i-1)}}^{x_{1i}} f_1(x_1) dx_1 \times \int_{x_{2(j-1)}}^{x_{2j}} f_2(x_2) dx_2} \right) \right.$$
$$\left. \times \left(\int_{x_{1(i-1)}}^{x_{1i}} f_1(x_1) dx_1 \times \int_{x_{2(j-1)}}^{x_{2j}} f_2(x_2) dx_2 \right) \right],$$

with respect to the $x_{1i}'s$ and the $x_{2j}'s$, which is reduced to finding the maximum of a continuous and differentiable function of $p + q - 2$ variables defined on a compact subset of $\mathbb{R}^{p+q-2}$ under the constraints $0 < x_{11} < x_{12}... < x_{1i} < ... < x_{1(p-1)} < x_{1p} < 1$ and $0 < x_{21} < x_{22}... < x_{2j} < ... < x_{2(q-1)} < 1$. In order to solve this standard optimization problem, many methods are available and one has chosen the well-known method of feasible directions described in Zoutendijk (1960) and also in Bertsekas (1999).

Example 1. *We consider the function $\varphi(t) = (t-1)^2$ and a probability measure on $[0,1]^2$ which probability density function (with respect to the Lebesgue measure on $[0,1]^2$) given by:*

$$f(x_1, x_2) = (x_1 + x_2) \mathbb{I}_{[0,1]^2}(x_1, x_2) \ .$$

A straightforward calculation shows that:

$$\mathcal{I}_{(t-1)^2}(X_1, X_2) = 9.7 \times 10^{-3} \, .$$

As $\mathcal{I}_{(t-1)^2}(X_1, X_2)$ is a continuously differentiable function with respect to the variables (x_{ij_i}), we obtain easily the following optimal partition of $[0,1]^2$ for $p = q = 3$:

$$x_{11} = x_{21} = .2211 \ ; \ x_{12} = x_{22} = .54 \, .$$

Since the probability density function of the vector (X_1, X_2) is symmetric with respect to x_1 and x_2, the random variables $\xi_1^{\mathcal{P}^*}$ and $\xi_2^{\mathcal{P}^*}$ are identically distributed and the probability measures $\mathbb{P}_{\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}}$ and $\mathbb{P}_{\xi_1^{\mathcal{P}^*}} = \mathbb{P}_{\xi_2^{\mathcal{P}^*}}$ are given by:

$\mathbb{P}_{\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}}$	$[0, \ .2211[$	$[.2211, \ .54[$	$[.54, \ 1]$
$[0, \ .2211[$	.0108	.0346	.0895
$[.2211, \ .54[$	.0346	.0774	.1687
$[.54, \ 1]$	.0895	.1687	.3258
$\mathbb{P}_{\xi_1^{\mathcal{P}^*}} = \mathbb{P}_{\xi_2^{\mathcal{P}^*}}$	.1349	.2807	.5840

Let $\{p_{ij}\}, \{p_{i\cdot}\}$ *and* $\{p_{\cdot j}\}$ *be respectively the probability density functions of* $(\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}), \xi_1^{\mathcal{P}^*}$ *and* $\xi_2^{\mathcal{P}^*}$*. For* $\varphi(t) = (t-1)^2$ *one has:*

$$\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}\right) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}} - 1\,,$$

and after calculation one obtains: $\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}\right) = 7{\times}10^{-3}$ *which represents* 72% *of the initial mutual information. As comparison, in the case of a regular partition (each class has the same width), the probability measure* $\mathbb{P}_{\xi_1^{\mathcal{P}_{reg}}, \xi_2^{\mathcal{P}_{reg}}}$ *is given by:*

$\mathbb{P}_{\xi_1^{\mathcal{P}_{reg}}, \xi_2^{\mathcal{P}_{reg}}}$	$[0, .3333[$	$[.3333, .6666[$	$[.6666, 1]$
$[0, .3333[$	$.0370$	$.0740$	$.1111$
$[.3333, .6666[$	$.0740$	$.1111$	$.1481$
$[.6666, 1]$	$.1111$	$.1481$	$.1852$
$\mathbb{P}_{\xi_1^{\mathcal{P}_{reg}}} = \mathbb{P}_{\xi_2^{\mathcal{P}_{reg}}}$	$.2221$	$.3332$	$.4444$

while this probability measure is, for an "equipartition" (each marginal class has the same frequency), equals to:

$\mathbb{P}_{\xi_1^{\mathcal{P}_{equi}}, \xi_2^{\mathcal{P}_{equi}}}$	$[0, .4574[$	$[.4574, .7583[$	$[.7583, 1]$
$[0, .4574[$	$.0957$	$.1152$	$.1224$
$[.4574, .7583[$	$.1152$	$.1100$	$.1081$
$[.7583, 1]$	$.1224$	$.1081$	$.1028$
$\mathbb{P}_{\xi_1^{\mathcal{P}_{equi}}} = \mathbb{P}_{\xi_2^{\mathcal{P}_{equi}}}$	$.3333$	$.3333$	$.3333$

For this two last partitions one has:

$$\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}_{reg}}, \xi_2^{\mathcal{P}_{reg}}\right) = 5.44{\times}10^{-3}\ (56\%\ \textit{of}\ \mathcal{I}_{(t-1)^2}(X_1, X_2))\,,$$

in the first case and:

$$\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}_{equi}}, \xi_2^{\mathcal{P}_{equi}}\right) = 5.6{\times}10^{-3}\ (58\%\ \textit{of}\ \mathcal{I}_{(t-1)^2}(X_1, X_2))\,,$$

in the second one. Finally, it is obvious that if we choose another function φ*, we will usually obtain another optimal partition of* $[0,1]^2$*. As an illustration of the Remark 3, one can check that:*

$$c(u_1, u_2) = \frac{2(\sqrt{1+8u_1} + \sqrt{1+8u_2} - 2)}{\sqrt{1+8u_1}\sqrt{1+8u_2}} \mathbb{I}_{[0,1]^2}(u_1, u_2)\,,$$

is the probability density function of the copula corresponding to $f(x_1, x_2)$ *and that* $u_{11} = u_{21} = 0.135$ *and* $u_{12} = u_{22} = 0.416$ *are, in this case, solutions to the problem of optimization. Furthermore,* $x_{11} = .2211 = F_{X_1}^{-1}(u_{11})$ *and* $x_{12} = .54 = F_{X_1}^{-1}(u_{12})$ *with* $F_{X_1}(x_1) = \frac{x_1(x_1+1)}{2}$.

Now, we use this elementary example as a $(3{\times}3)$ two-way contingency table to illustrate the consequences of the choice of a $(3,3)$-partition of $[0,1]^2$ on a χ^2-test of independence. Let us suppose that one has n independent observations from a random vector (ξ_1, ξ_2) where ξ_1 and ξ_2 are two categorical variables with respectively p and q categories and let $N = \{n_{ij}\}$, be the corresponding contingency table. The usual χ^2-statistics for testing independence between ξ_1 and ξ_2 is, in a standard notation, given by:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q} n\frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} ,\\ &= n\left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}\right] = n\mathcal{I}_{(t-1)^2}\left(\xi_1, \xi_2\right) .\end{aligned}$$

This shows that, up to a factor n, χ^2 and $\mathcal{I}_{(t-1)^2}(\xi_1, \xi_2)$ are the same. So if, as above, the random vector (ξ_1, ξ_2) is obtained from a (p,q)-partition of the support of the continuous random vector (X_1, X_2) one has:

$$\chi^2_{\mathcal{P}^*} = n\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*}, \xi_2^{\mathcal{P}^*}\right) \geq \chi^2_{\mathcal{P}} = n\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}}, \xi_2^{\mathcal{P}}\right) ,$$

where $\chi^2_{\mathcal{P}^*}$ and $\chi^2_{\mathcal{P}}$ are respectively arising from an (p,q)-optimal partition $\mathcal{P}^*$, or from an unspecified (p,q)-partition $\mathcal{P}$. Therefore, for the usual χ^2 test of independence, the inequality:

$$\chi^2 \leq \chi^2_c ,$$

corresponding to the acceptance of the hypothesis $\mathcal{H}_0$ of independence between the variables X_1 and X_2, is more easily checked under $\mathcal{P}$ than under $\mathcal{P}^*$ and if:

$$\chi^2_{\mathcal{P}} < \chi^2_c < \chi^2_{\mathcal{P}^*} ,$$

$\mathcal{H}_0$ is accepted under $\mathcal{P}$ and rejected under $\mathcal{P}^*$, which is not very satisfactory. For simplicity reasons, it will be assumed in this example, that the above optimal partition has been obtained by the means of a density estimator, for example, and that an independent sample of (X_1, X_2) of size n, has given rise to joint empirical frequencies equal to $\mathbb{P}_{\xi_1^{\mathcal{P}_{reg}}, \xi_2^{\mathcal{P}_{reg}}}$ or to

$\mathbb{P}_{\xi_1^{\mathcal{P}_{equi}},\xi_2^{\mathcal{P}_{equi}}}$. Thus with $n = 1200$, one has $\chi^2_{0.9} = 7.78$ for a χ^2 with 4 degrees of freedom, and one easily checks that $\mathcal{H}_0$ is accepted under $\mathcal{P}_{reg}$ and $\mathcal{P}_{equi}$ while $\mathcal{H}_0$ is rejected under $\mathcal{P}^*$. Unsurprisingly, $\mathcal{H}_0$ is rejected under any one of these partitions as soon as n is greater than 1430 since, as n increases, the conditions for independence are more and more difficult to fulfill.

Example 2. *Let $X = (X_1, X_2) \sim \mathcal{E}_2(\theta)$ $(-1 \leq \theta \leq 1)$ be a bivariate exponential random vector, whose probability density function is given by:*

$$f(x_1,x_2) = e^{-x_1-x_2}\left[1+\theta-2\theta(e^{-x_1}+e^{-x_2}-2e^{-x_1-x_2})\right]\mathbb{I}_{\mathbb{R}_+^2}(x_1,x_2)\,,$$

and let $C(u_1,u_2)$ be its copula whose probability density function $c(u_1,u_2)$ is:

$$c(u_1,u_2) = [1+\theta(1-2u_1)(1-2u_2)]\mathbb{I}_{[0,1]^2}(u_1,u_2)\ .$$

This family of distributions, also known as Farlie-Gumbel-Morgenstern class, is used, among others, in reliability theory as a life-lengths joint distribution of dependent components of a system operating in a random environment. Others applications appear in Elandt-Johnson (1976). As an illustration of a practical case, this example shows how to discretize simultaneously in a given numbers of categories, the life-lengths scale of each component, in order to qualify for example, the reliability of the system from "very low" to "very high". For various φ and for $\hat{\theta} = .5$ and $-.75$, which are considered here as estimates of the parameter θ, one has:

$\varphi(t)$	$\mathcal{I}_\varphi(X_1,X_2)$	$\mathcal{I}_\varphi(X_1,X_2)$
	$\hat{\theta} = .5$	$\hat{\theta} = -.75$
$(t-1)^2$: *(χ^2-metrics)*	$2.77{\times}10^{-2}$	$6.25{\times}10^{-2}$
$t\ln t$: *(Kullback)*	$1.41{\times}10^{-2}$	$3.24{\times}10^{-2}$
$(t-1)\ln t$: *(Kullback-Leibler)*	$2.87{\times}10^{-2}$	$6.77{\times}10^{-2}$
$\|t-1\|$: *(distance in variation)*	$12.5{\times}10^{-2}$	$18.75{\times}10^{-2}$
$(\sqrt{t}-1)^2$: *(Hellinger)*	$0.71{\times}10^{-2}$	$1.68{\times}10^{-2}$
$-\frac{1}{t}\ln t$: *(Leibler)*	$4.78{\times}10^{-2}$	$13.48{\times}10^{-2}$

One can observe that, maybe except for the distance in variation, all the mutual information figures are, for each $\hat{\theta}$, of the same order of magnitude. For $\varphi(t) = (t-1)^2$, one obtains the following optimal partitions on $[0,1]^2$ for $\hat{\theta} = .5$ and for $\hat{\theta} = -.75$:

- $\hat{\theta} = .5\,;\, p = q = 5$

$\mathbb{P}_{\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}}$	$[0,.2[$	$[.2,.4[$	$[.4,.6[$	$[.6,.8[$	$[.8,1]$	$\mathbb{P}_{\xi_1^{\mathcal{P}^*}}$
$[0,.2[$	.0528	.0464	.0400	.0336	.0272	.2
$[.2,.4[$	.0464	.0432	.0400	.0368	.0336	.2
$[.4,.6[$	.0400	.0400	.0400	.0400	.0400	.2
$[.6,.8[$	.0336	.0368	.0400	.0432	.0464	.2
$[.8,1]$	.0272	.0336	.0400	.0464	.0528	.2
$\mathbb{P}_{\xi_2^{\mathcal{P}^*}}$	.2	.2	.2	.2	.2	

with: $\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}\right) = 2.56{\times}10^{-2}$ *(*92.2% *of the initial mutual information). In this case, due to uniform marginal and due to properties of symmetry of* $c(u_1,u_2)$ *on* $[0,1]^2$*, it is not surprising that* $\mathcal{P}^*$*,* $\mathcal{P}_{reg}$ *and* $\mathcal{P}_{equi}$*, are the same. In fact, for every integers* p *and* q*, one considers the function* $\mathcal{I}_{(t-1)^2}(\xi_1^{\mathcal{P}},\xi_2^{\mathcal{P}})$ *given by:*

$$\begin{aligned}
&\mathcal{I}_{(t-1)^2}(\xi_1^{\mathcal{P}},\xi_2^{\mathcal{P}})\\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}\left[\frac{\left[\int_{u_{1i-1}}^{u_{1i}}\int_{u_{2j-1}}^{u_{2j}}(1+\theta\,(1-2u_1)\,(1-2u_2))du_1du_2\right]^2}{(u_{1i}-u_{1i-1})(u_{2j}-u_{2j-1})}\right]-1\\
&= \theta^2\left(\sum\nolimits_{i=1}^{p}u_{1i}u_{1i-1}\,(u_{1i}-u_{1i-1})\right)\left(\sum\nolimits_{j=1}^{q}u_{2j}u_{2j-1}\,(u_{2j}-u_{2j-1})\right),
\end{aligned}$$

where $u_{10}=u_{20}=0$ *and* $u_{1p}=u_{2q}=1$*. It is then easy, though somewhat tedious, to prove that, regardless* θ*, the optimal choice of classes on* $[0,1]^2$ *is given by the Cartesian product of two regular partitions on* $[0,1]$ *with respectively* p *and* q *classes. Moreover, the value of the maximum is:*

$$\begin{aligned}
\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}\right) &= \frac{\theta^2pq(p+2)(q+2)}{9(p+1)^2(q+1)^2}\\
&= \mathcal{I}_{(t-1)^2}\left(X_1,X_2,\theta\right)\left[\frac{p(p+2)q(q+2)}{(p+1)^2(q+1)^2}\right],
\end{aligned}$$

and, as $p\to\infty$ *and* $q\to\infty$*, then* $\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}\right)\to\mathcal{I}_{(t-1)^2}\left(X_1,X_2,\theta\right)$ $=\frac{\theta^2}{9}$ *as one can check easily.*

Since $u_1=F_1(x_1)=1-e^{-x_1}$ *and* $u_2=F_2(x_2)=1-e^{-x_2}$*, it follows that for every* i *and* j *one has:* $x_{1i}=F_1^{-1}(u_{1i})$ *and* $x_{2j}=F_2^{-1}(u_{2j})$*. Therefore, for* $p=q=5$*, the optimal choice of classes on* $\mathbb{R}_+^2$*, is rising from two identical partitions of* $[0,\infty[$ *given by:*

$$[0,.223[,[.223,.511[,[.511,.916[,[.916,1.609[,[1.609,\infty[\;.$$

- $\hat{\theta}=-.75\;;p=4,q=5$

$\mathbb{P}_{\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}}$	$[0,.2[$	$[.2,.4[$	$[.4,.6[$	$[.6,.8[$	$[.8,1]$	$\mathbb{P}_{\xi_1^{\mathcal{P}^*}}$
$[0,.25[$	.0275	.0388	.0500	.0612	.0725	.25
$[.25,.5[$	.0425	.0463	.0500	.0537	.0575	.25
$[.5,.75[$	.0575	.0538	.0500	.0462	.0425	.25
$[.75,1]$	.0725	.0613	.0500	.0387	.0275	.25
$\mathbb{P}_{\xi_2^{\mathcal{P}^*}}$	.2	.2	.2	.2	.2	

with: $\mathcal{I}_{(t-1)^2}\left(\xi_1^{\mathcal{P}^*},\xi_2^{\mathcal{P}^*}\right) = 5.62\times10^{-2}$ (89.9% *of* $\mathcal{I}_{(t-1)^2}(X_1,X_2)$). *Then the optimal choice of classes on* $\mathbb{R}^{+^2}$ *is given by the Cartesian product of the elements of the partition:*

$$[0,.223[, [.223,.511[, [.511,.916[, [.916,1.609[, [1.609,\infty[\text{ , for } X_1\,,$$

by the elements of the partition:

$$[0,.29[, [.29,.69[, [.69,1.38[, [1.38,\infty[\text{ , for } X_2\,.$$

5. Conclusions

Let $X = (X_1, X_2, ..., X_k)$ be a random vector in $\mathbb{R}^k$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with a joint probability measure $\mathbb{P}_X$ absolutely continuous with respect to a measure λ (usually, λ is the *Lebesgue* measure of $\mathbb{R}^k$). 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 $\mathcal{S}_{\mathbb{P}_X}$ of $\mathbb{P}_X$ 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, $n_1, n_2, ..., n_k$ classes. This procedure allows us 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 or statistical inference tasks 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 practical point of view in mind, we have paid attention to the semiparametric case (Example 2) for which one can assume that the probability measure $\mathbb{P}_X$ is a member of a given family depending on an unknown parameter θ.

References

ACZÉL, J., and DARÓCZY, Z. (1975), *On Measures of Information and Their Characterizations,* New York: Academic Press.

ADHIKARI, B.P., and JOSHI, D.D. (1956), "Distance, Discrimination et Résumé exhaustif"*, Publications de l'Institut de Statistique de l'Université de Paris, 5*, 57–74.

ALI, S.M., and SILVEY, S.D. (1966), "A General Class of Coefficients of Divergence of One Distribution from Another"*, Journal of the Royal Statistical Society, B28*, 131–142.

BEIRLANT, J., DUDEWICZ, E.J., GYORFI, L., and VAN DER MEULEN, E.C. (1997), "Nonparametric Entropy Estimation: An Overview"*, International Journal of Mathematical and Statistical Sciences, 6 (1)*, 17–39.

BERTSEKAS, D.P. (1999), *Nonlinear Programming* (2nd ed.), Belmont MA: Athena Scientific.

BOSQ, D., and LECOUTRE, J.P. (1987), *Théorie de l'Estimation fonctionnelle,* London: Economica.

CSISZÁR, I. (1967), "Information-Type Measures of Difference of Probability Distributions and Indirect Observations", *Studia Scientiarum Mathematicarum Hungarica, 2*, 299–318.

CSISZÁR, I. (1972), "A Class of Measures of Informativity of Observation Channels"*, Periodica Mathematica Hungarica, 2(1-4)*, 191–213.

CSISZÁR, I. (1977), "Information Measures : A Critical Survey"*, Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Vol. A*, Prague: Publishing House of the Czechoslovak Academy of Sciences, pp. 73–86.

DARBELLAY, G.A. (1999), "An Estimator of the Mutual Information Based on a Criterion for Conditional Independence"*, Computational Statistics & Data Analysis, 32*, 1–17.

DARBELLAY, G.A., and VAJDA, I. (1999), "Estimation of the Information by an Adaptive Partitioning of the Observation Space", *IEEE Transactions on Information Theory 45(4)*, 1315–1321.

ELANDT-JONHSON, R.C. (1976), "Conditional Failure Time Distributions Under Competing Risk Theory with Dependent Failure Times and Proportional Hazard Rates", *Scandinavian Actuarial Journal, 1*, 37–51.

FRECHET, M. (1951), "Sur les tableaux de corrélation dont les marges sont données", *Annals of the University of Lyon, 3(14)*, 53–77.

GAVURIN, M.K. (1968), "On the Value of Information", *Vestuik Leningrad University Series, 4, 1963*, 27–34, and *Translation in Selected Translations in Mathematical Statistics and Probability, 7,* 193–202.

GOËL, P.K. (1981), "Information Measures and Bayesian Hierarcichal Models", Technical Report, #81-4-1, Department of Statistics, Purdue University, West Lafayette, IN.

HAUSSLER, D., and OPPER, M. (1997)"Mutual information, Metric Entropy and Cumulative Relative Entropy Risk", *The Annals of Statistics, 25(6),* 2451–2492.

LEHMAN, E.L (1991), *Theory of Point Estimation,* Pacific Grove CA: Wadsworth & Brooks.

LIESE, F., and VAJDA, I. (2006), "On Divergences and Information in Statistics and Information Theory", *IEEE Transactions on Information Theory, 52(10)*, 4394–4412.

LIESE, F., MORALES, D., and VAJDA, I. (2006), "Asymptotically Sufficient Partitions and Quantizations", *IEEE Transactions on Information Theory, 52(12)*, 5599–5606.

MCELIECE, R.J. (1977), "The Theory of Information Coding", in *Encyclopedia of Mathematics and Its Applications*, eds. R. Doran, M. Ismail, T.-Y. Lam, and E. Lutwak, Reading MA: Addison Wesley.

MODDEMEIJER, R. (1999), "A Statistic to Estimate the Variance of the Histogram-Based Mutual Information Estimator, Based on Dependent Pairs of Observations", *Signal Processing, 75*, 51–63.

MODDEMEIJER, R. (1989), "On Estimation of Entropy and Mutual Information of Continuous Distributions", *Signal Processing, 16,* 233–248.

MORALES, D., PARDO, L., and VAJDA, I. (1995), "Asymptotic Divergence of Estimates of Discrete Distributions", *Journal of Statistical Planning and Inference, 48*, 347–369.

ÖSTERREICHER, F., and VAJDA, I. (2003), "A New Class of Metric Divergences on Probability Spaces and Its Applicability in Statistics", *Annals of the Institute of Statistical Mathematics, 55(3)*, 639–653.

PINSKER, M.S. (1964), *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden-Day.

PÖTZLBERGER, K. (2003), "Asymptotic Quantization of Probability Distribution", *Analysis in Theory and Applications, 19(4)*, 355–364.

RÉNYI, A. (1959), "On Measures of Dependence"*, Acta Mathematica Hungararica, 10*, 441–451.

RÉNYI., A. (1961), "On Measures of Entropy and Information", *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, (1)*, Berkeley: University of California Press, pp. 547–561.

SERFLING, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

SKLAR, A. (1959), "Fonctions de répartition à n dimensions et leurs marges", *Publications de l'Institut de Statistique de l'Université de Paris, 8,* 229–231.

VAJDA, I. (2002), "On Convergence of Information Contained in Quantizied Observations", *IEEE Transactions on Information Theory, 48(8)*, 2163–2172.

ZAKAI, J., and ZIV, M. (1973), "On Functionals Satisfying a Data-Processing Theorem", *IEEE Transactions, IT-19*, 275–282.

ZOUTENDIJK, G. (1960), *Methods of Feasible Directions*, Amsterdam: Elsevier, and Princeton NJ: D. VanNostrand.