# Measuring Instantaneous Directed Dependencies in Interacting Oscillators

Bruno Bauwens
University Gent
Dep. EESA, machinelearning
Technologiepark 913,
B-9052 Zwijnaarde, Belgium
Bruno.Bauwens@Ugent.be

Bart Wyns
University Gent
Dep. EESA, machinelearning
Technologiepark 913,
B-9052 Zwijnaarde, Belgium
Bart.Wyns@Ugent.be

Dieter Devlaeminck
University Gent
Dep. EESA, machinelearning
Technologiepark 913,
B-9052 Zwijnaarde, Belgium
Dieter@autoctrl.ugent.be

Luc Boullart
University Gent
Dep. EESA, machinelearning
Technologiepark 913,
B-9052 Zwijnaarde, Belgium
Luc.Boullart@Ugent.be

Georges Otte
P.C.Dr. Guislain Institute
Dep. of Neurophysiology
F. Ferrerlaan 88A
B-9000 Ghent, Belgium
Georges.Otte@telenet.be

Patrick Santens
University Gent
Internal medicine, neurology
De Pintelaan 185,
B-9000 Ghent, Belgium
Patrick.Santens@ugent.be

### Abstract

Algorithmic information transfer has been theoretically shown to detect directed dependency or causality in bivariate signals in an optimal way. Here its practical use in a non-ideal setting is investigated. First we show the close connection to the common tests for directed interactions. Subsequently, consequences of model non-ideality is described and a deep connection between regularization and (directed) independence tests is concluded. Thereafter, a directed independence test is constructed to detect the presence of phase coupling between two oscillators using support vector regression. Finally the algorithm is used to determine moments of interaction for dynamic-coupled harmonic oscillators, by exploiting full length information of the signals.

## 1 Introduction

Characterization of bivariate interaction has been investigated in different scientific fields such as climatology [20], electronics [2] and the cardio-respiratory system [12]. Of special interest is the detection of communication between different human brain areas [22, 5, 3] relative to different pathologies.

From a mathematical point of view, measuring directed dependencies or causalities is studied in a purely statistical sense in [13, 19, 14, 11]. In [6] the idea of predictability improvement is used to define general Granger causality. In [1, 9] predictability improvement is defined by data compression lengths. In this way the detection of directed dependence is linked to general machine learning principles: minimum description length and universal sequence prediction [7, 18].

From the practical point many algorithms exist to detect non-directed interactions: cross-correlation, coherence and mutual information. To detect directed interactions, linear Granger causality, Geweke's spectra and information transfer is used. Recently new nonlinear techniques have been studied based on analysis in state spaces and investigation of phase evolution mapping which have been compared in [16] but confidence bounds are very cumbersome.

In [1] optimal independence tests are investigated which have the notion of confidence bounds as defining property. The optimal test is shown to be algorithmic mutual information and can be decomposed as the sum of tree tests: common simultaneous information and directed independence in both directions. To evaluate these tests in practice, actual data compressors are used approximating ideal compressors. It has been shown that improved modeling corresponds to improved compression and otherwise around [21, 7].

- General modeling techniques can be used to evaluate the tests. Any improvement in modeling results in an improved test.
- Different modeling techniques are easily combined to obtain improved tests.
- Provided an accurate modeling, the test has simple generic confidence bounds.
- The system is parameter free, the optimal value of any model-parameter can be found by optimizing description length.

An important problem is the investigation of *nonstationary* oscillatory signals. Traditionally one tries to construct tests that conclude directed independence using as few data as possible [17]. By applying the algorithm in a limited time window one tries to answer the question at which moments an influence exists. Here we take an alternate approach and take advantage of information during previous time intervals to increase the temporal resolution of the found interactions and show that even for weak interactions a serious decrease in sample length is possible.

# 2 Algorithmic Information Transfer

Algorithmic information transfer is introduced as a directed independence tests. It is part of a decomposition of the universal undirected independence test. First we start with a short introduction on algorithmic complexity. Definitions and proofs of theorems can be found in [10, 1]

## 2.1 Definitions

The algorithmic complexity $K(x)$ of $x$ is the length of the shortest executable file that generates $x$ on a computer $U$.

**Definition 2.1.**

$$K(x) = min\{l(p)|U(p) \downarrow = x\}$$

In the same way we define $K(x|y)$ as the length of the shortest program that produces $x$ if it has access to a file that contains $y$. $K(x|y \uparrow)$ is the length of the shortest program that generates $x$ using $y$ in such a way that $x_k$ is generated before any of the bits of $y_{k...l(y)}$ has been read. Algorithmic mutual information equals: $I(x;y) = K(x) - K(x|y)$.

A total function $f$ dominates a total function $g$ if there is a $c$ such that $f(x)+c > g(x)$ for all $x$. Let $S$ be a set of total functions and $f \in S$, $f$ is universal in $S$ if $f$ dominates all elements in $S$. A universal element in a set of functions is the element of the set that is bigger up to a constant relative to any other function of the set.

## 2.2 Undirected independence tests

**Definition 2.2.** *Let $P, Q$ be probability distributions. A total function $d : \mathcal{B}^* \times \mathcal{B}^* \to \mathbb{N}$ is a $(P, Q)$-independence test for $P, Q$ iff:*

$$\sum_{x,y} P(x)Q(y)2^{d(x,y)} < 2$$

An independence test $d$ rejects the hypotheses of $x$ and $y$ being independent with confidence $2^{-d(x,y)+1}$. In practical situations we do not know the underlaying distribution $P$ and $Q$. An independence test is *general purpose independence test* (gpi-test) if for all enumerable $P, Q$ a constant $c_{P,Q}$ exists such that $d - c_{P,Q}$ is a $(P, Q)$-independence test.

Is there a universal gpi-test in the set $S$ of monotonically approximatable functions? The answer is positive by theorem 2.3. Denote $\Sigma_2^0$ as the set of functions that are enumerable given an oracle for the halting problem.

**Theorem 2.3.** *The set of gpi-tests in $\Sigma_2^0$ has a universal element which equals $I(x; y) + I(x, y|\xi)$.*

The term $I(x, y|\xi)$ is the mutual information of $x, y$ with the Halting problem. $I(x, y)$ typically grows with the length of the data while $I(x, y|\xi)$ is believed to be bounded by a small constant for real systems and is therefor assumed to be negligable.

## 2.3 Directed Tests

Information flowing from $x$ to $y$ is represented as:

**Definition 2.4.** *Algorithmic information transfer $IT(x \leftarrow y) = K(x) - K(x|y \uparrow)$.*

Denote $p_{x \leftarrow y}$ as the shortest program that generates $x$ from $y$ as in the definition of $IT$. The information arriving in $x$ and $y$ simultaneously is represented as:

**Definition 2.5.** *Common simultaneous common information $IT(x = y) = I(p_{x \leftarrow y}; p_{y \leftarrow x})$*

Algorithmic mutual information $I(x; y)$ can be decomposed now into directed tests and simultaneous common information.

**Theorem 2.6.**

$$I(x; y) =^{+c} IT(x \leftarrow y) + IT(y \leftarrow x) + I(x = y) + dIT(x, y)$$

*with $dIT(x, y) = K(K(x|y \uparrow), K(y|x \uparrow)|x, y) <^c l^*(l(x)) + l^*(l(y))$ and $c$ a constant independent from $x$ and $y$.*

## 2.4 Comparison with Literature

$y$ is said to Granger-cause $x$ if the average predictability of $x_t$ is better using $x_{1...t-1}$ and $y_{1...t-1}$ compared to the the predictability of $x_t$ using only $x_{1...t-1}$ for $t = 1...n$. This idea has been developed in many practical algorithms such as [5, 8, 15, 12]. The approach above fits into this framework by defining predictability improvement as data compression-lengths $K(x)$, $K(x|y \uparrow)$.

On the other side, Shannon information transfer (SIT) uses the definitions of information content relative to a probability distribution to measure the rate of transmitted information between two signals: $SIT = H(X^+|X^-) - H(X^+|X^-Y^-)$ with $X^+$,$X^-$,$Y^-$

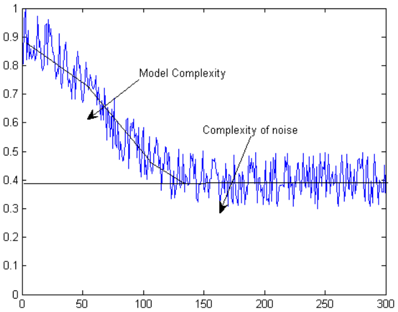

Figure 1: Complexity of Modal and Noise

stochastic variables representing the future and past of signals $x$ and $y$ [15, 12]. The quality of the measure is determined by the quality of the underlaying distribution but the better the probability distribution describes the data, the more over-fitting is present and the measure becomes unreliable. For any distribution $P$ datacompression is linked to Shannon entropies by the formula:

$$\lim_{n\to\infty} \frac{E_P[K(x_{1\ldots n})]}{n} = H(P)$$

# 3 The Influence of Compression non-Ideality

To build practical tests using IT, one uses a datacompressor $C$ to estimating the algorithmic complexities $K(x)$ and $K(x|y$ by $C(x)$ and $C(x|y)$. The approximation of $IT$ is denoted as $CIT$. An ideal compressor returns for every $x$ the shortest executable file that prints $x$. In practice we cannot approximate it's length in a reasonable time, therefor non-ideality of actual compressors are unaviodable. To discuss them, we show a typical plot of the incremental non-ideal compression length $C_x(n) = C(x_{1\ldots n}|n) - C(x_{1\ldots n-1}|n-1)$ for signals with a stationary behaviour. A typical shape is plotted in figure 1. The total area under the plot represents the total compression length. This area can be divided in two parts. The highest part is only visible in the initial bits of the compression. During this stage, the modeling for the compression is significantly adapted resulting in definite improvements for the prediction of the following bits. After sample 150 the changes in the model does not decrease or increase the predictability of the samples significantly. The predictable part represents the complexity of the noise according to the current model. The first area represents the complexity of the model as seen by the compressor and the second part represents the noise in the system.

The model complexity can be very high although the data has a simple structure. This effect can be explained easily by the following example. Assume the given signal is $x_t = (-1)^t + \xi_t$ with $\xi_t$ some noise and we apply auto-regression to approximate $C(x)$: $\hat{x}_t = a_0 + a_1 x_{t-1} + \ldots + a_k x_{t-k}$. In a non-regularised training algorithm any possible value for the vector $a$ has the same a priory probability. During the prediction of the first data samples the training algorithm will over-fit because it prefers a vector $a'$ that also fits the noise over the ideal $a = [0, -1, 1, -1, 1 \ldots 1]/k$.

To approximate $C(x|y)$ one typically has a larger model complexity even if no interactions are present. In the same example we can fit a multivariate auto-regression model $\hat{x}_t = a_0 + a_1 x_{t-1} + \ldots + a_k x_{t-k} + b_1 y_{t-1} + \ldots + b_k y_{t-k}$. But now there are $2k+1$ coefficients resulting in a much bigger model complexity. To cure this problem partly, the model length is not taken into account by ignoring the contribution of the first 150 samples.

The second type of non ideality is induced by fluctuations of $\delta_x(n) = C_x(n) - K_x(n)$ and $\delta_{x|y\uparrow} = C_{x|y\uparrow} - K_{x|y\uparrow}$ the differences between ideal and non ideal compression rates. Typically $x_t$ is distributed approximately independently from $x_{t+s}$ for $s$ large enough, therefore $C(x)$ is distributed with variance of the order $\sqrt{n}\sigma_\delta$, which decreases the confidence according to $\sqrt{n}$ in contrast with the ideal theory that has confidence bounds independent from the length of the segment under consideration.

In conclusion we have that both types of errors can be suppressed by improved normalization of the models. This indicates that better regularisation corresponds to improved tests for independence. The other way around, measuring independence can be used to improve model generalization.

# 4 Measuring Time Dependent Interactions

Suppose two noisy oscillators with momentary interactions. The aim is to detect the time intervals of the interactions. Define the instantaneous ideal and non ideal information transfer $IT_x(n) = K_x(n) - K_{x|y\uparrow}(n)$ and $CIT_x(n) = C_x(n) - C_{x|y\uparrow}(n)$. Notate $Z(m...n) = \sum_{i=m}^{n} Z(i)$ for $Z = C_x, C_{x|y\uparrow}, IT, CIT$.

If no interactions are present for the samples $n$ to $n+\tau$ then $C_x(n...n+\tau) > C_{x|y\uparrow}(n...n+\tau)$. But in the ideal case, the compression with side information can never be lower than without. This is caused by over-fitting and the presence of noise. On the other hand, if interactions are present, $C_{x|y\uparrow}(n)$ decreases. A new compressor $C'$ is build to ensure $C(x) > C'(x|y\uparrow)$ by the results of the compression $C_x$ and $C_{x|y\uparrow}$. At each time instance $n$ a classification algorithm decides to predict according to $C_x$ or to $C_{x|y\uparrow}$. It turns out that if $C'(x|y\uparrow) << C(x)$ the quantity $C'IT_x(n...n+\tau) = C_x(n...n+\tau) - C_{x|y\uparrow}(n...n+\tau)$ returns the expected exponential confidence bounds. Any parameters for the classification algorithm can be trained by optimizing the compression length.

We can construct a third compressor $C''$ by using two compressors $C^{int}$ and $C^{noInt}$. $C^{int}$ is trained by the samples which were classified as interacting by the classification algorithm above and $C^{noInt}$ was trained by the other samples. In this way, the modelling can be improved iteratively, but a second iteration did not result in a significant improvement.

Finally we decide at each time point the presence of interactions by the formula:

$$\frac{C''IT_{x|y\uparrow}(n-k...n+k)}{2k+1} > \frac{C''IT(x|y\uparrow)}{l(x)}$$

# 5 Tests and results

The algorithm described above was implemented using support vector regression. Support vector regression outperformed linear regression and neural networks because because in errors of both types as described in section 3. This is not surprising because support vector regression was invented using a rigorous mathematical regularisation theory: structural risk minimization. The compressor $C'$ classified each $n$-th sample according to the sign of $CIT_{x|y\uparrow}(n-k...n-1)$. $k$ was trained to have optimal compression length $C'(x)$.

To relate mean square error into datacompression, one estimates the variance $\hat{\sigma}$ of the data and assumes for each prediction $\hat{x}_n$ that $x_n - \hat{x}_n$ a Gaussian distribution $\sigma$. In this simple case the compression length is closely related to mean square error.

The algorithm was tested by data sampled from coupled oscillators with momentary interactions $i(n) = 0, 1$. To detect the presence of interactions:

$$\begin{aligned}\dot{\phi}_1 &= \omega_1 + \epsilon i(t) \sin(\phi_2 - \phi_1) + \xi_1 \\ \dot{\phi}_2 &= \omega_2 + \xi_2\end{aligned}$$

The noise $\xi_1$ and $\xi_2$ are independent and Gaussian distributed with mean 0 and standard deviation 0.2. The frequencies are $\omega_1 = 0.9$ and $\omega_2 = 1.1$. The oscillators were integrated using the Euler method in time steps of 0.005. The signals were sampled at times $0, \pi/7, 2\pi/7, ...$. The interaction $i(m)$ was switched on first for a long period and interchanged rapidly after this, see figure 2f.

The average accuracy of $\hat{i}(t)$ relative to $i(t)$ versus interaction strength was plotted for an interaction length $n = 100$ and $n = 500$. In the second case the accuracy is bigger because most errors occur near the borders of the interaction regions in 2a. For very short interactions periods and strong interaction, the accuracy remains remarkable high as can be seen in 2b.

The method was compared to SIT as described in [12]. It turned out that for these short signals estimated entropies with 4 bins outperformed the choice of any other bin number. To compare both methods we asked them to classify the non-interacting and interacting intervals. To compare the reliability of the classification we consider the area under the ROC curve (AUC) to evaluate both IT and SIT as a classification parameter. AUC are standardly used to compare the performance of ranking variables for classification by a ranking parameter [4].

The AUC of SIT and IT were plotted in figure 2c and 2d. In figure 2c we find AUC slightly below 0.5 for both SIT and AIT in the case of very weak interactions. This slight remarkable effect can be explained to remark that the support vector regression model for $C(x|y \uparrow)$ and the histograms for calculating $H(X^+|X^-Y^-)$ are not able to discover any relationship in the short data in case of short very weak interactions. Moreover, there performance decreases disproportionate due to an increase of noise in $\phi_1$ during the interaction ($\xi_1$ vs. $\xi_1 + \epsilon\xi_2$). Because the increase of noise is bigger for $C(x|y \uparrow)$ and $H(X^+|X^-Y^-)$ than for $C(x)$ and $H(X^+|X^-)$ as discussed in section 3, both SIT and IT suffer from reverse ordering and biased classification in the case of short and very weak interactions. However, in the case of AIT this effect is clearly smaller.

Figure 2d demonstrates the validity of the theoretical confidence for the classification of segments with a large variety of lengths and interaction strengths.

# 6 Conclusion

Algorithmic information transfer has been shown to be a useful concept for constructing and investigating directed independence tests. A deep connection between modeling-non-ideality and regularization was demonstrated. An implementation using support vector regression showed to be a reliable method for classifying segments of interaction and non-interaction that outperformed Shannon information transfer. Finally we showed the possibility to detect the moments of interaction even for very short interaction lengths.

## Acknowledgment

Bruno Bauwens was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The authors thank Willem Waegeman for discussions.

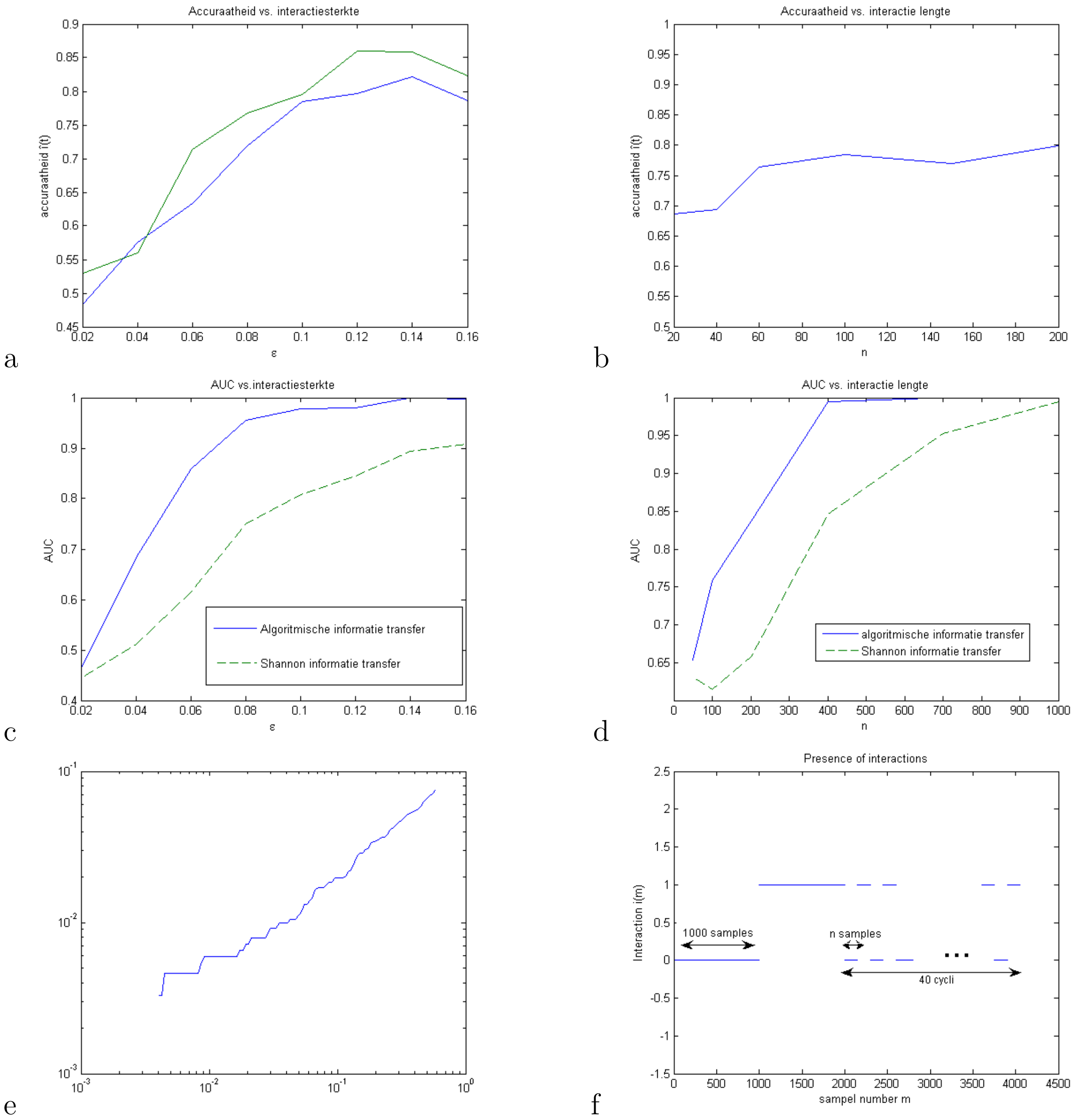

Figure 2: Performance of algorithmic information transfer

(a) accuracy of $\hat{i}(n)$ with respect to $i(n)$ versus interaction strength $\epsilon$ for interaction lengths 100 and 500. (b) same accuracy versus interaction length for strong interaction strength$\epsilon = 0.1$. (c) AUC of algorithmic and Shannon information transfer for the classification of interacting and non-interacting signals with length 200 versus interaction strength. (d) Same AUC for an interaction strength of 0.05 and varying interaction length. (e) Actual confidence versus theoretic confidence for the classification of 600 interacting and non-interacting signals with lengths interaction strengths $\epsilon$ =0.035,0.06,0.08 en 0.1 and interaction lengths $n$=20, 50, 100, 200, 500, 700. (f) Presence of interactions versus samples.

# References

[1] Bruno Bauwens, Bart Wyns, Dieter Devlaminck, Georges Otte, Luc Boullart, and Patrick Santens. Mutual information and algorithmic information transfer as ideal undirected and directed independence tests. In *Proceedings of the 2007 International Conference on Foundations of Computer Science, LasVegas*, 2007.

[2] B. Bezruchko, V. Ponomarenko, M. G. Rosenblum, and A. S. Pikovsky. Characterizing direction of coupling from experimental observations. *Chaos*, 13:179–184, March 2003.

[3] M. Chavez, J. Martinerie, and M.L.V. Quyen. Statistical assessment of nonlineair causality: Application to epileptic eeg signals. *Journal of Neuroscience Methods*, 123:113–128, 2003.

[4] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[5] U. Feldmann and J. Bhattacharya. Predictability improvement as an asymmetrical measure of interdependence in bivariate time series. *International Journal of Bifurcation and Chaos*, 14(2):505–514, 2004.

[6] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969.

[7] P.D. Grunwald, I.J.Myung, and M.A.Pitt. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[8] M. Kaminski, M. Ding, W.A. Truccolo, and S.L. Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 2001.

[9] Jan Lemeire and Erik Dirkx. Causal models as minimal descriptions of multivariate systems. In *Causality and Probability in the Sciences*, 2006.

[10] M. Li and P.M.B. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 1993.

[11] I. Martel. *Probabilistic Empiricism: In Defence of a Reichenbachian Theory of Causation and the Direction of Time*. PhD thesis, University of Colorado, 2000.

[12] M. Palus and A. Stefanovska. Direction of coupling from phases of interacting oscillators: an information theoretic approach. *Physical Review E, Rapid Communications*, 67:055201(R), 2003.

[13] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[14] J.M. Brett R.J. Lawrence, S.A. Mulaik. *Causal Analysis*. Sage Publications, 1983.

[15] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2), 2000.

[16] D.A. Smirnov and R.G. Andrzejak. Detection of weak directional coupling: Phase-dynamics approach versus state-space approach. *Physical Review E*, 71(3):036207–+, March 2005.

[17] Dmitry A. Smirnov and Boris P. Bezruchko. Estimation of interaction strength and direction from short and noisy time series. *Phys. Rev. E*, 68(4):046209, Oct 2003.

[18] Ray J. Solomonoff. A formal theory of inductive inference. part II. *Information and Control*, 7(2):224–254, 1964.

[19] P. Suppes. *A Probabilistic Theory of Causality*. North-Holland Publishing Company, 1970.

[20] P. F. Verdes. Assessing causality from multivariate time series. *Physical Review E*, 72(2):026222–+, August 2005.

[21] N.K. Vereshchagin and P.M.B. Vitanyi. Kolmogorov's structure functions and model selection. *IEEE Trans. Inform. Theory, 3265- 3290*, 2004.

[22] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, R. Bauer, J. Timmer, and H. Witte. Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing*, 85:2137–2160, 2005.