PERFORMANCE ANALYSIS OF THE MUTUAL INFORMATION FUNCTION FOR NONLINEAR AND LINEAR SIGNAL PROCESSING

Hans-Peter Bernhard

Institute for Communications
and Radio-Frequency Engineering,
Vienna University of Technology,
Gusshausstrasse 25/389,
A-1040 Vienna, Austria.
E-mail: H.P.Bernhard@ieee.org

Georges A. Darbellay

Institute of Information Theory
and Automation, AV ČR
P.O. Box 18, CZ-18208 Prague 8,
Czech Republic
E-mail: dbe@utia.cas.cz

ABSTRACT

Nonlinear signal processing is now well established both in theory and applications. Nevertheless, very few tools are available for the analysis of nonlinear systems. We introduce the mutual information function (MIF) as a *nonlinear correlation function* and describe the practicalities of estimating it from data. Even if an estimator is consistent, it is of great interest to check what the bias and variance are with a finite sample. We discuss these questions, as well as the computational efficiency, for two estimators. Both algorithms are of the complexity $N \log_2 N$, where N is the sample length, but they use different methods to find the histogram for the estimation of the mutual information. An efficient implementation makes it possible to apply the algorithm on real time signal processing problems where the linear correlation analysis breaks down. Current applications are: mobile radio channels, load curve forecasting, speech processing, nonlinear systems theory.

1. INTRODUCTION AND MOTIVATION

Most work concerning nonlinear signal processing is done on nonlinear methods to simulate and design systems. The analysis of nonlinear phenomena suffers from the lack of adequate methods. Higher order statistics are used in this area, but apart from this there are few tools for nonlinear analysis.

Nonlinear systems are often investigated with linear methods. Obviously, the results are restricted to the scope of the linear theory, which may be helpful, but it withholds information related to the nonlinear properties of the system.

Higher order statistics gain more information from the observed nonlinear signal, but we are limited by the order of the statistic being used. Furthermore, the higher the order, the more restrictive become the conditions that must be imposed for the existence of the moments.

To overcome these limitations we introduced the mutual information function (MIF) as a nonlinear correlation function for signal processing and time series analysis [1, 2]. We developed the MIF on the basis of the theoretical work of Kolmogorov, Gelfand, Yaglom, Perez, Dobrushin [3] (and references therein) and Sinai [4], as well as the practical work of Fraser [5]. The MIF extends the correlation function, suitable for linear systems, to nonlinear systems. With the help of the mutual information function it is possible to calculate the maximum prediction gain for virtually every signal [6].

Calculating the mutual information between discrete random variables does not pose any serious problems. For the analysis of random variables taking continuous values, the calculation of the mutual information is quite challenging. In this paper we address the case of continuous random variables.

2. DEMONSTRATION EXAMPLES

To show the potential capability of the MIF we consider two simple examples. The first experiment is a linear channel with an amplifier a and white Gaussian additive noise $n(t)$ with variance σ_n^2, which represents an AWGN-channel. The source signal is a zero mean, white Gaussian signal $x(t)$ with the variance σ_x^2. The received signal $y(t)$ is

$$y(t) = a\, x(t) + n(t). \tag{1}$$

Experiment 2 is similar to experiment 1 but with an ideal AC-DC converter as nonlinear channel. The received signal $y(t)$ is now

$$y(t) = a\, |x(t)| + n(t). \tag{2}$$

We consider the complete channel as a black box, hence we can investigate only $x(t)$ and $y(t)$. In the next section we show that it is possible to measure the information transmitted over the channel using the MIF, even if the channel is nonlinear. The corresponding analysis using the correlation function cannot detect any usable information. In this paper we concentrate on the MIF and leave out all other methods for nonlinear analysis.

3. THEORY

We now analyse the two channels introduced above by means of the linear correlation and the mutual information. Let $X(t)$ and $Y(t)$ be the random variables taking the values $x(t)$ and $y(t)$ respectively.

For the linear channel the cross correlation function is

$$R(X(t), Y(t+\tau)) \equiv R_{X,Y}(t,\tau) = \begin{cases} a\sigma_x^2 & \text{for} \quad \tau = 0 \\ 0 & \text{for} \quad \tau \neq 0 \end{cases}, \tag{3}$$

and the coefficient of linear correlation is

$$r_{X,Y}(t,\tau) = \begin{cases} \frac{a\sigma_x}{\sqrt{a^2\sigma_x^2+\sigma_n^2}} & \text{for} \quad \tau = 0 \\ 0 & \text{for} \quad \tau \neq 0 \end{cases}. \tag{4}$$

In the nonlinear case the cross correlation function is zero for all τ

$$R_{X,Y}(t,\tau) = 0 \tag{5}$$

The same applies to the coefficient of linear correlation $r_{X,Y}(t,\tau)$. Obviously, there is some information contained in the input signal about the output signal. Yet, it is not possible to detect it with the standard correlation function. To overcome these problems we consider the mutual information function.

Definition 1 *The mutual information function is defined as*

$$M_{\vec{X},\vec{Y}}(t,\tau) = I(\vec{X}(t); \vec{Y}(t+\tau)) \tag{6}$$

with $\tau \in \mathbb{R}$ for continuous-time processes $\vec{X}$ and $\vec{Y}$.

The right hand side of (6) is the mutual information

$$I(\vec{X};\vec{Y}) = \int \cdots \int p(\vec{x},\vec{y}) \log_2 \frac{p(\vec{x},\vec{y})}{p(\vec{x})p(\vec{y})} d\vec{x}d\vec{y}. \tag{7}$$

The mutual information measures the full dependence between the vectors of random variables $\vec{X}$ and $\vec{Y}$, and not only the linear component of the dependence. The functions $p(\vec{x},\vec{y})$, $p(\vec{x})$, $p(\vec{y})$ are, respectively, the joint probability density of $(\vec{X},\vec{Y})$ and the marginal probability densities of $\vec{X}$ and $\vec{Y}$. The MIF is introduced for n-dimensional signals. In this paper, we focus on one-dimensional processes, which is sufficient for the analysis of the two experiments under consideration. The signals are time invariant independent Gaussian processes. Hence, the MIF is zero for all τ except $\tau = 0$ and does not change for different t.

For the linear channel, the mutual information function can be shown to be

$$M_{X,Y}(t,\tau) = \begin{cases} \frac{1}{2}\log_2 \frac{a^2\sigma_x^2+\sigma_n^2}{\sigma_n^2} & \text{for} \quad \tau = 0 \\ 0 & \text{for} \quad \tau \neq 0 \end{cases}. \tag{8}$$

which may be written as

$$M_{X,Y}(t,\tau) = -\frac{1}{2}\log_2\left(1 - r_{X,Y}^2(t,\tau)\right) \tag{9}$$

where $r_{X,Y}(t,\tau)$ is given by (4).

For the nonlinear channel, the mutual information defies any analytical calculation. However the MIF may be reduced to a one-dimensional integral, which easily can be evaluated numerically. The probability density function of $Y(t)$ can be expressed as

$$p_Y(y) = \frac{e^{-\frac{y^2}{2((a\sigma_x)^2+\sigma_n^2)}}}{\sqrt{2\pi((a\sigma_x)^2+\sigma_n^2)}} \mathbf{erfc}\left(-\frac{a\sigma_x y}{\sqrt{2}\sigma_n\sqrt{((a\sigma_x)^2+\sigma_n^2)}}\right) \tag{10}$$

The calculation of the MIF then reduces to

$$M_{X,Y}(t,\tau) = \begin{cases} -\int_{-\infty}^{\infty} p_y(u)\log_2 p_y(u)du - \frac{1}{2}\log_2\left(2\pi e\sigma_n^2\right) & : \tau = 0 \\ 0 & : \tau \neq 0. \end{cases}$$

The MIF tells us how much information there is in X about Y. Some numerical values may be found in Section 5. The MIF bears no relation to the coefficient of linear correlation, which is blind to nonlinearities.

4. ESTIMATION

In practice we often do not know a priori what the probability distributions are. One is thus faced with the problem of estimating the mutual information from data. We assume that we have a sample of N i.i.d. observations $(\vec{x},\vec{y})$ of the pair $(\vec{X},\vec{Y})$.

It is well known in information theory that the mutual information between two continuous random variables (or more generally, vectors of random variables) is the limit of the mutual information between their quantized versions. The quantization of a random variable is obtained by partitioning the range, i.e. the observation space, into non-intersecting intervals covering the whole range. For vectors of random variables the intervals become hyper rectangles.

Partitioning both the observation space of $\vec{X}$ and the observation space of $\vec{Y}$ induces a partition of the joint observation space of $(\vec{X},\vec{Y})$, which is called a product partition. It is simply a regular grid of hyper rectangles.

There exists, however, a far more general result, which states that the restriction to product partitions is unnecessary [3]. This is very important because it allows the construction of adaptive, i.e. data-dependent, partitions. The product partition is built in a single step as it just involves putting a regular grid over the data. Our partition will be constructed through a multi-step procedure, i.e. a sequences of finer and finer partitions. This gives us the flexibility of adapting both the location and the size of the cells (the hyper rectangles) of the partition to the data distribution. The benefits of using such data-dependent partitions over product partitions, is a dramatic decrease in the bias of the estimator, because the data points are used much more efficiently. In higher-dimensional spaces we are thus far more able to resist the curse of dimensionality, a well-known problem of nonparametric estimation. Another advantage is that the multi-step procedure may be optimized so as to be significantly faster than the single-step procedure.

The different ways of implementing a multi-step procedure is the first difference between the two algorithms presented below. A second difference is that algorithm A stops the partitioning procedure when uniformity has been achieved, while algorithm B does it when conditional independence has been achieved.

4.1. Algorithms

The multi-step procedure starts with the whole n-dimensional observation space of $(\vec{X},\vec{Y})$ as a single rectangular cell.

		mutual information analysis						correlation analysis		
		Algorithm A			Algorithm B					
a	theory	mean	var	bias	mean	var	bias	mean	var	bias
0.0	0.0	-4.01e-3	1.27e-4	-4.01e-3	0.00e+0	0.00	3.99e-15	6.47e-005	8.70e-9	6.47e-5
0.2	2.82e-2	2.55e-2	1.22e-4	-2.72e-3	2.85e-2	1.03e-5	2.13e-4	2.86e-002	7.27e-6	3.68e-4
1.0	4.99e-1	5.00e-1	2.56e-4	6.93e-4	5.00e-1	1.11e-4	6.69e-4	5.01e-001	9.91e-5	1.21e-3
5.0	2.35	2.34	3.94e-4	-6.22e-3	2.34	2.54e-4	-5.00e-3	2.35	1.94e-4	1.19e-3
10.0	3.32	3.32	3.75e-4	-8.28e-3	3.30	2.55e-4	-1.95e-2	3.33	2.00e-4	1.26e-3
100.0	6.63	6.59	6.02e-4	-4.42e-2	6.46	3.25e-4	-1.72e-1	6.64	2.01e-4	6.80e-3

Table 1: Estimated values and statistics for the mutual information of the linear channel ($\sigma_x = \sigma_n = 1$)

		mutual information analysis						correlation analysis		
		Algorithm A			Algorithm B					
a	theory	mean	var	bias	mean	var	bias	mean	var	bias
0.0	0.0	-4.06e-3	1.39e-4	-4.06e-3	3.51e-6	2.47e-9	3.51e-6	7.42e-5	1.07e-8	7.42e-5
0.2	1.04e-2	6.61e-3	1.45e-4	-3.80e-3	9.87e-3	7.28e-6	-5.36e-4	7.37e-5	1.11e-8	-1.03e-2
1.0	2.21e-1	2.19e-1	2.20e-4	-2.24e-3	2.21e-1	7.87e-5	-8.33e-4	1.00e-4	2.81e-8	-2.21e-1
5.0	1.55	1.55	3.52e-4	-7.08e-3	1.55	2.32e-4	-7.19e-3	1.85e-4	1.13e-7	-1.55
10.0	2.43	2.43	3.77e-4	-6.26e-3	2.42	2.80e-4	-1.71e-2	1.98e-4	1.26e-7	-2.43
100.0	5.65	5.59	4.86e-4	-5.91e-2	5.45	3.18e-4	-1.98e-1	2.04e-4	1.30e-7	-5.65

Table 2: Estimated values and statistics for the mutual information of the nonlinear channel ($\sigma_x = \sigma_n = 1$)

Then the following two rules are applied recursively.

Algorithm A	Algorithm B
(R1) Subpartition a cell into 2^n subcells by dividing each one of its n edges into two *equidistant* intervals.	(R1) Subpartition a cell into 2^n subcells by dividing each one of its n edges into two *equiprobable* intervals.
(R2) Stop the subpartitioning of a cell if the vectors of random variables $\vec{X}$ and $\vec{Y}$ are *uniformly distributed* on it.	(R2) Stop the subpartitioning of a cell if the vectors of random variables $\vec{X}$ and $\vec{Y}$ are *conditionally independent* on it.

In practice, algorithm B sorts each one of the n arrays of data so as to speed up the division into equiprobable intervals. Once the partitioning procedure is stopped, the estimate of the mutual information is simply calculated as a finite sum over all the cells $A \times B$ of the partition, A being a subset of the observation space of $\vec{X}$ and B a subset of the observation space of $\vec{Y}$,

$$\hat{I}(\vec{X};\vec{Y}) = \frac{1}{N}\sum_{A\times B} N(A\times B)\log_2\frac{N(A\times B)}{N(A)N(B)} + \log_2 N\ . \tag{11}$$

Here, $N(A \times B)$ denotes the number of points $(\vec{x}, \vec{y})$ falling in the hyper rectangle $A \times B$, $N(A)$ the number of points $\vec{x}$ falling in the hyper rectangle A and $N(B)$ the number of points $\vec{y}$ falling in the hyper rectangle B.

- **Algorithm A**
  The distribution of points, contained in the current cube is tested, and if a uniform distribution is found, the entire cube will be treated as a homogeneous class. The distribution is tested using a χ^2 test, where the null hypothesis is a uniform distribution of the state vectors. For the test all possible subcubes are considered. Then we test if the investigated distribution is uniform and therefore the estimated probability of one subcube is compared to the test probability $p_t(A_k \times B_j) = \frac{1}{2^n}$ of all possible 2^n cubes. This comparison is done by the χ^2 test where

$$\chi^2 = \sum_{k,j}\frac{(N(A_k\times B_j) - N(A\times B)p_t(A_k\times B_j))^2}{N(A\times B)p_t(A_k\times B_j)} \tag{12}$$

  represents the quantity that is tested against a confidential level χ^2_α. If $\chi^2 < \chi^2_\alpha$ then the null hypothesis, uniform distribution, holds with a error probability of α. The values for χ^2_α can be found in statistical tables as in [7]. The used confidential levels depend on the number of investigated data. We applied an adaptive optimization of the χ^2_α-level for different data length. It turns out that the values for α have to vary from 20% to 34% in the two dimensional case, i.e. 1000 samples lead to $\alpha = 29\%$.

- **Algorithm B**
  For testing independence on a given cell $A \times B$ a χ^2 test is used. For this purpose product partitions are good enough. Let $\{A_k \times B_j\}$ be a product partition of the cell $A \times B$. If the statistic

$$\chi^2 = \sum_{k,j}\frac{(N(A_k\times B_j) - N(A\times B)\frac{N(A_k)N(B_j)}{N(A)N(B)})^2}{N(A\times B)\frac{N(A_k)N(B_j)}{N(A)N(B)}} \tag{13}$$

  is "small enough", then the cell $A \times B$ will not be partitioned any further. "Small enough" means that the critical values of the χ^2 test should be chosen such that the significance level of the test does not exceed 5% [2].

	N number of samples				
	250	500	1000	10000	100000
var	3.0e-2	6.9e-3	3.2e-3	3.5e-4	2.7e-5
bias	-2.46e-1	-1.3e-1	-6.5e-2	-7.1e-3	-3.4e-4

Table 3: Dependence of the variance and bias on increasing N (Nonlinear channel with $a = 5$).

5. EXPERIMENTS

For the experiments we assume a very common situation in communication engineering. The channel is unknown but we know input and output signals. The aim of the mutual information analysis is to investigate if the channel is able to transmit information. With the experiments we test the tools, if they are suitable for the analysis.

As we have shown, the parameter a of the demonstration channels affects the resulting correlation or mutual information value. The variation of a is used to show, if the algorithm is able to find the relation between input and output even if the parameter a is varied over a wide range. In our experiments we used $a = \{0.0, 0.2, 1, 5, 10, 100\}$. Additionally, the accuracy of the estimated value depends on the number of samples N considered in the estimating procedure. Therefore we show the dependence between the variance of the estimation error and N.

In Table 1 the results for 100 trials of the linear channel experiment is shown. We use 10000 samples and different parameters to point out how the error variance is influenced by a. The linear correlation analysis is compared to the mutual information analysis by the value of equation (9). On the other hand Table 2 displays the results for the nonlinear channel. As was expected, the correlation analysis is unable to find the dependences and the mutual information analysis remains correct.

Table 3 shows the dependency of the number of samples N. As one can clearly see, the variance and the absolute value of the bias are decreasing with N. We have shown this dependence for just one parameter a but the others behave similarly.

6. GUIDELINES

We presented results for Gaussian and related distributions. Extensive simulations have shown that our partitioning estimator work equally well for non-Gaussian smooth densities. Usually, statistical estimators have difficulties with distributions having fat tails or sharp discontinuities. Our partitioning procedure has no problem with fat tails. It does, however, suffer when sharp discontinuities are present, in the sense that the bias will decrease more slowly with the sample size than for smooth densities. As far as we know, the only densities for which our estimators will not be consistent are the "exotic" densities which display symmetries with respect to the subpartitions used in the χ^2 test. This leads to a systematic underestimation of the mutual information. Other techniques would be needed to deal with such highly symmetric objects (should they be more than just academic curiosities).

Purely deterministic channels are another special case. For them the mutual information diverges to infinity. For a sample of N points, the estimated mutual information is upper-bounded by $\log_2 N$. Again, our estimators remain consistent in this case.

There is a difference in the accuracy between the algorithms. For mutual information results < 1 Algorithm B produces less variance and bias than A. The bias for mutual information results > 1 is smaller if we use algorithm A. So it is possible to select the proper algorithm after one first estimate.

Basically, our multi-step partitioning procedure has a tree structure. With respect to other estimation techniques, this means simplicity and speed. Indeed our estimators are extremely fast. On an ordinary PC, calculations with samples of 10000 points as above take a fraction of a second. The calculation time is of the order $N \log_2 N$. More complicated estimators, such as kernel estimators, would be unable to compete.

7. CONCLUSION

The excellent results of the experiments demonstrate that the transmission of information over a nonlinear channel can be detected. This is in contrast to the failure of the linear correlation analysis. The availability of an accurate, precise and fast MIF estimator provides a good new tool for nonlinear signal processing.

8. REFERENCES

[1] H.-P. Bernhard. The Mutual Information Function and its Application to Signal Processing . Doctoral thesis, Technische Universität Wien, Vienna (Austria), 1997.

[2] G.A. Darbellay. Predictability: An information-theoretic perspective. In P.J.W. Rayner A. Procházka, J. Uhlíř and N.G. Kingsbury, editors, *Signal Analysis and Prediction*, pages 249–262. Birkhäuser, Boston, 1998.

[3] R.L. Dobrushin. General formulation of Shannon's main theorem in information theory. *Am. Math. Soc. Trans.*, 33:323–438, 1959.

[4] Y.G. Sinai. *Topics in Ergodic Theory.* Princeton University Press, Princeton, New Jersy, 1994.

[5] A.M. Fraser. Information and entropy in strange attractors. *IEEE Transactions on Information Theory*, IT-35(2):245–262, March 1989.

[6] H.-P. Bernhard. A tight upper bound on the gain of linear and nonlinear predictors for stationary stochastic processes. *IEEE Transactions on Signal Processing*, November 1998.

[7] I.N. Bronstein and K.A. Semendjajew. *Taschenbuch der Mathematik.* BSB B.G. Teubner Verlagsgesellschaft, Leipzig (Germany), 1983.