

## ON ESTIMATION OF ENTROPY AND MUTUAL INFORMATION OF CONTINUOUS DISTRIBUTIONS

R. MODDEMEIJER

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

Received 2 September 1987

Revised 25 May 1988 and 1 June 1988

**Abstract.** Mutual information is used in a procedure to estimate time-delays between recordings of electroencephalogram (EEG) signals originating from epileptic animals and patients. We present a simple and reliable histogram-based method to estimate mutual information. The accuracies of this mutual information estimator and of a similar entropy estimator are discussed. The bias and variance calculations presented can also be applied to discrete valued systems. Finally, we present some simulation results, which are compared with earlier work.

**Zusammenfassung.** Die Auswertung von Elektroenzephalogrammen bei Epilepsie-Kranken erfordert die Bestimmung von Totzeiten zwischen den einzelnen Aufzeichnungen. Hierzu sind Verfahren geeignet, die am informationstheoretischen Begriff "mittlere Transinformation" anknüpfen. In diesem Beitrag wird eine einfache und zuverlässige Methode beschrieben, die mittlere Transinformation auf der Basis experimenteller Daten zu schätzen. Diskutiert werden Erwartungstreue und Varianz der vorgeschlagenen Schätzfunktionen. Die theoretisch erzielten Ergebnisse werden mit experimentellen Daten verglichen, die an Signalfolgen mit Gauß-verteilten Amplituden gewonnen wurden.

**Résumé.** L'information mutuelle est utilisée dans une procédure pour estimer les retards temporels entre les enregistrements de signaux électroencéphalographique (EEG) provenant des animaux et patients épileptiques. Nous présentons une méthode simple et fiable basée sur l'histogramme pour estimer l'information mutuelle. La précision de cet estimateur d'information mutuelle et celle des estimateurs d'entropie sont discutées. Les calculs de biais et de variance présentés peuvent également être appliqués aux systèmes à valeurs discrètes. Finalement, nous présentons quelques résultats de simulations qui sont comparés aux travaux antérieurs.

**Keywords.** Mutual information, entropy, bias, variance, estimator, delay estimation.

### 1. Introduction

To estimate time-delays between recordings of electroencephalogram (EEG) signals we use a method based on maximum mutual information [8]. This procedure to locate epileptic foci has produced promising results [9, 10]. The method can be compared with the cross-correlation method, because both methods search for the maximum correspondence of pairs of samples  $(x(t), y(t - \tau))$  as a function of a time-shift  $\tau$ . In the case of a binormally distributed  $x$  and  $y$ , theoretically both methods are equivalent. These methods do not take the statistical dependence of subsequent sample pairs into account. An example of a method which does, is the maximum likelihood delay estimation method of Knapp et al. [7]. The mutual information method can handle those cases in which  $y$  is a corrupted non-linear response of  $x$ . Recently, this method was used to analyse human electroencephalogram signals [15].

We assume the  $x$ - and the  $y$ -signal are corrupted responses to a common cause and all signals originate from stationary stochastic processes. The probability density function (pdf) of the pair  $(x(t), y(t - \tau))$  is

denoted by  $f_{xy}(x, y; \tau)$  (notational conventions with respect to stochastic variables and estimators are explained in Appendix A). The time-shift maximizing the mutual information  $I\{\underline{x}(t); \underline{y}(t - \tau)\}$  is regarded as the delay of  $x$  with respect to  $y$ . In our present analysis the time-shift  $\tau$  does not play any role, so we suppress it.

The definitions of entropy and mutual information go back to Shannon [14] and these are for continuous distributions

$$H\{\underline{x}\} = H_x = - \int_{-\infty}^{\infty} f_x(x) \log f_x(x) dx, \quad (1.1)$$

$$H\{\underline{x}, \underline{y}\} = H_{xy} = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) \log f_{xy}(x, y) dx dy, \quad (1.2)$$

$$I\{\underline{x}; \underline{y}\} = I_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) \log \frac{f_{xy}(x, y)}{f_x(x)f_y(y)} dx dy. \quad (1.3)$$

We assume base “e” for the logarithm, so the unit of measurement is “nat”. We will use the shorthand notation  $I_{xy}$  instead of  $I\{\underline{x}; \underline{y}\}$  and similarly for the entropy. The mutual information is a function of entropies

$$I_{xy} = H_x + H_y - H_{xy}. \quad (1.4)$$

Because of (1.4) we have to develop a theory to estimate  $H_{xy}$  only: the modifications to estimate  $H_x$  or  $I_{xy}$  are almost trivial, so we will concentrate on  $H_{xy}$  and we only present the results for  $H_x$  and  $I_{xy}$ .

The estimation of mutual information and entropy is a two-step process: first the pdf is estimated and thereafter the mutual information or entropy is calculated. In his work Mars estimated  $f_{xy}(x, y)$  by a kernel method [3, 13] and he calculated from this estimate  $\hat{I}_{xy}$  by numerical integration. The main disadvantages of this method were: its complexity, the inefficient computation (mainly because he determined the optimal kernel-width iteratively), and the lack of understanding of the statistical properties.

We discretize  $x$  and  $y$  and estimate the pdf, represented by a histogram; thereafter we calculate  $\hat{I}_{xy}$  from this estimate. The problem of choosing the optimal kernel-width is replaced by the problem of choosing the optimal rectangular grid dividing the  $xy$ -plane into cells.

## 2. The estimator function

We define a rectangular grid in the  $xy$ -plane by lines parallel to the axes dividing a part of the  $xy$ -space into  $(I \times J)$  equally sized  $(\Delta x \times \Delta y)$  cells with coordinates  $(i, j)$ . The origin and the grid are chosen with the histogram covering the area  $|x - \bar{x}| < 3\sigma_x$  and  $|y - \bar{y}| < 3\sigma_y$ . We define a probability  $p_{ij}$  of observing a sample in cell  $(i, j)$  with

$$p_{ij} = \iint_{\text{cell}(i,j)} f_{xy}(x, y) dx dy \approx f_{xy}(x_i, y_j) \Delta x \Delta y \quad (2.1)$$

and  $(x_i, y_j)$  representing the centre of the cell. The number of samples observed in cell  $(i, j)$  is  $k_{ij}$ : the total number of samples equals  $N$ . We neglect the probability of observing a sample outside the histogram. Row and column sums are denoted by

$$k_{i.} = \sum_{j=1}^J k_{ij}, \quad k_{.j} = \sum_{i=1}^I k_{ij}. \quad (2.2)$$

If the samples are independent, the stochastic variables  $\underline{k}_{ij}$  are multinomially distributed [1] with expectations

$$E\{\underline{k}_{ij}\} = \bar{k}_{ij} = Np_{ij} \tag{2.3}$$

and covariances

$$\begin{aligned} \text{COV}\{\underline{k}_{ij}, \underline{k}_{mn}\} &= Np_{ij}(1 - p_{ij}), \quad \text{if } i = m \text{ and } j = n, \\ &= -Np_{ij}p_{nm}, \quad \text{all others.} \end{aligned} \tag{2.4}$$

The third- and fourth-order central moments of the multinomial distribution are

$$\begin{aligned} E\{(\underline{k}_{ij} - \bar{k}_{ij})^3\} &= N(2p_{ij}^3 - 3p_{ij}^2 + p_{ij}), \\ E\{(\underline{k}_{ij} - \bar{k}_{ij})^4\} &= N^2 3(p_{ij}^4 - 2p_{ij}^3 + p_{ij}^2) + O\{N\}. \end{aligned} \tag{2.5}$$

If the pdf is approximately constant within a cell, a reasonable approximation of the entropy would be

$$\begin{aligned} H_{xy} &\approx -\sum_{i,j} \Delta x \Delta y \cdot f_{xy}(x_i, y_j) \log f_{xy}(x_i, y_j) \\ &\approx -\sum_{i,j} p_{ij}(\log p_{ij} - \log(\Delta x \Delta y)). \end{aligned} \tag{2.6}$$

If we omit the boundaries of a summation over  $i$  or  $j$ , we mean summation from  $i = 1$  to  $I$ , respectively from  $j = 1$  to  $J$ . For the time being we assume that replacing  $p_{ij}$  by  $k_{ij}/N$  results in a proper estimator function of  $H_{xy}$ . Similarly, we define the estimator functions of  $H_x$  and  $I_{xy}$

$$\hat{H}_x = -\sum_i \left( \frac{k_{i\cdot}}{N} \log \frac{k_{i\cdot}}{N} \right) + \log \Delta x, \tag{2.7}$$

$$\hat{H}_{xy} = -\sum_{i,j} \left( \frac{k_{ij}}{N} \log \frac{k_{ij}}{N} \right) + \log(\Delta x \Delta y), \tag{2.8}$$

$$\hat{I}_{xy} = \sum_{i,j} \frac{k_{ij}}{N} \log \frac{k_{ij}N}{k_{i\cdot}k_{\cdot j}}. \tag{2.9}$$

Replacing  $k_{ij}$ ,  $k_{i\cdot}$ , respectively  $k_{\cdot j}$ , by  $k_{ij}$ ,  $k_{i\cdot}$  and  $k_{\cdot j}$  provides us with estimates of entropy and mutual information. We study the bias and the variance of the estimators. If necessary, we improve these estimators by bias correction.

### 3. Bias

The bias is considered as a sum of two components:

- (1)  $R$ -bias, caused by insufficient representation of the pdf by the histogram, and
- (2)  $N$ -bias, due to the finite sample size.

The behaviour of both components is different:  $R$ -bias is constant and a priori determined by the pdf and by the estimation method,  $N$ -bias depends on the sample size ( $N$ -) and tends to zero for  $N \rightarrow \infty$ . After sub-division of the  $R$ -bias into two separate sources (a) and (b) we end up with three causes:

- (a) limited integration area ( $R$ -),
- (b) finite resolution ( $R$ -), and
- (c) non-linear transformation of unbiased local density estimates to contributions to the entropy ( $N$ -).

In reverse order we present an approximation of the bias caused by these sources.

(c) The entropy estimator according to (2.8) is in fact a non-linear function of probability estimators  $\underline{k}_{ij}/N$ . Because in the case of a (non-linear) concave function like  $-a \log a$ :  $E\{-a \log a\} \leq -E\{a\} \log E\{a\}$  if  $a > 0$ , we expect a biased estimator. The stochastic variable: probability estimator  $\underline{k}_{ij}/N$  is affected by a random estimation error. Because  $-\underline{k}_{ij}/N \log \underline{k}_{ij}/N$  is a concave function both positive and negative deviations of  $\underline{k}_{ij}/N$  from its mean  $\bar{k}_{ij}/N$  will cause a less than proportional deviation. Therefore the entropy will on the average be underestimated due to cause (c). We approximate this bias for every cell  $(i, j)$  by taking the first four terms of a Taylor expansion of (2.8) in  $k_{ij} = \bar{k}_{ij}$

$$\hat{H}_{xy} = \sum_{i,j} \left( -\frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} - \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{ij}}{N} \right) (k_{ij} - \bar{k}_{ij}) - \frac{(k_{ij} - \bar{k}_{ij})^2}{2N\bar{k}_{ij}} + \frac{(k_{ij} - \bar{k}_{ij})^3}{6N\bar{k}_{ij}^2} + R_{ij}^4(k_{ij}) \right) + \log(\Delta x \Delta y), \quad (3.1)$$

in which  $R_{ij}^4(k_{ij})$  is the remainder of the Taylor expansion. We replace the formal parameters  $k_{ij}$  by the stochastic variables  $\underline{k}_{ij}$ , we assume independent samples so  $\underline{k}_{ij}$  is multinomially distributed and we take the expectation

$$E\{\hat{H}_{xy}\} = \sum_{i,j} \left( -\frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} - \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{ij}}{N} \right) E\{(k_{ij} - \bar{k}_{ij})\} - \frac{E\{(k_{ij} - \bar{k}_{ij})^2\}}{2N\bar{k}_{ij}} + \frac{E\{(k_{ij} - \bar{k}_{ij})^3\}}{6N\bar{k}_{ij}^2} + E\{R_{ij}^4(k_{ij})\} \right) + \log(\Delta x \Delta y). \quad (3.2)$$

The linear term vanishes and for the variance of  $\underline{k}_{ij}$  we substitute (2.4). If  $N \rightarrow \infty$ , the third-order term is due to (2.5) of the order  $N^{-2}$  and the expectation of the remainder is in the same order, see Appendix B

$$E\{\hat{H}_{xy}\} = \sum_{i,j} \left( -\frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} \right) - \frac{IJ-1}{2N} + O\left\{ \frac{1}{N^2} \right\} + \log(\Delta x \Delta y), \quad (3.3)$$

and similarly for  $\hat{H}_x$  and  $\hat{H}_y$ . The expression  $IJ-1$  represents the number of degrees of freedom of the histogram:  $I \times J$  cells with probabilities  $p_{ij}$  with the condition that the sum of the probabilities equals one. Due to bias (c) the entropy will, on average, be underestimated and the underestimation deteriorates with decreasing number of samples per degree of freedom. In 1955, by a slightly different method, Miller [11] derived a first- and second-order approximation of the bias caused by (c). Our approximations confirm Miller's; for a discussion see Section 5. If  $\bar{k}_{ij} \rightarrow 0$ , because of the slowly converging Taylor expansion (3.1), the approximation (3.3) loses its usefulness; this case will occur if there are too many cells and some of them are almost empty.

(b) Also without the statistical effect in (c), the finite resolution leads to a bias of  $\hat{H}_{xy}$  because  $f_{xy}(x, y)$  is not constant within a cell. We study this bias locally (for one cell). Assume for the moment the pdf is approximately linear within the cell (see Fig. 1). Because of the concaveness of  $-f_{xy}(x, y) \log f_{xy}(x, y)$ , the integral over one cell of this function will be smaller than the approximation  $-f_{xy}(x_i, y_j) \cdot \log(f_{xy}(x_i, y_j)) \Delta x \Delta y$ , which causes bias (b). To determine this bias we determine the difference between the entropy contribution  $h_{ij}$  of cell  $(i, j)$  calculated by

$$h_{ij} = - \iint_{\text{cell}(i,j)} f_{xy}(x, y) \log f_{xy}(x, y) dx dy, \quad (3.4)$$

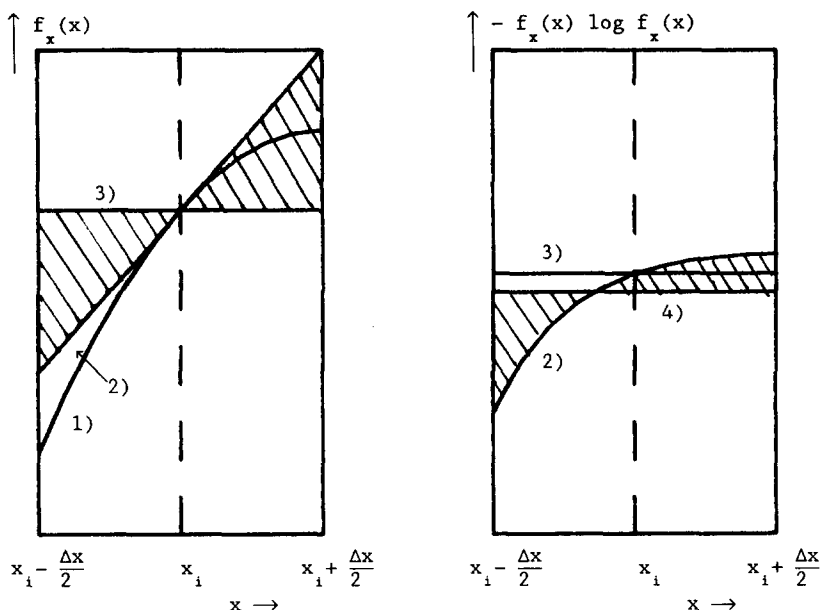


Fig. 1. Bias (b) caused by finite resolution for a one-dimensional pdf. On the left-hand figure curve (1) is the true pdf, (2) its linear approximation in  $x = x_i$ , and (3) the average within a cell of curve (2). On the right-hand figure the entropy contributions of the corresponding left-hand curves are shown. Curve (4) is the average within a cell of the right-hand curve (2).

and its approximation according to (2.6)

$$\tilde{h}_{ij} = -p_{ij}(\log p_{ij} - \log(\Delta x \Delta y)). \tag{3.5}$$

First we express (3.5) as a function of  $f_{xy}(x, y)$  and its derivatives in  $(x_i, y_j)$  and thereafter we do the same for (3.4). The difference between  $\tilde{h}_{ij}$  and  $h_{ij}$  leads, for each cell separately, to an expression for the bias (b). We calculate the probability  $p_{ij}$  by approximating the integrand of (2.1) by a second-order Taylor expansion in  $(x_i, y_j)$

$$p_{ij} \approx (f_{xy}(x_i, y_j) + \Delta f_{xy}(x_i, y_j)) \Delta x \Delta y, \tag{3.6a}$$

with

$$\begin{aligned} \Delta f_{xy}(x_i, y_j) \Delta x \Delta y = & \iint_{\text{cell}(i,j)} \left( \frac{\partial}{\partial x_i} f_{xy}(x_i, y_j)(x - x_i) + \frac{\partial}{\partial y_j} f_{xy}(x_i, y_j)(y - y_j) + \frac{1}{2} \frac{\partial^2}{\partial x_i^2} f_{xy}(x_i, y_j)(x - x_i)^2 \right. \\ & \left. + \frac{\partial^2}{\partial x_i \partial y_j} f_{xy}(x_i, y_j)(x - x_i)(y - y_j) + \frac{1}{2} \frac{\partial^2}{\partial y_j^2} f_{xy}(x_i, y_j)(y - y_j)^2 \right) dx dy. \end{aligned} \tag{3.6b}$$

Note that integral (3.6b) only depends on the third and the fifth term of the integrand because terms linear in  $x$  or  $y$  or  $x$  and  $y$  vanish. The result is small and will be shown to be irrelevant in our calculations. Now we are able to express (3.5) as a function of  $f_{xy}(x_i, y_j)$  and its derivatives:

$$\begin{aligned} \tilde{h}_{ij} & \approx -(f_{xy}(x_i, y_j) + \Delta f_{xy}(x_i, y_j)) \Delta x \Delta y \log(f_{xy}(x_i, y_j) + \Delta f_{xy}(x_i, y_j)) \\ & \approx -(f_{xy}(x_i, y_j) \log f_{xy}(x_i, y_j) + (1 + \log f_{xy}(x_i, y_j)) \Delta f_{xy}(x_i, y_j)) \Delta x \Delta y. \end{aligned} \tag{3.7}$$

The approximation of  $\tilde{h}_{ij}$  by an expression linear in  $\Delta f_{xy}(x_i, y_j)$  is consistent with the second-order approximations we usually make, because the term quadratic in  $\Delta f_{xy}(x_i, y_j)$ , which is omitted, is proportional to  $\Delta x^n \Delta y^m$  with  $n + m = 4$ .

We approximate the integrand of (3.4), as we did to calculate  $p_{ij}$ , by a second-order Taylor expansion in  $(x_i, y_j)$

$$\begin{aligned} h_{ij} \approx & \iint_{\text{cell}(i,j)} \left( -f_{xy}(x_i, y_j) \log f_{xy}(x_i, y_j) - \frac{1}{2f_{xy}(x_i, y_j)} \right. \\ & \times \left. \left( \frac{\partial}{\partial x_i} f_{xy}(x_i, y_j)(x - x_i) + \frac{\partial}{\partial y_j} f_{xy}(x_i, y_j)(y - y_j) \right)^2 \right) dx dy \\ & - (1 + \log f_{xy}(x_i, y_j)) \Delta f_{xy}(x_i, y_j) \Delta x \Delta y. \end{aligned} \quad (3.8)$$

We substitute (3.7) into (3.8) and integrate. The term dependent on  $\Delta f_{xy}(x_i, y_j)$  vanishes

$$\tilde{h}_{ij} \approx h_{ij} + \frac{1}{24f_{xy}(x_i, y_j)} \left( \left( \frac{\partial f_{xy}(x_i, y_j)}{\partial x_i} \right)^2 (\Delta x)^2 + \left( \frac{\partial f_{xy}(x_i, y_j)}{\partial y_j} \right)^2 (\Delta y)^2 \right) \Delta x \Delta y. \quad (3.9)$$

Substitution of (3.9) into (3.3) and approximation of the summation for all  $(x_i, y_j)$  by an integral over  $x$  and  $y$  leads to an approximation of  $E\{\hat{H}_{xy}\}$ :

$$E\{\hat{H}_{xy}\} \approx H_{xy} - \frac{IJ-1}{2N} + \iint_{-\infty}^{\infty} \frac{1}{24f_{xy}(x, y)} \left( \left( \frac{\partial f_{xy}(x, y)}{\partial x} \right)^2 (\Delta x)^2 + \left( \frac{\partial f_{xy}(x, y)}{\partial y} \right)^2 (\Delta y)^2 \right) dx dy. \quad (3.10)$$

A similar expression can be derived for  $E\{\hat{H}_x\}$

$$E\{\hat{H}_x\} \approx H_x - \frac{I-1}{2N} + \int_{-\infty}^{\infty} \frac{1}{24f_x(x)} \left( \frac{\partial f_x(x)}{\partial x} \right)^2 (\Delta x)^2 dx. \quad (3.11)$$

The bias caused by (b) depends on the cell sizes and deteriorates with increasing cell sizes. The integral expressions in (3.10) and (3.11) measure the smoothness of the pdfs. If these are smooth, the first derivatives are almost zero and the squared first derivatives hardly contribute to the result, so the bias reaches a minimum. Substitution of a normal distribution with arbitrary mean, variances  $\sigma_x$  and  $\sigma_y$  and correlation coefficient  $\rho$  into (3.10) and (3.11) leads to

$$E\{\hat{H}_x\} \approx H_x - \frac{I-1}{2N} + \frac{1}{24} \left( \frac{\Delta x}{\sigma_x} \right)^2, \quad (3.12)$$

$$E\{\hat{H}_{xy}\} \approx H_{xy} - \frac{IJ-1}{2N} + \frac{1}{24(1-\rho^2)} \left( \left( \frac{\Delta x}{\sigma_x} \right)^2 + \left( \frac{\Delta y}{\sigma_y} \right)^2 \right), \quad (3.13)$$

$$E\{\hat{I}_{xy}\} \approx I_{xy} + \frac{(I-1)(J-1)}{2N} - \frac{\rho^2}{24(1-\rho^2)} \left( \left( \frac{\Delta x}{\sigma_x} \right)^2 + \left( \frac{\Delta y}{\sigma_y} \right)^2 \right). \quad (3.14)$$

Note, in case of normal distributions, the absolute values of both the  $R$ -bias and the  $N$ -bias are smaller in  $\hat{I}_{xy}$  than in  $\hat{H}_{xy}$ .

(a) Although the integration variables of (1.2) run from  $-\infty$  to  $\infty$ , the histogram only covers a finite area. In case of a binormal distribution and a histogram with  $|x - \bar{x}| < 3\sigma_x$  and  $|y - \bar{y}| < 3\sigma_y$ , this leads to a bias of  $\hat{I}_{xy}$  in absolute value smaller than  $0.011I_{xy} + 0.019|\rho|$  [12]. This error is small compared to the bias caused by (b) and (c).

**4. Variance**

To calculate the variance we use a method similar to the calculation of bias (c). We approximate the entropy contribution of every cell  $(i, j)$  by taking the first two terms of a Taylor expansion of (2.8) in  $k_{ij} = \bar{k}_{ij}$ , instead of the four terms of (3.1), and replace  $k_{ij}$  by  $\bar{k}_{ij}$

$$\hat{H}_{xy} = \sum_{i,j} \left( -\frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} - \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{ij}}{N} \right) (k_{ij} - \bar{k}_{ij}) + R_{ij}^2(k_{ij}) \right) + \log(\Delta x \Delta y). \tag{4.1}$$

The variance of an entropy estimator equals

$$\text{VAR}\{\hat{H}_{xy}\} = E\{\hat{H}_{xy}^2\} - E\{\hat{H}_{xy}\}^2. \tag{4.2}$$

We substitute the Taylor expansion (4.1) into (4.2). Terms containing the first and last term on the right-hand side of (4.1) vanish, because the variance is independent of additional constants. All terms containing  $R_{ij}^2(k_{ij})$  lead to an approximation error of the order  $N^{-2}$

$$\begin{aligned} \text{VAR}\{\hat{H}_{xy}\} = & \sum_{i,j} \sum_{m,n} \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{ij}}{N} \right) \cdot \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{mn}}{N} \right) \cdot E\{(k_{ij} - \bar{k}_{ij})(k_{mn} - \bar{k}_{mn})\} \\ & - \left( \sum_{i,j} \left( \frac{1}{N} + \frac{1}{N} \log \frac{\bar{k}_{ij}}{N} \right) E\{(k_{ij} - \bar{k}_{ij})\} \right)^2 + O\left\{ \frac{1}{N^2} \right\}. \end{aligned} \tag{4.3}$$

Substitution of (2.3) and (2.4) into (4.3) results in a variance approximation of the entropy  $\hat{H}_{xy}$ . Similarly, we approximate the variance of  $\hat{H}_x$  and  $\hat{I}_{xy}$

$$\text{VAR}\{\hat{H}_x\} = \frac{1}{N} \left( \sum_i \frac{\bar{k}_i}{N} \log^2 \frac{\bar{k}_i}{N} - \left( \sum_i \frac{\bar{k}_i}{N} \log \frac{\bar{k}_i}{N} \right)^2 \right) + O\left\{ \frac{1}{N^2} \right\}, \tag{4.4}$$

$$\text{VAR}\{\hat{H}_{xy}\} = \frac{1}{N} \left( \sum_{i,j} \frac{\bar{k}_{ij}}{N} \log^2 \frac{\bar{k}_{ij}}{N} - \left( \sum_{i,j} \frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} \right)^2 \right) + O\left\{ \frac{1}{N^2} \right\}, \tag{4.5}$$

$$\text{VAR}\{\hat{I}_{xy}\} = \frac{1}{N} \left( \sum_{i,j} \frac{\bar{k}_{ij}}{N} \log^2 \frac{\bar{k}_{ij}N}{\bar{k}_i \bar{k}_j} - \left( \sum_{i,j} \frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}N}{\bar{k}_i \bar{k}_j} \right)^2 \right) + O\left\{ \frac{1}{N^2} \right\}. \tag{4.6}$$

These results lead to the following observations:

(1) Replacing the expected number of samples  $\bar{k}_{ij}$  by the observed number of samples  $k_{ij}$  results in a variance estimator independent of the distribution.

(2) Approximating the summation by an integration leads to

$$\begin{aligned} \text{VAR}\{\hat{H}_{xy}\} \approx & \frac{1}{N} \left( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) \log^2 f_{xy}(x, y) \, dx \, dy \right. \\ & \left. - \left( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y) \log f_{xy}(x, y) \, dx \, dy \right)^2 \right). \end{aligned} \tag{4.7}$$

(3) We can use a short notation for the entropy according to (1.2)

$$H_{xy} = E\{-\log f_{xy}(\underline{x}, \underline{y})\}. \quad (4.8)$$

Similarly, we can rewrite the variance approximations (4.4)–(4.7) of the entropy and the mutual information:

$$\text{VAR}\{\hat{H}_x\} \approx \frac{1}{N} \text{VAR}\{-\log f_x(\underline{x})\}, \quad (4.9)$$

$$\text{VAR}\{\hat{H}_{xy}\} \approx \frac{1}{N} \text{VAR}\{-\log f_{xy}(\underline{x}, \underline{y})\}, \quad (4.10)$$

$$\text{VAR}\{\hat{I}_{xy}\} \approx \frac{1}{N} \text{VAR}\left\{\log \frac{f_{xy}(\underline{x}, \underline{y})}{f_x(\underline{x})f_y(\underline{y})}\right\}. \quad (4.11)$$

(4) Remarkably, the variance is approximately independent of the cell sizes. Only in extreme cases like  $I = J = 1$  and  $I = J \rightarrow \infty$  will the variance both theoretically and in practice be zero.

(5) In the case of binormally distributed random variables the variances of entropy and mutual information are independent of the mean and the variance

$$\text{VAR}\{\hat{H}_x\} \approx \frac{1}{2}N^{-1}, \quad (4.12)$$

$$\text{VAR}\{\hat{H}_{xy}\} \approx N^{-1}, \quad (4.13)$$

$$\text{VAR}\{\hat{I}_{xy}\} \approx \rho^2 N^{-1}. \quad (4.14)$$

(6) In theory, the first-order variance approximation will be zero in the case of uniform distributions with the edges coinciding with the grid, therefore a uniform distribution is an excellent choice to study second-order effects. A more thorough investigation of (4.4)–(4.6) leads to the astonishing conclusion that the variance of such uniform distributions must be of the order  $N^{-2}$  instead of  $N^{-1}$ .

(7) We compare for a normal distribution  $\text{VAR}\{\hat{I}_{xy}\}$  with  $\text{VAR}\{\hat{\rho}\}$ . The variance of the maximum likelihood estimator of the correlation coefficients  $\rho$  for large  $N$  equals [6]

$$\text{VAR}\{\hat{\rho}\} = (1 - \rho^2)^2 / N. \quad (4.15)$$

Using the relation between  $\rho$  and  $I_{xy}$  in the case of a binormal distribution

$$I_{xy} = -\frac{1}{2} \log(1 - \rho^2), \quad (4.16)$$

we can demonstrate by error propagation, for large  $N$ , that the variances of both estimators  $\hat{\rho}$  and  $\hat{I}_{xy}$  are equivalent; in this respect there is no preference to determine  $I_{xy}$  via  $\rho$  or directly.

Determination of the optimal cell sizes is difficult because a priori knowledge of the distributions is needed. Because the number of cells, the sizes of these cells, and the area of the  $xy$ -plane covered by the histogram are related by  $6\sigma_x = I \Delta x$  and  $6\sigma_y = J \Delta y$ , a  $\rho$ -dependent grid can be found with the  $R$ -bias (b) and the  $N$ -bias (c) compensating each other (3.14). This grid is the optimal grid with a minimum mean square error, because the variances are almost independent of the cell sizes. For mutual information estimation in the case of a binormal distribution this optimal grid as a function of  $\rho$  is given in Fig. 2. Of course, for  $\rho = 0$  the optimal grid will have one cell, because then under all circumstances  $x$  and  $y$  in the histogram are guaranteed to be statistically independent. If the pdf becomes peaked, which is the case if  $\rho \rightarrow 1$ , the optimal number of cells increases reducing the smoothing caused by finite resolution.



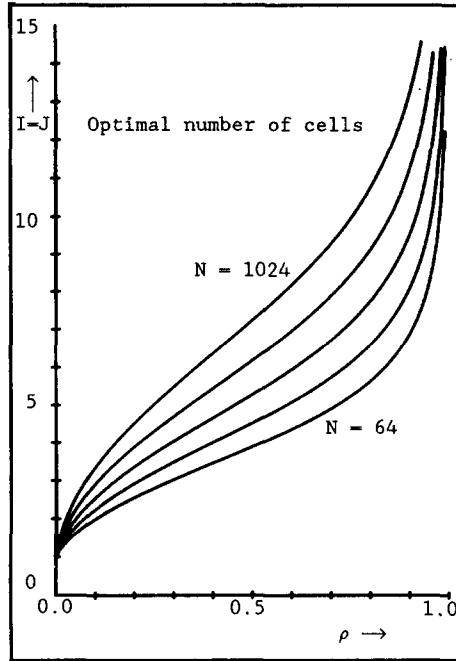


Fig. 2. The optimal number of cells  $I=J$  to estimate the mutual information in the case of a normal distribution as function of  $\rho$  for  $N=64, 128, 256, 512$  and  $1024$ .

**5. Exact  $N$ -bias calculation**

In Section 3 we approximated the  $N$ -bias of the entropy estimator, contenting ourselves with four terms of the Taylor expansion (3.1). However, central moments of the multinomially distributed  $k_{ij}$ 's are functions of the  $N$  and the  $p_{ij}$ 's, which can be calculated using the moment-generating function [4]. We can, for the  $N$ -bias and also for the variance, derive higher order approximations. If convergence is assured, this strategy seems to be attractive to calculate the exact  $N$ -bias. First we point out that this strategy is condemned to fail and can only be used for low order approximations. Then we determine the exact  $N$ -bias by a straightforward calculation of the expectation of the entropy estimator using the multinomial distribution.

Instead of using the first four terms of the Taylor expansion of (2.8), as in (3.1) we calculate an improved  $N$ -bias approximation by taking any odd number  $2L+1$  ( $L > 1$ ) of terms into account. Note that if  $N \rightarrow \infty$  both the  $2L$ th and  $2L+1$ st term contribute to an additional term of  $O\{N^{-L}\}$ . The  $2L+2$ nd term is in the same order as the remainder. As in Appendix B, the remainder  $R_{ij}^{2L+2}$  is bounded as follows

$$-\frac{1}{(2L+1)N} E\left\{\frac{(k_{ij} - \bar{k}_{ij})^{2L+2}}{\bar{k}_{ij}^{2L+1}}\right\} \leq E\{R_{ij}^{2L+2}(k_{ij})\} \leq 0. \tag{5.1}$$

For large  $N$ ,  $k_{ij}$  follows approximately a normal distribution with mean  $\bar{k}_{ij} = Np_{ij}$  and variance  $\sigma_{ij}^2 = Np_{ij}(1-p_{ij})$ . The even central moments of that normal distribution are given by [2]

$$E\{(k_{ij} - \bar{k}_{ij})^{2L+2}\} = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2L+1) \sigma_{ij}^{2L+2}. \tag{5.2}$$

If the remainder converges to zero, convergence is assured for increasing  $L$ ; this is not the case. For large  $L$  the central moments of the normal distribution increase rapidly, so there exists only a  $p_{ij}$ -dependent

maximum order  $M(p_{ij})$  with the expectation of the remainder converging as long as  $L \ll M(p_{ij})$ . Therefore only low order approximations are useful.

If we take  $L = 2$  we find a second-order  $N$ -bias approximation after taking the expectation of the first five terms of the Taylor expansion of (2.8) and substituting (2.4)-(2.5)

$$\begin{aligned} N\text{-BIAS}\{\hat{H}_{xy}\} &= \sum_{i,j} \left( -\frac{1-p_{ij}}{2N} + \frac{1}{12N^2} \left( p_{ij} - \frac{1}{p_{ij}} \right) \right) + O\left\{ \frac{1}{N^3} \right\} \\ &= -\frac{IJ-1}{2N} + \frac{1}{12N^2} \left( 1 - \sum_{i,j} \frac{1}{p_{ij}} \right) + O\left\{ \frac{1}{N^3} \right\}. \end{aligned} \tag{5.3}$$

Both the first- and the second-order  $N$ -bias approximations are in agreement with the earlier work of Miller [11]. It is likely Miller was not aware of the multinomial distribution of all  $k_{ij}$ , otherwise he would not have used the argument: “the entropy estimator is approximately chi-squared distributed”, to calculate the  $N$ -bias.

The second strategy to determine the exact  $N$ -bias is to calculate the expectation of the entropy estimator using the multinomial distribution. We calculate

$$E\{\hat{H}_{xy}\} = \sum_{i,j} E\{\hat{h}_{ij}\}, \quad \text{with } \hat{h}_{ij} = -\frac{k_{ij}}{N} \log \frac{k_{ij}}{N \Delta x \Delta y}. \tag{5.4}$$

For one cell the multinomial distribution reduces to a binomial distribution: a sample is observed inside or outside a cell. We calculate the  $N$ -bias of the entropy estimator

$$N\text{-BIAS}\{\hat{h}_{ij}\} = \frac{\bar{k}_{ij}}{N} \log \frac{\bar{k}_{ij}}{N} - \sum_{k=0}^N P(k|p_{ij}) \left( \frac{k}{N} \log \frac{k}{N} \right). \tag{5.5}$$

Substitution of the binomial distribution leads to the exact  $N$ -bias for cell  $(i, j)$

$$N\text{-BIAS}\{\hat{h}_{ij}\} = p_{ij} \log p_{ij} - \sum_{k=0}^N \binom{N}{k} p_{ij}^k (1-p_{ij})^{N-k} \left( \frac{k}{N} \log \frac{k}{N} \right). \tag{5.6}$$

In Table 1 we present the contribution to the  $N$ -bias for every cell. The  $N$ -bias contribution is always negative and reaches, in absolute value, a maximum in the neighbourhood of  $p_{ij} = 1/N$ . For  $N = 256$  we also present the first- and second-order  $N$ -bias approximation. If  $p_{ij}$  is large, a first-order approximation is sufficient, but in the neighbourhood of the extremum a second-order approximation is desirable.

**6. Simulations**

To verify our theory we generate 100 sequences of  $N = 256$  binormally distributed samples. We estimated  $\hat{I}_{xy}$  using different cell sizes. The averaged results over the sequences are presented in Fig. 3. Ideally the estimator  $\hat{I}_{xy}$  as a function of the true value  $I_{xy}$  is a straight line. Characteristically,  $I_{xy}$  is overestimated due to cause (c); this overestimation increases with the number of cells. However, if  $I_{xy}$  is large,  $I_{xy}$  is underestimated, due to the domination of cause (b). This underestimation decreases with the number of cells, or in other words with an increasing resolution. After full bias correction (3.14) we obtain the improved graphs of Fig. 4. The large deviations for  $I = J = 4$  and  $I_{xy}$  large are probably caused by the approximation of the summation by an integration in the calculation of the  $R$ -bias (b); it cannot be

Table 1

$N$ -bias contribution of every cell  $(i, j)$  as a function of  $p_{ij}$  and  $N$ . Presented is  $-N\text{-BIAS}\{\hat{I}_{ij}\}$  in  $10^{-3}$  nats. The left-hand columns are exact  $N$ -bias calculations according to (5.6), and the right-hand columns are the first- and second-order approximations for  $N = 256$  based on (5.3)

$p_{ij}$	Exact $N$ -bias				First approximation	Second approximation
	$N = 64$	128	256	512	256	256
0.0001	0.51	0.44	0.37	0.30	1.95	14.67
0.0002	0.87	0.74	0.60	0.47	1.95	8.31
0.0005	1.73	1.40	1.07	0.77	1.95	4.50
0.001	2.79	2.14	1.28	0.99	1.95	3.22
0.002	4.28	3.06	1.98	1.12	1.95	2.58
0.005	6.72	4.18	2.26	1.08	1.94	2.20
0.01	8.33	4.50	2.15	1.01	1.93	2.06
0.02	8.93	4.27	2.00	0.97	1.91	1.98
0.05	8.09	3.84	1.88	0.93	1.86	1.88
0.1	7.28	3.57	1.77	0.88	1.76	1.77
0.2	6.36	3.15	1.57	0.78	1.56	1.57
0.5	3.94	1.96	0.98	0.49	0.98	0.98

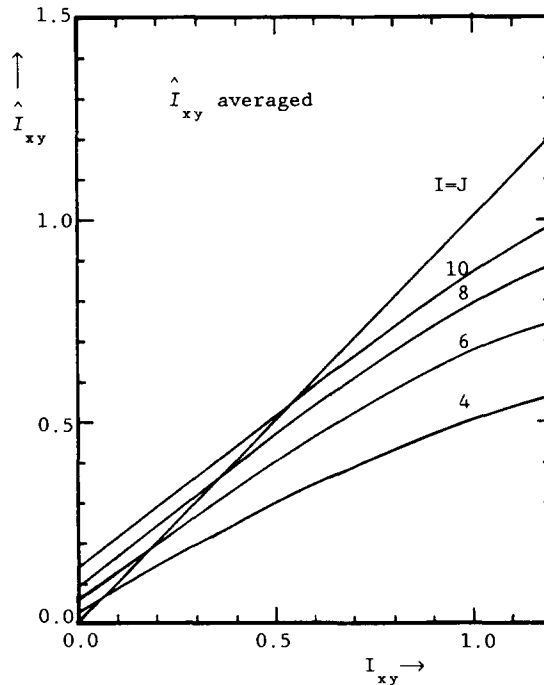


Fig. 3. The averaged (over 100 sequences) mutual information estimates  $\hat{I}_{xy}$  (in nats) for  $N = 256$  binormally distributed samples as a function of  $I_{xy}$ . For an ideal estimator these curves coincide with the line  $\hat{I}_{xy} = I_{xy}$ .

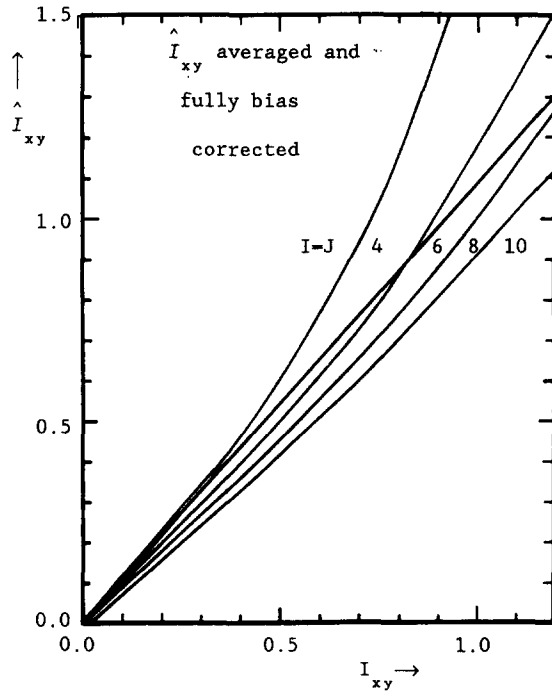


Fig. 4. As in Fig. 3: the mutual information estimates corrected for  $R$ -bias (b) and for the part  $(I-1)(J-1)/2N$  of  $N$ -bias (c).

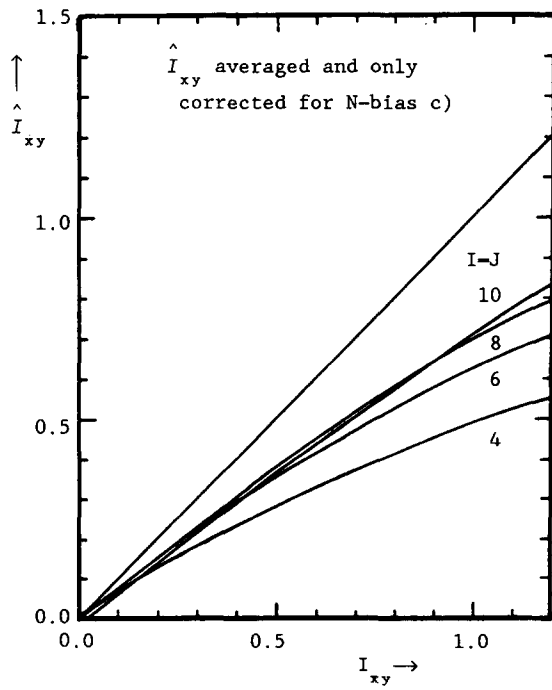


Fig. 5. As in Fig. 3: the mutual information estimates only corrected for the part  $(I-1)(J-1)/2N$  of  $N$ -bias (c).

justified to replace a summation over 4 cells of size  $\frac{3}{2}\sigma$  by an integration. If we only correct the distribution independent part of  $N$ -bias (c), we obtain the graphs of Fig. 5.

According to the graphs of Figs. 3-5  $E\{\hat{I}_{xy}\}$  is a monotonously increasing function of  $I_{xy}$ . This shows that the maximum of  $\hat{I}_{xy}$ , as a function of  $\tau$ , does not depend on the bias. This is a strong argument not to change the grid for estimations involving different  $\tau$ . We expect a non-optimal grid will not change the delay estimate.

The estimated standard deviation obtained from 100 sequences is given in Table 2 for different values of  $N$ . The average standard deviation calculated using our variance estimator (4.6) is presented in Table 3. We see a good agreement between these tables and the variance approximation of (4.14) (Table 4), except for Table 4 and  $\rho = 0$ ; which results cannot be realistic.

Table 2  
Standard deviation of  $\hat{I}_{xy}$  (in nats) experimentally obtained from 100 sequences

N	I = J	$\rho$	0.00	0.30	0.45	0.70	0.85	0.95
		$I_{xy}$	0.00	0.05	0.11	0.34	0.64	1.16
128	6		0.020	0.037	0.035	0.054	0.061	0.075
	8		0.037	0.035	0.042	0.062	0.065	0.084
256	6		0.014	0.018	0.024	0.037	0.047	0.060
	8		0.019	0.021	0.027	0.042	0.049	0.063
512	6		0.007	0.013	0.019	0.027	0.031	0.041
	8		0.009	0.013	0.021	0.029	0.032	0.042

Table 3  
Estimated standard deviation of  $\hat{I}_{xy}$ , estimated using (4.6), averaged

N	I = J	$\rho$	0.00	0.30	0.45	0.70	0.85	0.95
128	6		0.030	0.036	0.042	0.056	0.065	0.079
	8		0.039	0.043	0.048	0.060	0.068	0.077
256	6		0.016	0.023	0.027	0.039	0.046	0.056
	8		0.022	0.026	0.031	0.041	0.048	0.055
512	6		0.009	0.014	0.018	0.027	0.033	0.040
	8		0.012	0.016	0.020	0.029	0.034	0.038

Table 4  
Approximate standard deviation of  $\hat{I}_{xy}$  according to (4.14)

N	$\rho$	0.00	0.30	0.45	0.70	0.85	0.95
128		0.000	0.027	0.035	0.062	0.075	0.084
256		0.000	0.019	0.028	0.044	0.053	0.059
512		0.000	0.013	0.020	0.031	0.038	0.042

## 7. Discussion

Our bias and variance calculations agree with the simulation results. Further improvements can be achieved by more accurate calculations. We can, for example, take an exact sum instead of an approximation by an integral. We doubt whether such extensions to improve the estimate justify the effort. For bias (c) an exact expression is derived by a straightforward calculation of the expectation of the entropy estimator.

Including the dependence of subsequent sample pairs by trying to estimate the mutual information of a pair of samples  $(x(t), x(t + \Delta t))$  of the  $x$ - and a pair of samples  $(y(t - \tau), y(t + \Delta t - \tau))$  of the  $y$ -signal leads to a histogram of  $I^2 \times J^2$  cells. For an EEG the number of cells obtained in this way exceeds the number of samples for which the EEGs can be considered to be stationary. The only solution of this problem is the reduction of the number of degrees of freedom by additional assumptions, such as the signals are normally distributed.

Comparing our histogram results with the kernel results of Mars et al. [8], we conclude his mutual information estimates are affected by similar bias. This shows his iterative method to find the optimal kernel-width hardly improves the estimate. This conclusion is not surprising, because in the case of a reasonable number of cells, the entropy or mutual information estimates are hardly sensitive to small changes in the cell sizes. After further tests, also with real EEG data, we concluded both methods—Mars' and ours—lead to equivalent delay estimates. Our variance estimator enables us to judge the significance of a maximum in  $\hat{I}_{xy}$ . The derivation of a covariance estimator of  $\hat{I}_{xy}$ 's belonging to different  $\tau$ 's is a problem, because this estimator depends on the dependence between subsequent data samples of our signals. Because of this dependence a priori knowledge about the correlation function of the mutual information is needed.

Our calculations should be tested using different pdfs in order to obtain a better understanding of the validity of our estimation procedure and of our corrections; such work is reported by Henning et al. [5]. To reduce the  $N$ -bias (c) and the variance we consider equalizing the expected number of samples per cell by choosing a non-equidistant grid.

Our methods can be applied to entropy and mutual information estimation of discrete systems, then only  $N$ -bias (c) and the variance are relevant.

Presumably, the approach of splitting the bias into two components of different origin:  $N$ -bias of statistical origin and  $R$ -bias due to insufficient representation can be used in other fields.

## Acknowledgements

Its a pleasure to thank Prof. Ir. E.W. Gröneveld and Dr. M.R. Best for the fruitful discussions leading to this work.

## Appendix A. Notational convention

For a stochastic variable  $\underline{x}$  we use the notational convention:  $x$  for a formal parameter,  $\underline{x}$  for the stochastic variable,  $\bar{x}$  for the mean and  $x$  for an observation of that variable. We define the estimate of a quantity  $A$  by an estimation operation defined by the function  $\hat{A}(x_1, \dots, x_N)$  of formal parameters  $x_1, \dots, x_N$  representing  $N$  observations. The stochastic *estimator* is defined by  $\hat{A} = \hat{A}(\underline{x}_1, \dots, \underline{x}_N)$  and the observation *estimate* by  $\hat{A} = \hat{A}(x_1, \dots, x_N)$ .

## Appendix B. The remainder of the Taylor expansion

In this appendix we determine the order of the expectation of the remainder of the Taylor expansion (3.1) if  $N \rightarrow \infty$ . We bound this remainder by a minorant function and a majorant function:  $R_{ij}^{4-}(k_{ij}) \leq R_{ij}^4(k_{ij}) \leq R_{ij}^{4+}(k_{ij})$  with  $0 \leq k_{ij} \leq N$ . Thereafter we take the expectation of both the minorant and the majorant and determine their order. The order of  $E\{R_{ij}^4(k_{ij})\}$  must be smaller than the largest order so obtained.

To determine the majorant we consider the remainder of the Taylor expansion

$$R_{ij}^4(k_{ij}) = -\frac{(k_{ij} - \bar{k}_{ij})^4}{12N(k_{ij} - \theta(k_{ij} - \bar{k}_{ij}))^3}, \quad (\text{B.1})$$

with  $0 \leq \theta \leq 1$ . This remainder is negative for all  $\theta$ , so we can take

$$R_{ij}^{4+}(k_{ij}) = 0. \quad (\text{B.2})$$

To determine a minorant we distinguish two situations:  $\bar{k}_{ij} \leq k_{ij} \leq N$  and  $0 \leq k_{ij} < \bar{k}_{ij}$ . For the first interval we can determine a minorant by taking the worst case, occurring if the denominator reaches a minimum, i.e. if  $\theta = 1$ . We can use  $R_{ij}^4(k_{ij})$  with  $\theta = 1$  as minorant on the first interval. To determine the minorant for the second interval we use the remaining terms of the *full* Taylor expansion

$$R_{ij}^4(k_{ij}) = \sum_{m=4}^{\infty} \frac{(-1)^{m-1}}{Nm(m-1)} \frac{(k_{ij} - \bar{k}_{ij})^m}{\bar{k}_{ij}^{m-1}}, \quad (\text{B.3})$$

which converges for any  $k_{ij}$  on this interval. As the minorant of this remainder we use

$$R_{ij}^{4-}(k_{ij}) = -\sum_{m=4}^{\infty} \frac{1}{Nm(m-1)} \frac{(k_{ij} - \bar{k}_{ij})^4}{\bar{k}_{ij}^3}, \quad (\text{B.4})$$

because on this interval  $|(k_{ij} - \bar{k}_{ij})/\bar{k}_{ij}| \leq 1$ , so we can replace every term of (B.3) by a fourth-order minorant of that term. The sum of these minorants equals

$$R_{ij}^{4-}(k_{ij}) = -\frac{1}{3N} \frac{(k_{ij} - \bar{k}_{ij})^4}{\bar{k}_{ij}^3}, \quad (\text{B.5})$$

which is smaller than the minorant proposed for  $\bar{k}_{ij} \leq k_{ij} \leq N$ , so it can be used for the whole  $k_{ij}$  range  $0 \leq k_{ij} \leq N$ .

We determine the order of  $E\{R_{ij}^4(k_{ij})\}$  if  $N \rightarrow \infty$ , with  $E\{R_{ij}^{4-}(k_{ij})\} \leq E\{R_{ij}^4(k_{ij})\} \leq E\{R_{ij}^{4+}(k_{ij})\}$  by calculating the order of  $E\{R_{ij}^{4-}(k_{ij})\}$ . We replace  $k_{ij}$  by  $\underline{k}_{ij}$  in (B.5), take the expectation and substitute (2.5); we find the order

$$E\{R_{ij}^{4-}(\underline{k}_{ij})\} = O\{N^{-2}\}. \quad (\text{B.6})$$

Because both the majorant and its expectation are zero it follows for the remainder of the Taylor expansion

$$O\{N^{-2}\} \leq E\{R_{ij}^4(k_{ij})\} \leq 0. \quad (\text{B.7})$$

## References

- [1] S. Brandt, *Statistical and Computational Methods in Data Analysis*, North-Holland, Amsterdam, 1976, 2nd edn., pp. 39-42.
- [2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1963 (10th printing), p. 212.

- [3] V.A. Epanechnikov, "Non-parametric estimation of a multivariate probability density", *Theory of Prob. and its Appl.*, Vol. 14, 1969, pp. 153-158.
- [4] W. Feller, *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, New York, 1970 (revised), p. 279.
- [5] K. Henning et al., "Schätzung von Kennwerten nicht-linearer dynamischer Systeme mit Entropiefunktionen", Report of the KDI der RWTH Aachen, No. He-1245, 1987 [in German].
- [6] M.G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, Griffin, London, 1961, pp. 292-294.
- [7] C.H. Knapp et al., "The generalized correlation method for estimation of time delay", *IEEE Trans. Acoust. Speech and Signal Process.*, Vol. 24, 1976, pp. 320-327.
- [8] N.J.I. Mars et al., "Time delay estimation in non-linear systems using average amount of mutual information analysis", *Signal Processing*, Vol. 4, 1982, pp. 139-153.
- [9] N.J.I. Mars et al., "Propagation of seizure activity in kindled dogs", *Electroencephalography and Clinical Neurophysiology*, Vol. 56, 1983, pp. 194-209.
- [10] N.J.I. Mars et al., "Spread of epileptic seizure activity in humans", *Epilepsia*, Vol. 26, 1985, pp. 85-94.
- [11] G.A. Miller, "Note on the bias of information estimates", in: H. Quastler, ed., *Information Theory in Psychology*, Glencoe, Illinois, 1955, pp. 95-100.
- [12] R. Moddemeijer, "AAMI bijdrage van buiten het  $3\sigma$ -gebied bij een binormale verdeling", Internal Report University of Twente, No. 077-25-19, 1984 [in Dutch].
- [13] E. Parzen, "On estimation of a probability density function and mode", *Ann. of Math. Statistics*. Vol. 33, 1962, pp. 520-531.
- [14] C.E. Shannon, "A mathematical theory of communications", *The Bell System Technical Journal*, Vol. 27, 1948, pp. 379-423 and pp. 623-656.
- [15] R.G. van Bergen, "A discrete information estimator of continuous signals", Internal Report University of Twente, No. 080-86-44 (1986).