AN INFORMATION THEORETICAL DELAY ESTIMATOR

R. Moddemeijer *

Abstract. *Mutual information is used in a model free and non-parametric general purpose method to define time-delays between stochastic signals. For each of two stationary stochastic signals we define a past and a future by cutting both infinite sample sequences into two parts: the past and the future. The time-delay is considered to be the time-shift between the cutting moments for which the mutual information between both infinite length past vectors and both infinite length future vectors reaches a minimum. For stationary stochastic signals and under certain convergence conditions this mutual information possesses one unique minimum. We present some theoretical elaborations and discuss this method in relation to other methods.*

1. Introduction.

For the estimation of time-delays between recordings of electroencephalogram (EEG) signals several methods are in use such as the cross-correlation method and the mutual information method of Mars [1,2]. These two methods search for the maximum correspondence of pairs of samples ($\underline{x}(t),\underline{y}(t+\tau)$) as a function of the time-shift τ, disregarding the dependence of subsequent sample pairs. Other more sophisticated methods, such as the maximum likelihood delay estimation method of Knapp and Carter [3] assume a signal model which is not realistic for EEG's. A large number of phase measurement methods, defined in the frequency domain and based on the same model, are reported [4].

We do not know of any general purpose method, implemented in the time domain, estimating time-delays using a minimum of a priori assumptions, other then cross-correlation. In this article we propose such a method.

2. Information theoretical definition of a delay.

In this section we present a mutual information based definition of a delay. Assume $\underline{x}_n$ and $\underline{y}_n$ are time discrete stationary

* University of Twente, department of Electrical Engeneering,
P.O.Box 217, NL-7500 AE Enschede, The Netherlands

stochastic processes. We consider a part of length 2M of the x signal surrounding the sample m: $m-M \le n \le m+M-1$. We cut this part of the stochastic signal $\underline{x}_n$ into two pieces: *the past vector* $\underline{P}_x(m;M)$ and *the future vector* $\underline{F}_x(m;M)$:

$$(2.1)\qquad \begin{aligned} \underline{P}_x(m;M) &= (\underline{x}_{m-M},\ldots,\underline{x}_{m-2},\underline{x}_{m-1})^T \\ \underline{F}_x(m;M) &= (\underline{x}_m,\underline{x}_{m+1},\ldots,\underline{x}_{m+M-1})^T \end{aligned}$$

We calculate *the mutual information between past and future*:

$$(2.2)\qquad I(\underline{P}_x(m);\underline{F}_x(m)) = \lim_{M\to\infty} I(\underline{P}_x(m;M);\underline{F}_x(m;M))$$

For infinite length vectors we omit the M in $\underline{P}_x(m;M)$ and $\underline{F}_x(m;M)$. For stationary processes this mutual information is independent of m. Under certain weak convergence conditions the limit exists [10]. If the limit exists the mutual information between past and future is a well defined quantity inherent to the stochastic process. It has the attractive property that it can reasonably well be approximated by choosing a finite M.

To define time-delays we extend the concept to a pair of an x and a y signal. We concatenate both past vectors of x and y, in which the segments of the y signal are j samples shifted towards the future with respect to the corresponding segments of x, according to the scheme:

$$(2.3)\qquad \underline{P}(m,j;M) = (\underline{P}_x^T(m;M),\underline{P}_y^T(m+j;M))^T$$

and in a similar way for the future. We determine, in the limit $M \to \infty$, the mutual information for x and y together between their joint past and their joint future as a function of j. If this mutual information reaches a minimum as a function of j, then we have created an optimal distinction between the past and the future. This means that there exists for this specific time-shift a joint process with a minimum transport of information between the past and the future. For this shift j the x and the y signal are, according to an information theoretical criterion, synchronous. We define: the y signal is j samples delayed with respect to the x signal if:

$$(2.4)\qquad I(\underline{P}(m,j);\underline{F}(m,j)) \le I(\underline{P}(m,i);\underline{F}(m,i)) \qquad \text{for } i \in \mathbb{Z}$$

with analoguous to (2.2):

$$(2.5) \qquad I(\underline{P}(m,j);\underline{F}(m,j)) = \lim_{M\to\infty} I(\underline{P}(m,j;M);\underline{F}(m,j;M))$$

This mutual information is symmetric for exchange of the roles of the past and the future. We execute the limiting process in two steps:

$$(2.6) \qquad I(\underline{P}(m,j);\underline{F}(m,j)) =$$

$$= \lim_{M_1,M_2\to\infty} \Big[H(\underline{F}(m,j;M_1)) - H(\underline{F}(m,j;M_1)|\underline{P}(m,j;M_2)) \Big]$$

$$= \lim_{M_1\to\infty} \Big[H(\underline{F}(m,j;M_1)) - M_1\cdot L \Big]$$

Due to the symmetry the joint limit is independent of the order in which we take these steps. Taking the limit for $M_2 \to \infty$ reduces the entropy of the future conditioned on the past to $M_1\cdot L$ [10]: M_1 times the entropy per sample pair $(\underline{x}_m, \underline{y}_{m+j})$ of the joint process, which is independent of the timeshift j.

We call this delay definition by a minimum mutual information between the past and the future an information theoretical delay definition: no other theory is involved than the information theory.

3. Discussion of the delay definition.

The delay estimate is unambiguous if there exists only one minimum of the mutual information function. This implies that no local maximum can exist. We prove unambiguity by proving the non-existence of that local maximum. Because of stationarity the mutual information is independent of a time-shift, so a local maximum is characterized by:

$$(3.1) \qquad \begin{aligned} &I(\underline{P}(m\;\;,j+1);\underline{F}(m\;\;,j+1)) \le I(\underline{P}(m+1,j);\underline{F}(m+1,j)) \\ &I(\underline{P}(m+1,j-1);\underline{F}(m+1,j-1)) < I(\underline{P}(m\;\;,j);\underline{F}(m\;\;,j)) \end{aligned}$$

Without loss of generality we assume $m = j = 0$. We separate the stochastic variables $\underline{x}_0$ and $\underline{y}_0$:

$$(3.2) \qquad \begin{aligned} &I(\underline{P}(0,0),\underline{y}_0;\underline{F}(1,0),\underline{x}_0) \le I(\underline{P}(0,0),\underline{x}_0,\underline{y}_0;\underline{F}(1,0)) \\ &I(\underline{P}(0,0),\underline{x}_0;\underline{F}(1,0),\underline{y}_0) < I(\underline{P}(0,0);\underline{F}(1,0),\underline{x}_0,\underline{y}_0) \end{aligned}$$

The mutual information is defined in the limit for $M \to \infty$, we write for finite M the mutual information as a sum and a difference of entropies and omit the joint entropy of $\underline{P}(0,0)$, $\underline{F}(1,0)$, $\underline{x}_0$ and $\underline{y}_0$. We rewrite the inequalities (3.2) for a finite M:

$$\begin{aligned}(3.3)\quad & H(\underline{P}(0,0;M),\underline{y}_0)+H(\underline{F}(1,0;M),\underline{x}_0) \le H(\underline{P}(0,0;M),\underline{x}_0,\underline{y}_0)+H(\underline{F}(1,0;M)) \\ & H(\underline{P}(0,0;M),\underline{x}_0)+H(\underline{F}(1,0;M),\underline{y}_0) < H(\underline{P}(0,0;M))+H(\underline{F}(1,0;M),\underline{x}_0,\underline{y}_0)\end{aligned}$$

These inequalities can be rewritten as:

$$\begin{aligned}(3.4)\quad & H(\underline{x}_0|\underline{F}(1,0;M)) \le H(\underline{x}_0|\underline{P}(0,0;M),\underline{y}_0) \\ & H(\underline{x}_0|\underline{P}(0,0;M)) < H(\underline{x}_0|\underline{F}(1,0;M),\underline{y}_0)\end{aligned}$$

In general if one adds conditions to an entropy, then the resulting conditional entropy is equal to or is less than the corresponding entropy without these conditions, so:

$$\begin{aligned}(3.5)\quad & H(\underline{x}_0|\underline{F}(1,0;M)) \le H(\underline{x}_0|\underline{P}(0,0;M),\underline{y}_0) \le H(\underline{x}_0|\underline{P}(0,0;M)) < \\ & < H(\underline{x}_0|\underline{F}(1,0;M),\underline{y}_0) \le H(\underline{x}_0|\underline{F}(1,0;M))\end{aligned}$$

Taking the limit for $M \longrightarrow \infty$ does not influence the contradiction. This proves that there can not exist any local maximum of the mutual information function, which proves unambiguity. For a finite M or an estimated mutual information function unambiguity is not assured.

Is our definition a legitimate definition of the delay $D(\underline{x},\underline{y})$ [5]? We have to prove that the definition satisfies the six criteria proposed for discrete processes:

1) $D(\underline{x},\underline{y}) = -D(\underline{y},\underline{x})$
2) $D(\underline{x},\underline{x}) = 0$
3) $D(\lambda\underline{x}+\mu,\underline{y}) = D(\underline{x},\underline{y})$ for all $\lambda,\mu \in \mathbb{R}$ and $\lambda \ne 0$
4) $D(T_k(\underline{x}),\underline{y}) = D(\underline{x},\underline{y}) + k$ for all $k \in \mathbb{Z}$
5) $D(R(\underline{x}),R(\underline{y})) = -D(\underline{x},\underline{y})$
6) if $\underline{n}$ is a stochastic signal independent of $\underline{x}$ and $\underline{y}$ then $D(\underline{x}+\underline{n},\underline{y}) = D(\underline{x},\underline{y})$

The operator T_k shifts the sequence of samples k positions to the past: $\underline{z} = T_k(\underline{x})$ means $\underline{z}_m = \underline{x}_{m+k}$ $m,k \in \mathbb{Z}$ and the operator R reverses the time: $\underline{z} = R(\underline{x})$ means $\underline{z}_m = \underline{x}_{-m}$ $m \in \mathbb{Z}$.

The information theoretical delay definition satisfies criterion 1) - 5) [10]. We have not been able to prove 6) in general. If for all i and M there exists a j such that the probability density functions of $\underline{F}(m,j+i;M)$ and of $\underline{F}(m,j-i;M)$ are identical, then the mutual information function is symmetric with respect to j: $I(\underline{P}(m,j+i);\underline{F}(m,j+i)) = I(\underline{P}(m,j-i);\underline{F}(m,j-i))$ and criterion 6) can be proved [10].

What is the relation of the minimum mutual information method to some other methods? We defined the delay by:

a) minimize $I(\underline{P}(m,j);\underline{F}(m,j))$

and consider three other methods:

b) minimize $I(\underline{P}(m,j;M);\underline{F}(m,j;M))$ (for finite M)

c) minimize $H(\underline{F}(m,j;M))$ (for finite M)

d) maximize $I(\underline{F}_x(m;M);\underline{F}_y(m+j;M)$ (for finite M)

Method b) is a practical approximation of a) for a finite M. For this method b) it is sufficient to determine for a large enough M the minimum of the first term of (2.6): $H(\underline{F}(m,j;M))$. This leads to method c), which seems to be related to the maximum likelihood approach [6]. The entropies $H(\underline{F}_x(m;M))$ and $H(\underline{F}_y(m+j;M))$ are independent of j and m. So method d) is equivalent to c). This method d) for $M = 1$ is called the mutual information delay estimation method of Mars [1,2]. We conclude that the methods b) - d) are approximations of a) for a finite M and that methods c) and d) are equivalent.

4. Implementation for normal distributions.

The estimation of the mutual information between the past and the future for infinite M, method a), is impossible. For finite M this method b) has some disadventages: its computational complexity and the poor approximation of M·L of (2.6) by $H(\underline{F}(m,j;M)|\underline{P}(m,j;M))$ if j is large. Also because only entropies of the future are involved and for historical reasons we prefer maximizing $I(\underline{F}_x(m;M); \underline{F}_y(m+j;M))$ d) to minimizing $I(\underline{P}(m,j;M);\underline{F}(m,j;M))$ b).

We assume $E(\underline{x}_n) = E(\underline{y}_n) = 0$ and the joint signal vectors are normally distributed:

$$(4.1)\quad f(\underline{F}(m,j;M)) = \frac{1}{(2\pi)^M\sqrt{\det C(j)}} \exp -\frac{1}{2}\,\underline{F}(m,j;M)^T C(j)^{-1}\underline{F}(m,j;M)$$

with C(j) the covariance matrix:

$$(4.2)\qquad E\left\{ \begin{bmatrix} \underline{F}_x(m;M) \\ \underline{F}_y(m+j;M) \end{bmatrix} \cdot \begin{bmatrix} \underline{F}_x(m;M) \\ \underline{F}_y(m+j;M) \end{bmatrix}^T \right\} = \begin{bmatrix} C_{xx} & C_{xy}(j) \\ C_{yx}(j) & C_{yy} \end{bmatrix}$$

According to Shannon [7] we calculate the joint entropie:

$$(4.3)\qquad H\{\underline{F}(m,j;M)\} = M \log 2\pi + \frac{1}{2} \log \det C(j) + M$$

and similar for both marginal entropies. Calculating the mutual information leads to:

$$(4.4)\qquad I\{\underline{F}_x(m;M);\underline{F}_y(m+j;M)\} = -\frac{1}{2} \log \frac{\det C(j)}{\det C_{xx} \cdot \det C_{yy}}$$

Searching for the maximum of $I\{\underline{F}_x(m;M);\underline{F}_y(m+j;M)\}$ is equivalent to searching for the minimum of det C(j). The devision of det C(j) by det C_{xx} and det C_{yy} is a time-shift independent normalization. For $M = 1$ the mutual information function is a monotonic increasing function of $|\rho(j)|$, with $\rho(j)$ the correlation coefficient of the binormal distribution. This proves that searching for the position of the maximum modulus of the correlation coefficient is equivalent to our method for $M = 1$.

According to (4.4) the calculation of the mutual information is reduced to a simple operation on determinants of covariance matrices. We estimate by a non-parametric method the cross- and auto-correlation functions and we use these estimates to determine the covariance matrices. Substitution of these covariance matrices into (4.4) provides us with mutual information estimates. Searching for the position of the maximum as a function of j provides us with an estimate of the delay of y with respect to x.

The actual choice of the segment length is based on visual inspection. We use a large enough M, such that all ambiguities which are visible for small M have disapeared.

5. A theoretical elaboration.

For normal stationary stochastic discrete-time processes we calculate the mutual information function for the methods b) and d). The mutual information fuctions of signal model I are presented in figures 1 and 2. The x signal consists of filtered normal white noise using a second order filter with damping factor R = 0.98 and a frequency $f_0 = 0.05$ periods/sample [8, p. 166]. The y signal is the sum of the x signal and normal white noise with a signal to noise ratio $s_n = 1.0$. Most delay estimation methods are based on the evaluation of the extrema of a criterion as a function of the time-shift j. In the case of narrowband signals many of these methods, for example the generalized correlation method [9] and our three methods b) - d) for small M produce criteria with a large number of ambiguous extrema. Because our criteria have for $M \longrightarrow \infty$ only one extremum, we expect that all ambiguities resolve for increasing M. This effect is shown in figures 1 and 2.

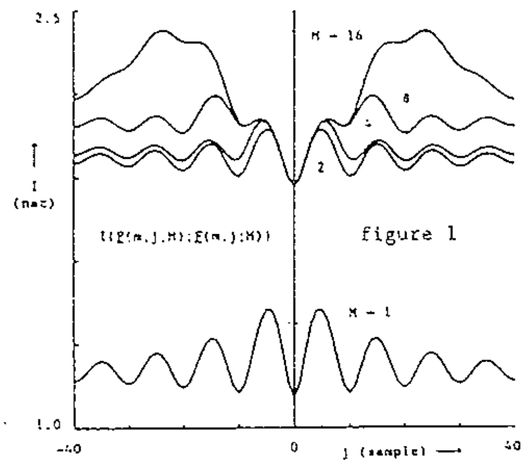
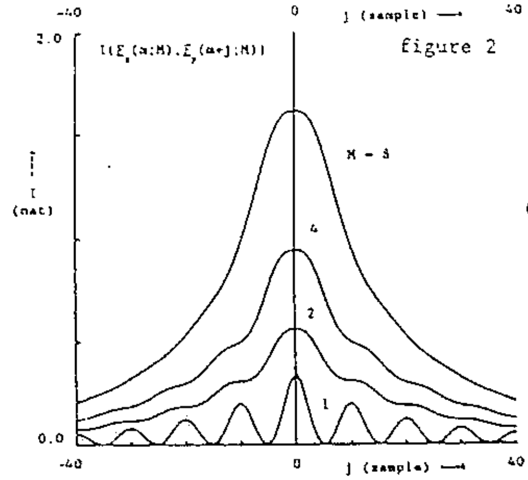
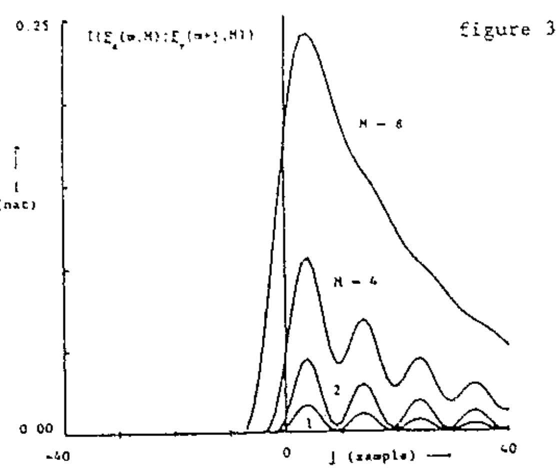


Figure 1. $I[\underline{P}(m,j:M):\underline{E}(m,j:M)]$ for increasing segment length M as a function of the time-shift j for signal model I.

Figure 2. $I(\underline{E}_x(m:M):\underline{E}_y(m+j:M))$ for increasing segment length M as a function of the time-shift j for signal model I.

Figure 3. $I(\underline{E}_x(m:M):\underline{E}_y(m+j:M))$ for increasing segment length M as a function of the time-shift j for signal model II.





If the delaying system is a second order auto-regressive filter we have the situation of model II. The x signal consists of normal white noise. The y signal is the sum of the second order filtered x signal, using the same filter as in model I, and normal white noise with a signal to noise ratio 1.0. The mutual information function for this model is not symmetrical (figure 3). The classical interpretation of a delay looses its meaning, but our information theoretical delay definition maintains its interpretation. In this case apparently y is delayed with respect to x and this delay corresponds approximately with the time of the first maximum of the impuls response of the filter.

6. Conclusions.

We have presented the information theoretical definition of a delay. Our first experiments indicate that for the delay estimation between recordings of EEG-signals it is a powerful method.

The only restrictive assumption is that it is sensible to use mutual information for the definition of a delay. Advantages of this definition are: unambiguity can be proved, the results are interpretable even if the delaying system is not a pure delay, the method is suitable for short data segments because no Fourier transform is involved and the method is model-free. A weakness is that it contains the limit for $M \longrightarrow \infty$. For practical applications this seems not to be a problem. The estimation of the criterion depends on the ability to estimate the mutual information. The estimator of the delay is based on a criterion for a finite M. Only for M = 1 [2] and for normal processes satisfactory solutions have been found. Method d) with M = 1 leads to well established methods: cross-correlation method and Mars' mutual information method [1].

References.

[1] N.J.I. Mars et.al., Signal Processing, 4 (1982) 139-153.
[2] R. Moddemeijer, Int. report University of Twente, nr. 080-87-33, 1987, (to be publ. in Signal Processing).
[3] C.H. Knapp et.al., IEEE Trans. on ASSP, 24 (1976) 320-327.
[4] C.Y. Wuu et.al., IEEE Trans. on ASSP, 32 (1984) 829-835.
[5] R. Moddemeijer, Int. report University of Twente, nr. 080-87-43, 1987.
[6] R. Moddemeijer, Eighth symposium on information theory in the Benelux, Deventer, 21-22 May 1987, The Netherlands.
[7] C.E. Shannon, The Bell System Technical Journal, 27 (1948) 379-423 and 623-656.
[8] G.M. Jenkins and D.G. Watts, Spectral analysis and its applications, San Francisco: Holden-Day, 1968.
[9] J.C. Hassab and R.E. Boucher, IEEE Trans. on ASSP, 29 (1981) 549-555.
[10] R. Moddemeijer, Int. report University of Twente, nr. 080-87-45, 1987, (subm. for publ. in IEEE Trans. on IT).