J. Appl. Prob. **36,** 607–610 (1999)
Printed in Israel

ON AN ENTROPY CONSERVATION PRINCIPLE

JÉRÔME MANUCEAU,*
MARYLÈNE TROUPÉ* AND
JEAN VAILLANT,* ** *University of Antilles-Guyane*

Abstract

We present an entropy conservation principle applicable to either discrete or continuous variables which provides a useful tool for aggregating observations. The associated method of modality grouping transforms a variable Z_1 into a new variable Z_2 such that the mutual information $I(Z_2, Y)$ between Y, a variable of interest, and Z_2 is equal to $I(Z_1, Y)$.

Keywords: Entropy; mutual information; covariable

AMS 1991 Subject Classification: Primary 62B10; 94A17
Secondary 94A15; 94A24

1. Introduction

Entropy-based methods are often used in different fields (see [2, 5, 8]) in order to take into account information carried simultaneously by several variables. Since the pioneer work presented by Shannon [9], many authors have studied the relationship between measures of entropy and statistical analyses (see for example [3, 5, 6, 7]). For example, Ebrahimi and Pellerey proposed in [1] a partial ordering of survival functions based on the differential entropy of residual lifetime distributions.

In a recent paper, Manuceau *et al.* [4] have presented an entropy-based method for analysing the influence of covariables on breast cancer survival time. In applying this procedure, it is often necessary to reduce the initial number of modalities to a value k by grouping k classes of consecutive modalities. One possibility is to use a maximum entropy principle so that these class frequencies are almost equal.

A more satisfactory method of modality grouping which provides a transformation of a variable Z into a modified variable Z' would be one which conserves the mutual information between Y and Z, say $I(Z, Y) = I(Z', Y)$, where Y is a variable of interest and $I(\cdot, \cdot)$ defined as below. This is the aim of the next paragraph.

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X and Y be two random variables taking values on the measure spaces $(\mathcal{E}_1, \mathcal{B}, \nu)$ and $(\mathcal{E}_2, \mathcal{A}, \mu)$ respectively. If:

(i) X (respectively Y) has a Radon–Nikodym derivative g (respectively f) with respect to ν (respectively μ),

(ii) for any x in $\mathcal{E}_1$, Y conditionally to $X = x$ has a Radon–Nikodym derivative f_x with respect to μ

Received 17 November 1997; revision received 30 April 1998.
* Postal address: UFR Sciences, Department of Mathematics, 97169 Pointe-à-Pitre, Guadeloupe, FWI.
**Email address: jean.vaillant@univ-ag.fr

then the *mutual information* between X and Y is

$$I(X, Y) = H(Y) - H(Y \mid X)$$

where

$$H(Y) = -\int_{\mathcal{E}_2} f(y) \ln(f(y)) \mu(\mathrm{d}y)$$

is the differential entropy of Y and

$$H(Y \mid X) = -\int_{\mathcal{E}_1} g(x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \nu(\mathrm{d}x)$$

is the differential entropy of Y conditional on X.

2. Entropy conservation principle

In this section, we give an entropy conservation result useful for data grouping. The associated method transforms a covariable X into a new covariable X' such that the differential entropy $H(Y \mid X')$ of Y conditional on X' equals $H(Y \mid X)$. The proposition presented is different from the information invariance proposition given by [3] for which X and Y have the same value set and are modified by the same transformation T. Let us consider the following lemma.

Lemma. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X be a random variable taking values on the measure space $(\mathcal{E}, \mathcal{B}, \nu)$. Let $K \in \mathcal{B}$ and $\alpha \in \mathcal{E}$. We denote by P_X the probability on $\mathcal{E}$ generated by X. Let X' be the random variable defined by*

$$\forall \omega \in \Omega, \quad X'(\omega) = X(w)\mathbf{1}_{X^{-1}(\mathcal{E}\setminus K)}(\omega) + \alpha \mathbf{1}_{X^{-1}(K)}(\omega). \tag{1}$$

If $P_X \ll \nu$, $\alpha \in K$ and $\nu(\{\alpha\}) \neq 0$, then $P_{X'} \ll \nu$.
The Radon–Nikodym derivative of $P_{X'}$ with respect to ν is then

$$\frac{\partial P_{X'}}{\partial \nu} = \frac{\partial P_X}{\partial \nu}\mathbf{1}_{\mathcal{E}\setminus K} + \frac{P_X(K)}{\nu(\{\alpha\})}\mathbf{1}_{\{\alpha\}}.$$

The following proposition provides a mutual information conservation principle:

Proposition. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X and Y be two random variables taking values on the measure spaces $(\mathcal{E}_1, \mathcal{B}, \nu)$ and $(\mathcal{E}_2, \mathcal{A}, \mu)$ respectively. We assume that X has a Radon-Nikodym derivative g with respect to ν. For any x in $\mathcal{E}_1$, we also assume that*

$$Y \text{ conditionally to } X = x$$

has a Radon-Nikodym derivative f_x with respect to μ. Let K be a non-empty set of $\mathcal{B}$. Then

$$\forall (x_1, x_2) \in K^2, \quad f_{x_1} = f_{x_2} \Longrightarrow I(X, Y) = I(X', Y)$$

where X' is a modification of X defined by:

$$X' = \begin{cases} X & \text{on} \quad X^{-1}(\mathcal{E}_1 \setminus K) \\ \alpha & \text{on} \quad X^{-1}(K) \end{cases}$$

where α is an element of K such that $\nu(\{\alpha\}) \neq 0$.

Proof. To show that $I(X, Y) = I(X', Y)$, we can simply show that $H(Y \mid X) = H(Y \mid X')$ since $I(X, Y) = H(Y) - H(Y \mid X)$.

$$\begin{aligned} H(Y \mid X) &= -\int_{\mathcal{E}_1} g(x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \nu(\mathrm{d}x) \\ &= -\int_{\mathcal{E}_1 \setminus K} g(x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \nu(\mathrm{d}x) \\ &\quad - \int_K g(x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \nu(\mathrm{d}x). \end{aligned}$$

Since $f_x(y)$ does not depend on x for x in K, we have

$$\begin{aligned} H(Y \mid X) = &-\int_{\mathcal{E}_1 \setminus K} g(x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \nu(\mathrm{d}x) \\ &- \int_K g(x) \nu(\mathrm{d}x) \int_{\mathcal{E}_2} f_x(y) \ln(f_x(y)) \mu(\mathrm{d}y) \end{aligned} \tag{2}$$

and the last term of the right member of (2) is equal to

$$-P_X(K) \int_{\mathcal{E}_2} f_K(y) \ln(f_K(y)) \mu(\mathrm{d}y)$$

where f_K stands for the single function f_x when $x \in K$.

Consequently, $H(Y \mid X) = H(Y \mid X')$ where X' is a random variable whose Radon–Nikodym derivative h with respect to ν is such that

$$h(x) = g(x) \mathbf{1}_{\mathcal{E}_1 \setminus K}(x) + \frac{P_X(K)}{\nu(\{\alpha\})} \mathbf{1}_{\{\alpha\}}(x) \quad \forall x \in \mathcal{E}_1.$$

From the above lemma, we know this is the case when X' equals X on $X^{-1}(\mathcal{E}_1 \setminus K)$, and is constant on $X^{-1}(K)$.

It is worth noticing that the different modifications X' of X generate the same sigma-algebra on $\mathcal{E}_1$.

Let us consider the following survival time example: Y is the survival time. X is a discrete covariable taking value in $\mathcal{E}_1$ influencing Y. For any modality $x \in \mathcal{E}_1$ of this covariable, f_x is the conditional probability density of Y associated with the survival curve conditional on x. If we have $f_{x_1} = f_{x_2}$ for (x_1, x_2) in $\mathcal{E}_1^2$, then these two modalities x_1 and x_2 can be grouped into a single modality since the two conditional survival curves are identical. Several groupings can be made in this way, which decreases the number of significant curves without modifying the mutual information between the survival time and the covariable. More precisely, let us set

$$\mathcal{E}_1 = [-n, n] \cap \mathbb{Z}^* \quad \text{and} \quad f_x(t) = \lambda(x) \, \mathrm{e}^{-\lambda(x)t}, \forall t \in \mathbb{R}^+$$

where $n \in \mathbb{N}^*$ and λ is a positive and even real function of x.

Then $\forall x \in \mathcal{E}_1$, $f_x = f_{-x}$, which implies that $X' = |X|$ is a modification of X such that $I(X, Y) = I(X', Y)$.

From X to X', the number of modalities of the covariable is divided by two.

3. Conclusions

The entropy conservation principle presented in this paper proves that aggregating observations or grouping modalities of a covariable is possible without losing information carried by this covariable on the variable of interest. One of the numerous possible applications of this principle is the aggregation of survival curves and the test of significant difference between them. Applying this method could minimize the number of relevant curves associated to different prognostic covariables.

References

[1] EBRAHIMI, N. AND PELLEREY, F. (1995). New partial ordering of survival functions based on the notion of uncertainty. *J. Appl. Prob.* **32,** 202–211.

[2] EL HASNAOUI, A. (1993). Le concept du gain d'information: une nouvelle approche en épidémiologie quantitative. Ph.D. Thesis, Université de Montpellier.

[3] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[4] MANUCEAU, J., TROUPÉ, M. AND VAILLANT, J. (1999). Information and prognostic value of some variables in the breast cancer. To appear in *European Series in Applied and Industrial Mathematics*.

[5] RAO C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya Ser. A.* **44,** 1–22.

[6] RAO, C. R. (1986). Generalization of ANOVA through entropy and cross entropy functions. In *Probability Theory and Mathematical Statistics*, Vol. 2. Sciences Press, Utrecht, pp. 477–494.

[7] RÉNYI, A. (1961). On measures of entropy and information. In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, CA, pp. 547–561.

[8] ROBERT, C. (1990). An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *J. Appl. Prob.* **27,** 303–313.

[9] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27,** 379–423 & 623–656.