

Data Redundancy Reduction Using Mutual Information

Riduzione di Ridondanza Usando la Funzione di Mutua Informazione

Silvio Damiani, Marco Congi

Statistics, Probability and Applied Statistics Dept. “La Sapienza” Univ. of Rome, Italy.
silvio.damiani@uniroma1.it marco.congi@uniroma1.it

Giovanni Calcagnini

Biomedical Engineering Lab. Italian National Institute of Health. Rome, Italy.
giovanni.calcagnini@iss.it

Riassunto: la frequenza di campionamento di segnali elettrocardiografici (ECG) è generalmente scelta, in base al teorema di Shannon, non inferiore al doppio della massima frequenza contenuta del segnale, risultando pertanto unica per un intero segmento. Tale scelta può condurre a sequenze temporali caratterizzate, da una significativa ridondanza. Obiettivo del presente lavoro è la messa a punto di una metodica di riduzione del campionamento temporale, basata sull'utilizzo della funzione di mutua informazione applicata a coppie temporalmente consecutive di campioni.

Keywords: data redundancy, mutual information, time series, electrocardiogram.

1. Introduction

Modern techniques for biomedical data collection, such as surface ECG mapping, produce massive data of which real informative content should be evaluated. The time-sampling of the data is usually carried out according to the Shannon theorem. This theorem defines a unique sampling frequency, based on the maximum frequency of the entire signal, without taking into account the information content of various segments of the signal. Thus, in particular portion of a signal the Shannon sampling frequency may result in a significant redundancy. Among various methods proposed in literature, Mutual Information (*m.i.*) has been suggested as a measure of the dependence among variables. P-wave is a particular portion of the ECG signal, expression of the electrical activity of the atria. This signal contains information able to predict the occurrence of certain pathologies such as atrial fibrillation. The aim of this paper is to evaluate the redundancy in surface P-waves by means of the mutual information function.

2. Methods and material

Statistical assumptions on the P-wave. The P-wave can be assumed as a stochastic process $\tilde{P}_t$ with $t=1,2,\dots, m$ where P_t is the true signal and U_t is a zero-mean noise: $\tilde{P}_t = P_t + U_t$. As for the second moment, the process is assumed to be not stationary. Let's name Γ the temporal variance and covariance matrix of the process: $VAR(\tilde{P}) = \Gamma = [\gamma_{ks}]$ with $k,s=1,2,\dots,m$. From the process $\tilde{P}_t$ we record n independent time series $(p_{1t},\dots,p_{kt},\dots,p_{nt})^T$ with $t=1,2,\dots,m$ and $n>m$. We define the estimates of the signal P_t and the matrix Γ :

$$\bar{p}_t = \frac{1}{n} \sum_{k=1}^n p_{k,t} \quad \text{for } t=1,2,\dots,m \quad (1)$$

$$G = \hat{V}AR(\tilde{P}_t) = [g_{ks}] \quad \text{where } g_{k,s} = \frac{1}{n-1} \sum_{i=1}^n (p_{k,i} - \bar{p}_k)(p_{s,i} - \bar{p}_s) \quad (2)$$

The extraction of a not- (time) redundant subset of $(p_{1t}, \dots, p_{kt}, \dots, p_{nt})^T$ will be presented.

Temporal dependence estimation by mutual information. The mutual information (*m.i.*) is a measure of the dependence between random variables (r.v.s). It takes into account the whole dependence structure of the variables. Let X and Y be continuous r.v.s with joint distribution $p(X,Y)$ and marginal distribution $p(X)$ and $p(Y)$. The mutual information $I(X,Y)$ is the relative entropy between the joint distribution and the product of the marginal distribution:

$$I(X,Y) = \iint_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy \quad (3)$$

Note that when X and Y are independent, $p(X,Y) = p(X)p(Y)$, so $I(X,Y) = 0$. Moreover the *m.i.* is invariant with respect to the translation. This implies that if we centre the r.v.s the *m.i.* doesn't change. An important interpretation of the *m.i.* is the following: $I(X,Y) = H(X) + H(Y) - H(X,Y)$ where $H(\bullet)$ is the Shannon entropy. Other important properties of the *m.i.* can be summarized by these statements:

$$i) I(X,Y) = H(X) - H(X|Y), \quad ii) I(X,Y) = H(Y) - H(Y|X) \quad iii) I(X,Y) = I(Y,X), \quad iv) I(X,X) = H(X).$$

The estimation of *m.i.* requires the knowledge of the distribution of $p(X,Y)$. The estimation of $p(X,Y)$, usually obtained by histograms, requires a large amount of data, often not available in real experiments. Alternatively, if the probability density is close to a Gaussian, it's possible to approximate the distribution by the Edgeworth expansion (Kendal, 1983). Given this result, Cardoso shows that *m.i.* can be estimate by:

$$I(\mathbf{X}) \approx \frac{1}{4} \sum_{ij \neq ii} (k_2^{ij}(\mathbf{X}))^2 + \frac{1}{48} \sum_{ijkl \neq iiii} (k_4^{ijkl}(\mathbf{X}))^2 \quad \text{with } i,j,k,l = 1,2 \quad (4)$$

where: $k_4^{i,j,k,l}(\mathbf{X}) = E(X_i X_j X_k X_l) - E(X_i X_j)E(X_k X_l) - E(X_i X_k)E(X_j X_l) - E(X_i X_l)E(X_j X_k)$, and $k_2^{i,j}(\mathbf{X}) = E(X_i X_j)$ are the 4th and 2nd cumulants of the zero-mean random vector $\mathbf{X} = [X_1, X_2]$. Note that (4) contains only off-diagonal terms, that is cross-cumulants. Since cross-cumulants of independent r.v.s identically vanish, it's not surprising to see that the *m.i.* is approximated by a sum of squared cross-cumulants. The procedure proposed is based on the assumption that if a couple of variables (P_i, P_j) with $i < j$ has a significant value of *m.i.*, one of the variables is predictable and thus redundant. A redundant variable can be removed from the original series. Since the mutual information is invariant to translation, there is the potential risk of removing variables carrying significant information. In order to avoid this risk, the procedure consists of two steps. First, variables are checked for their expected value at different times, then only those not significantly different are compared by means of their mutual information. In the first step we check if the difference $E(P_t) - E(P_{t+1})$, for each t , is statistically different from 0, by applying a t-Student test, if the spread results gaussian (Kolmogorov-Smirnov test), or the rank test (Wilcoxon-test) if the normality hypothesis is rejected ($p < 0.05$).

An alternative approach, which takes into account the multiple comparisons, would be, under normality assumption, the use of a multiple test (e.g. Sheffé, Tukey). In the 2nd step, over each constant trend partition, we select those variables to be removed on the basis of the estimation of the *m.i.* over the couple (P_i, P_j) with $i < j$. In order to evaluate whether a given *m.i.* value is significant we applied a Monte Carlo simulation, as follows. Let P_t and P_{t+1} two zero-mean r.v.s describing the P-wave at the time $t, t+1$ and $\mathbf{p}_t = (p_{1,t}, \dots, p_{k,t}, \dots, p_{n,t})^T$, $\mathbf{p}_{t+1} = (p_{1,t+1}, \dots, p_{k,t+1}, \dots, p_{n,t+1})^T$ are two n -dimensional samples drawn from P_t and P_{t+1} . Let $\hat{m}i_0$ the *m.i.* of P_t and P_{t+1} estimate with (4) using the sample cross cumulants. If we permute $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ we can accomplish the $(n!)^2$ links between P_t and P_{t+1} estimated by the *m.i.* of each shuffling. Note that this procedure doesn't change the marginal distributions but only the joint probability of P_t and P_{t+1} . If $\hat{m}i_0 > p_\alpha$, where p_α is an appropriate percentile of the *m.i.* average distribution, obtained from shuffled data, we can say that P_t and P_{t+1} are dependent. In this case it's possible to remove one of the two variables. The entropy can be evaluated using the property *iv*) of the *m.i.* and applying again the (4). Because the shuffling procedure does change the joint distribution of the original data, the approximation of the distribution operated by the Edgeworth expansion may somehow contribute to observe a difference in *m.i.* between original data and the shuffled counterparts. In order to evaluate and quantify this effect, of which the theoretical analysis is far beyond the aim of the present study, we calculated *m.i.* values for r.v.s with different distributions and degree of dependence.

3. Results

We simulated $n=1000$ P-wave with mean P_t $t=1,2,\dots,m$, and variance and covariance matrix (2) G/n . The matrix G has been estimated from a real recording. Each simulation consists of $m=490$ samples, which is a typical duration of human P-waves.

Table 1 illustrates the results of the test, which identified couples of variables having not different expected values, and thus eligible for the analysis by *m.i.*. The partition P1-P20 has a constant trend. Variables P20 and P21 have spread not equal to 0 (Wilcoxon test $p < 0.05$). To estimate the averaged distribution of *m.i.* we used 100 shuffling. From $(n!)^2$ available configurations, we drawn 100 rearrangements by random sampling with replacement, so that each permutation has the same probability of being drawn. The 95th percentile of this distribution was used to assess if the variables in the couple are dependent (*m.i.* observed $> 95^{\text{th}}$ of the *m.i.* average distribution - Table 2).

The choice of $N=100$ was based on the evaluation of the power of a hypothesis test on the *m.i.* average, obtained applying the Benjamin-Tchebycheff inequality. Over the

Table 1. Constant trend partitions

Spread	p-value
-----	-----
P17 - P18	0.477
P18 - P19	0.926
P19 - P20	0.129
P20 - P21	0.009
-----	-----

Table 2. *m.i.* over constant trend partitions

Number of the Shuffling	95 th perc. of <i>m.i.</i> average distribution	Matching	<i>m.i.</i> of original data (<i>m.i.</i> ₀)	Flag=1 if <i>m.i.</i> ₀ $> 95^{\text{th}}$ perc. of <i>m.i.</i> ave. distrib.	Variable of Matching with largest entropy
-----	-----	-----	-----	-----	-----
100	9.46E-17	P428 - P442	1.19E-17	0	P428
100	2.53E-17	P442 - P452	2.16E-19	0	P442
100	1.82E-16	P452 - P466	2.25E-16	1	P466
-----	-----	-----	-----	-----	-----

Table 3. Effect of data distribution and Edgeworth expansion on the estimation of $m.i.$

<i>Distribution of first variable</i>	<i>Distribution of second variable</i>	<i>Correlation Coefficient</i>	<i>5th perc. of m.i. ave. dis.</i>	<i>Average m.i.</i>	<i>95th perc. of m.i. ave. dist.</i>	<i>Observed m.i.</i>
GAUSSIAN	GAUSSIAN	0.589	2.65E-03	2.45E-02	7.37E-02	18.7E-02
GAUSSIAN	GAUSSIAN	0.287	1.29E-03	6.27E-03	1.24E-02	4.27E-02
GAUSSIAN	GAUSSIAN	independent	4.06E-04	2.58E-03	7.47E-03	4.30E-04
UNIFORM	UNIFORM	independent	3.04E-08	2.85E-06	1.13E-05	2.10E-06
UNIFORM	UNIFORM	0.700	1.52E-05	2.16E-04	5.67E-04	1.73E-03
GAUSSIAN	UNIFORM	0.036	9.35E-06	9.29E-05	2.77E-04	2.55E-05

Number of Shuffling: 100 - Number of samples: 500

trend constant partition P427-P471, the $m.i.$ procedure suggests to leave only the variables P428, P442 and P452. The variable P466 is removed because the observed $m.i.$ is larger than the 95th percentile of the $m.i.$ average distribution. At the end of the two steps the survived variables result 345 (almost 30% of reduction). The effect of different distributions and degrees of dependence is shown in table 3.

On simulated signals (fig. 1 and 2) we obtained a global reduction consistent with the signal complexity (30% for a biphasic wave, 47% for gaussian shape. Area information is preserved (0.22% and 1.02% differences from the original signals, respectively).

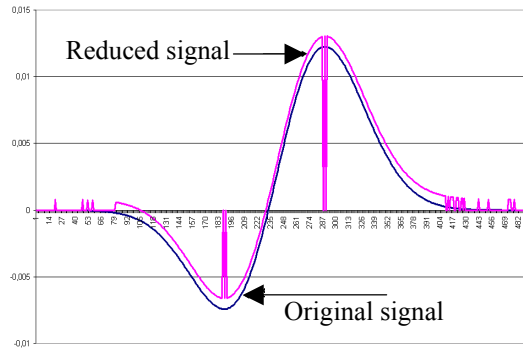


Figure 1: effect of time-reduction on a biphasic signal.

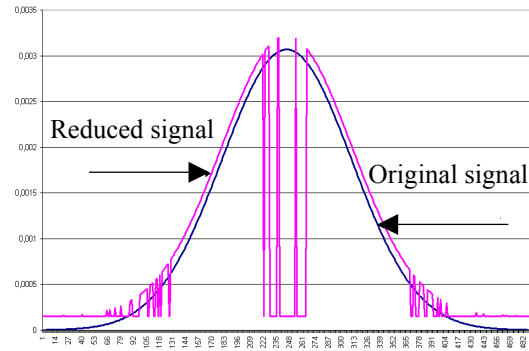


Figure 2: effect of time-reduction on a gaussian shape signal.

4. Conclusions

A new procedure to reduce a large set of time redundant variables has been proposed. This method is based on two steps: 1) detection of constant trend partitions of the signal; 2) application of Mutual Information to remove, in each partition, the redundant variables. The Mutual Information has been calculated using an approximation derived from Edgeworth expansion of the probability density. The reduction obtained reflects the complexity of the signal, i.e. higher the complexity, lower the reduction.

References

- Kendal M.G., Stuart A., Ord J.K. (1983) *The advanced theory of statistics*, Griffin, London.
 Cardoso J.F. (1999) High-Order Contrast for Independent Component Analysis, *Neural Computation*, 11, 157-192.
 Hyvärinen A., Karhunen J., Oja E. (2001) *Independent Component Analysis*, J. Wiley, N.Y.
 Papoulis A. (1991) *Probability, Random Variables and Stochastic Processes*, McGraw, N.Y.