Time-lagged causal information:
A new metric for effective connectivity analysis

Mark E. Pflieger
Source Signal Imaging, Inc.

mep@sourcesignal.com

Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19-22, 2003, New York, NY. Abstract available on CD-Rom in *NeuroImage*, Vol. 19, No. 2.

Introduction

Brain organization at the systems level may be conceived as a collection of semi-autonomous modules with modulated functional interconnections (constrained by anatomical pathways) that dynamically self-organize to coordinate cognition, emotion, and behavior (Bressler, 1995; Bressler & Kelso, 2001). Neuroimaging with cognitive paradigms typically identifies multiple active brain areas, which presumably cooperate somehow, via a network of functional connections, to accomplish a given task (Mesulam, 1990). Effective connectivity analysis aims to detect and characterize causal relationships in such networks (Friston 1994; Horwitz, Tagamets, & McIntosh, 1999).

A general problem is to assess causal relationships between processes based on the observation of multiple time series (e.g., obtained from a functional neuroimaging modality, such as fMRI or M/EEG). Granger (1969) formalized a causality concept essentially as follows: Process *A* does *not* cause process *B* if (and only if) the ability to predict *B*'s observables based on the histories of all observables is unaffected by the omission of *A*'s history. Metrics for Granger causality typically have been realized in the framework of multivariate Gaussian statistics via vector autoregressive (VAR) models (Kaminski et al., 2001). Using information theory, Diks and DeGoede (2001) have realized a more general Granger causality measure that accommodates in principle arbitrary statistical processes (whether linear or nonlinear, Gaussian or non-Gaussian).

In this work, the *state* of a process at time *t* is considered to encapsulate the relevant history of the process up to time *t*. Moreover, it is the state of a process that is to be predicted. Supposing that the process is governed by some stochastic differential equation, then—even if the equation remains unknown—the derivatives of the process at time *t* may be regarded as the state vector. For a discrete time series, the problem of determining the state becomes one of estimating the derivatives, which requires (a) a temporal window of some suitable length, and (b) an estimate of the order of the differential equation. Beyond this, no attempt is made here to obtain specific models of the processes or their couplings. Rather, the immediate objective is to quantify couplings and characterize time lags via empirical estimates of process entropies, which may be combined in ways to tease apart causal relationships.

Theory

Objective: Given observed time series at each node of a network, our objective is to characterize *causal timing relationships* between pairs of nodes at an abstract level, that is, in a model-neutral fashion. In particular, it is desired to handle both linear and nonlinear dynamic relationships without having prior linear/nonlinear models in mind. This would be a natural first step prior to more detailed modeling.

Predictive information is the reduction of uncertainty about the state vector of process *A* conditioned by observation of the state of process *B* at an earlier time. For a given time lag τ, predictive information may be quantified as the *mutual information* of the state of process *A* at time $t-\tau$ and the state of process *B* at time *t*. As will be discussed at the end of this presentation, predictive information is necessary, but not sufficient, for *causal information*. Because it is logically necessary to establish predictive information before establishing causality, and because the underlying methods used to estimate predictive information are essentially the same as those for estimating more complex causality constructs, this presentation focuses primarily on predictive information.

Formally, the predictive information as a function of time lag is

$$\mathrm{P}[\mathbf{A},\mathbf{B}](\tau) \equiv I(\mathbf{A}(t);\mathbf{B}(t-\tau)) \equiv H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A}(t),\mathbf{B}(t-\tau))$$

where I is the mutual information of two processes, and H is the *Shannon entropy* (Shannon & Weaver, 1949). Thus, estimation of predictive information boils down to estimation of a multivariate entropy, which is defined as

$$H(\mathbf{x}) \equiv -\int_{\mathbf{X}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

where $p(\mathbf{x})$ is the probability density of the vector variable $\mathbf{x}$. If the processes have linear dynamics, then their statistics are Gaussian. The "linear entropy", or entropy of a multivariate Gaussian process, is—except for a constant that doesn't depend on the data—given by

$$H_{lin}(\mathbf{x}) = \frac{1}{2}\sum_{d=1}^{D} \log w_d$$

where the w's are the eigenvalues of the covariance matrix that characterizes the multivariate Gaussian process.

If the data are “sphered” by the transformation of multiplying by the inverse square root of the covariance matrix, then the entropy will be reduced by the above quantity. That is, the linear entropy will have been extracted, and what remains may be rightly called the “nonlinear entropy”.

The problem that remains is to estimate the entropy of the sphered data sample. Of various numerical methods that are available (see Beirlant *et al*., 1997), the method implemented here is the leave-one-out resubstitution method (Ivanov & Rozhkova, 1981) that utilizes a Gaussian kernel probability density estimator:

$$\hat{H}_{nonlin}(Z) = -\frac{1}{I}\sum_{i=1}^{I}\log\frac{1}{I-1}\sum_{j\neq i}\frac{1}{(2\pi)^{D/2}\sigma}e^{-\frac{|\mathbf{z}_j-\mathbf{z}_i|^2}{2\sigma^2}}.$$

Here *I* is the number of samples, $\mathbf{z}_i$ is the *i*th sample vector, and σ is a *scaling parameter.*

Methods

Data consisting of two coupled time series were generated as described on the next page.

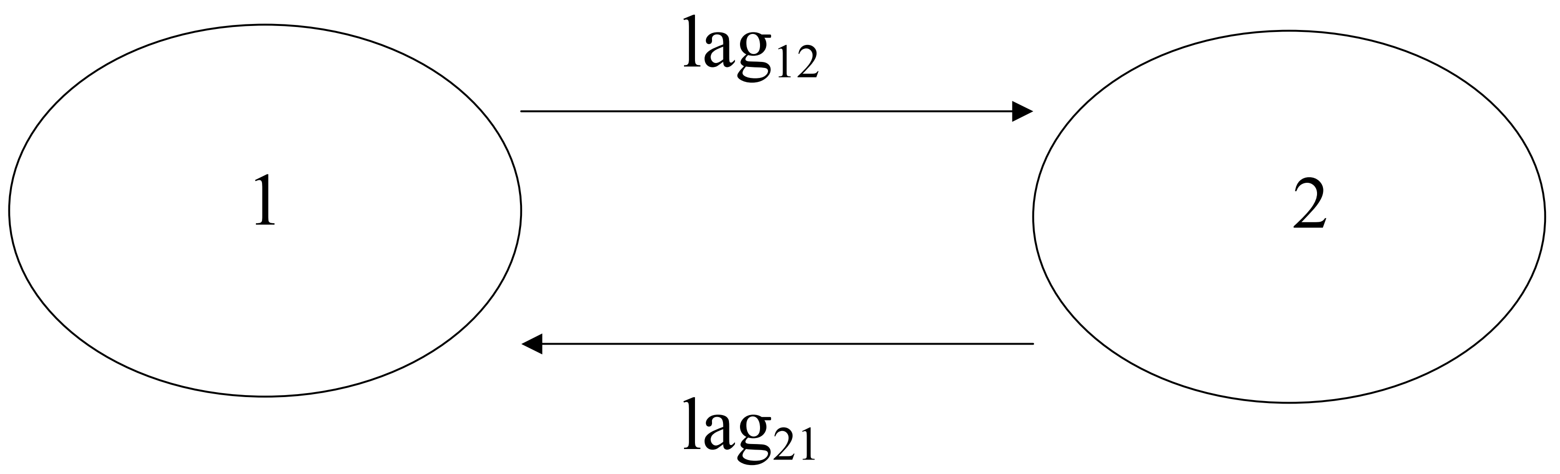
To estimate the state space of each process, *singular spectrum analysis* (SSA; Broomhead & King, 1986) was performed separately for each time series. An exponentially decaying window (from present to past) was used, with the decay coefficient and window length determined automatically based on the autocorrelation function. Singular Value Decomposition (SVD) was performed on the sliding windowed data, and the number of components retained was just adequate to account for at least 95% of the data variance. The obtained SSA functions were then used to characterize the “state” of the time series at any given time *t*.

A random sample of 1000 time points was drawn, subject to the constraint that no two times were closer than the maximum SSA window.

Linear and nonlinear predictive information (based on linear and nonlinear entropies) were computed as a function of time lag. The linear cross-correlation function was also computed for comparison.

Data Generation

Autoregressive moving average (ARMA) processes were used to simulate data in order to test the methods just described. Pairs of coupled processes were generated with specified time lags for inter-process interactions to occur, as illustrated below.



The general equation for a generated ARMA(*p, q*) process is

$$x_i(t) = \sum_{j=1}^{n}\left(\sum_{k=1}^{p}\alpha_{jik} f_{ji}(x_j(t-k-\tau_{ji})) + \sum_{k=1}^{q}\beta_{jik} g_{ji}(\varepsilon_j(t-k-\tau_{ji}))\right)$$

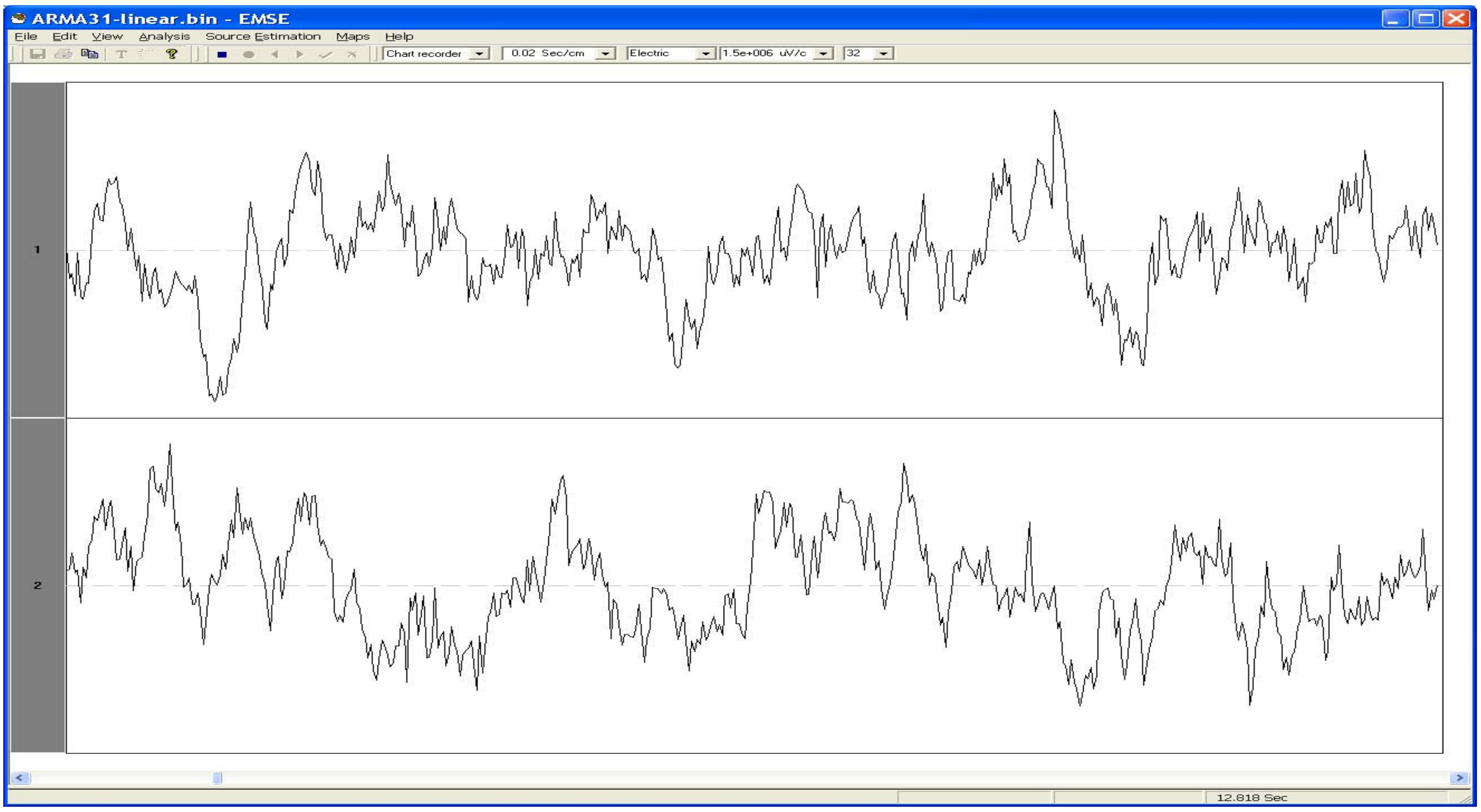
where $x_i(t)$ is the process value at node *i* at time *t*; $\varepsilon_j(t)$ is a Gaussian process with unit variance; *p* is the AR order; *q* is the MA order; the alphas and betas are AR and MA coefficients, respectively; τ_{ji} is the effective lag from node *j* to node *i*; and *f* and *g* are functions that may introduce nonlinear couplings.

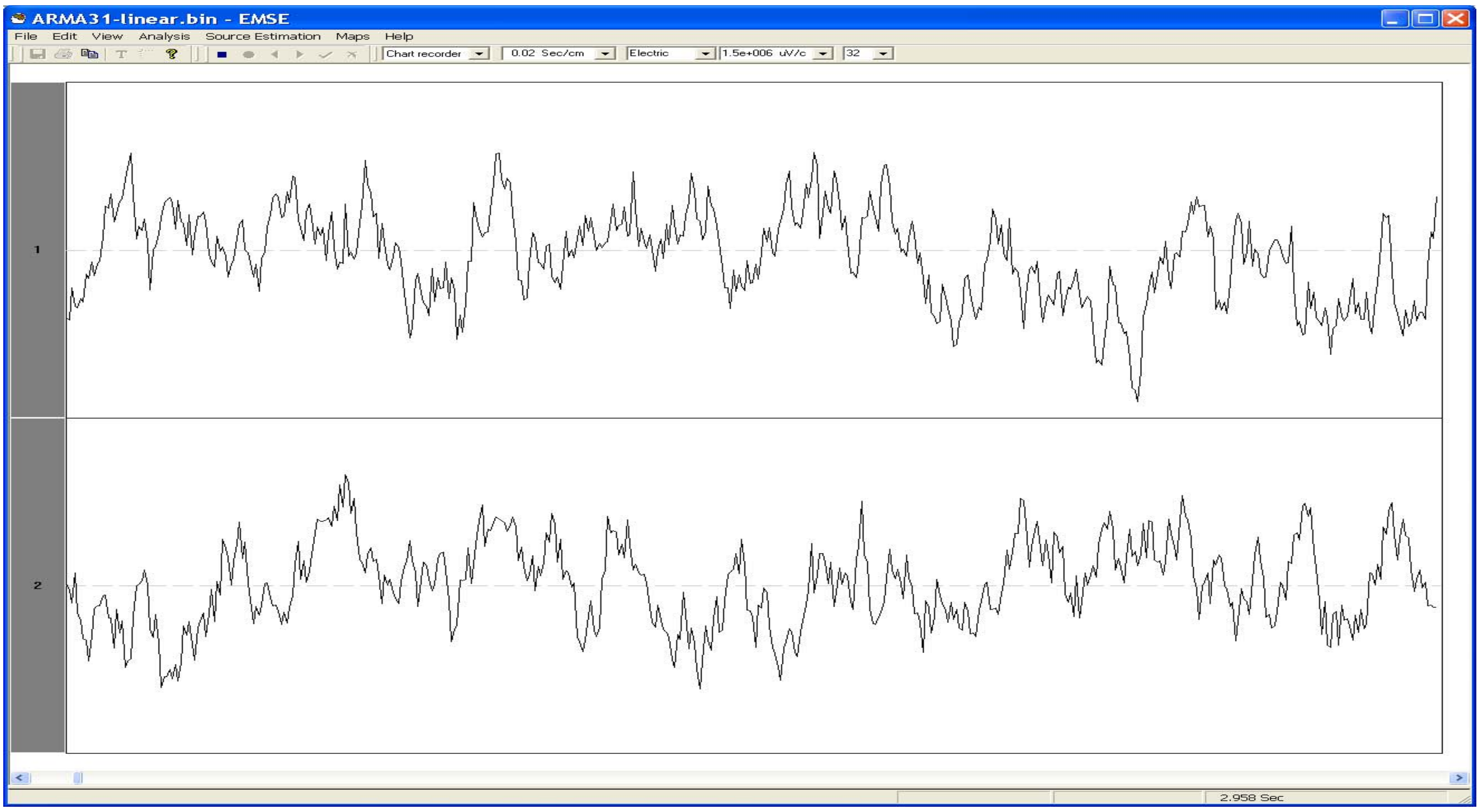
Two specific ARMA(3, 1) processes are illustrated on subsequent pages: one "linear" and the other "nonlinear". In both cases, *n*=2 time series, and both time series have identical internal dynamics, namely: the MA coefficient (beta) is 0.8 (with *g* the identity), and the AR coefficients (alpha) are 0.75, 0.15, and 0.05, respectively. The cross coupling alpha coefficients are –0.40, 0.00, and 0.40, respectively. ***The lag from 1 to 2 is 30 ms, and the lag from 2 to 1 is 40 ms***. The only difference between the "linear" and "nonlinear" processes is that the *f* function is the identity for the linear process, and is the following for the cross-coupling terms of the nonlinear process:

$$f(x) = |x|^{0.8}.$$

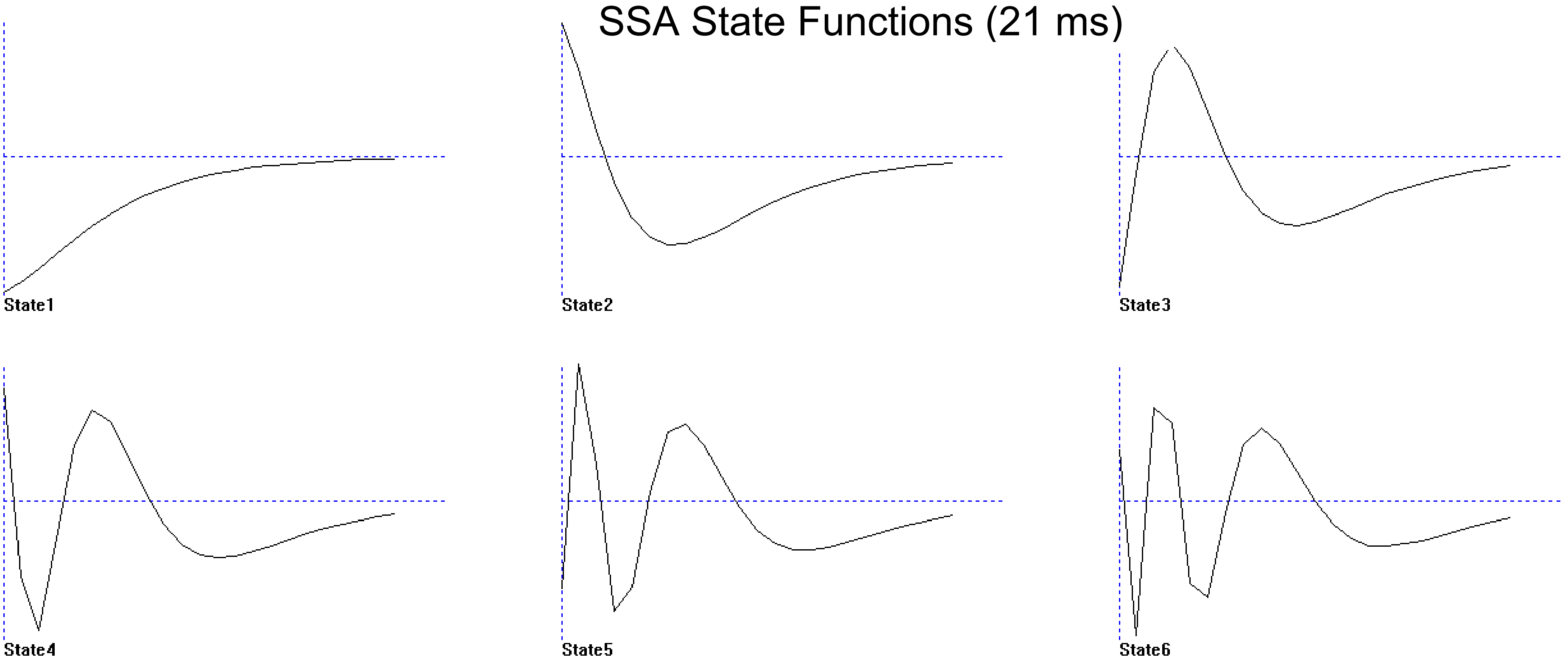
This nonlinearity (among others) was studied in Granger & Lin (1994) and Granger *et al.* (2000).

ARMA(3,1) Processes with **Linear** Coupling (500 ms)

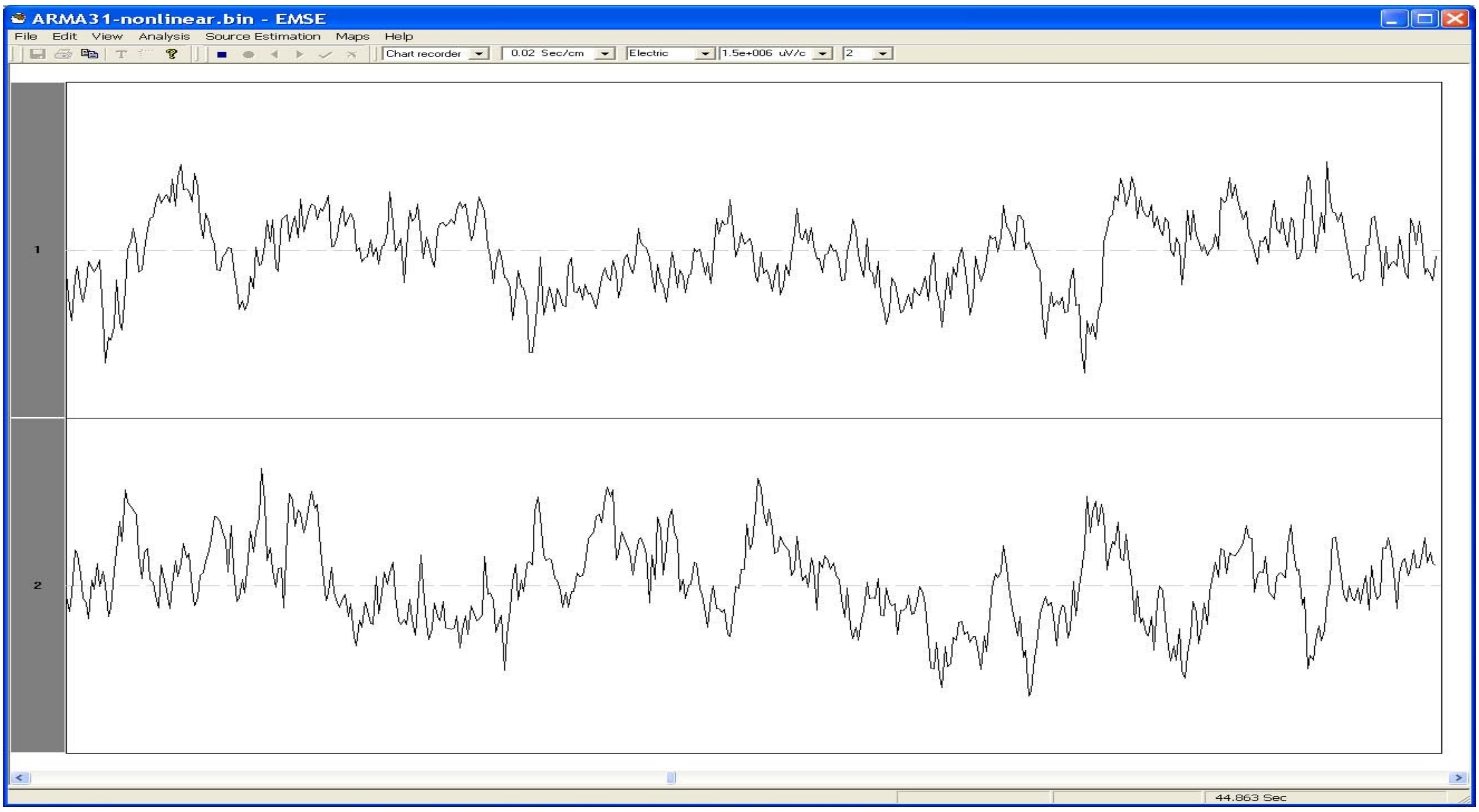


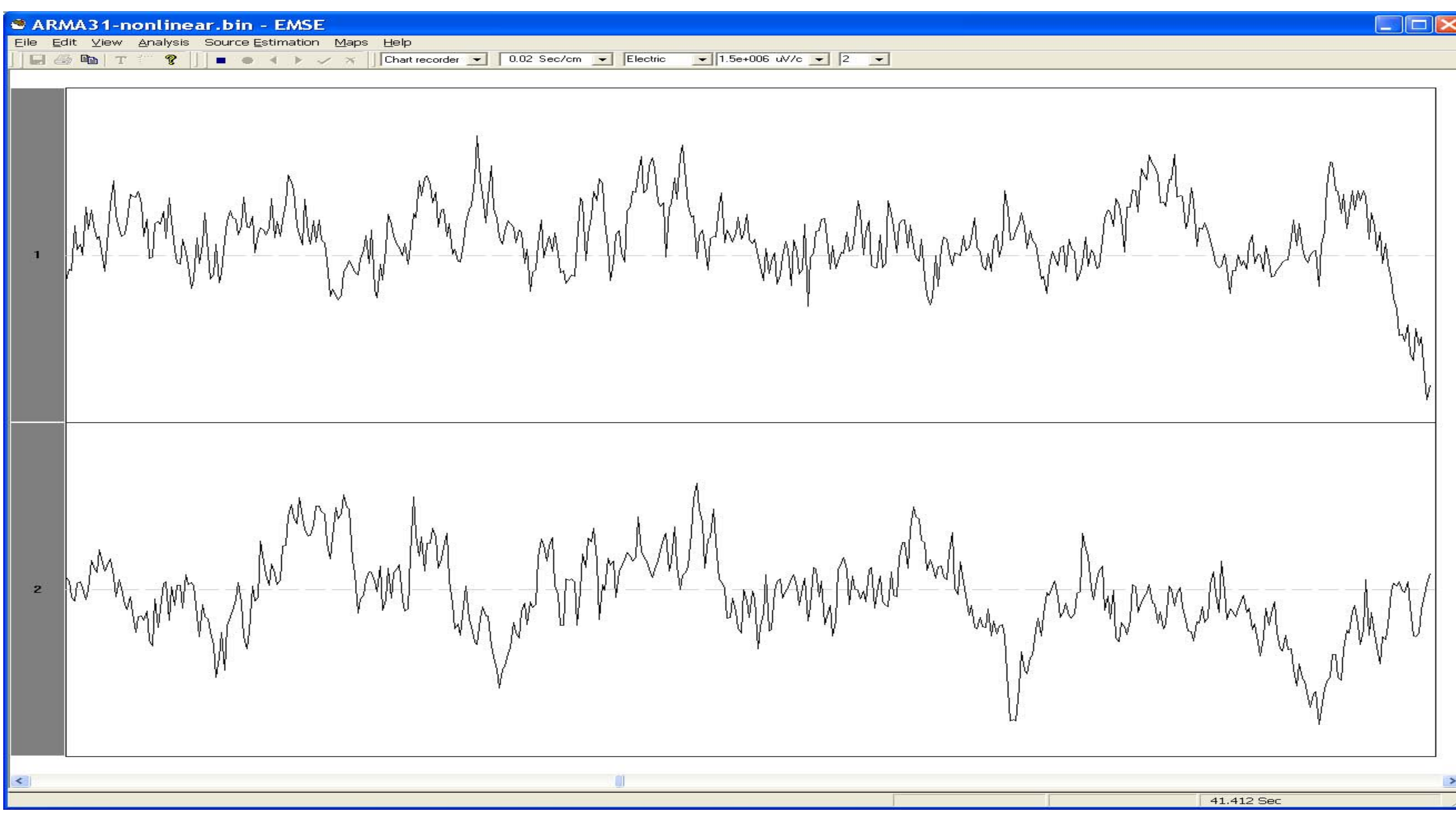


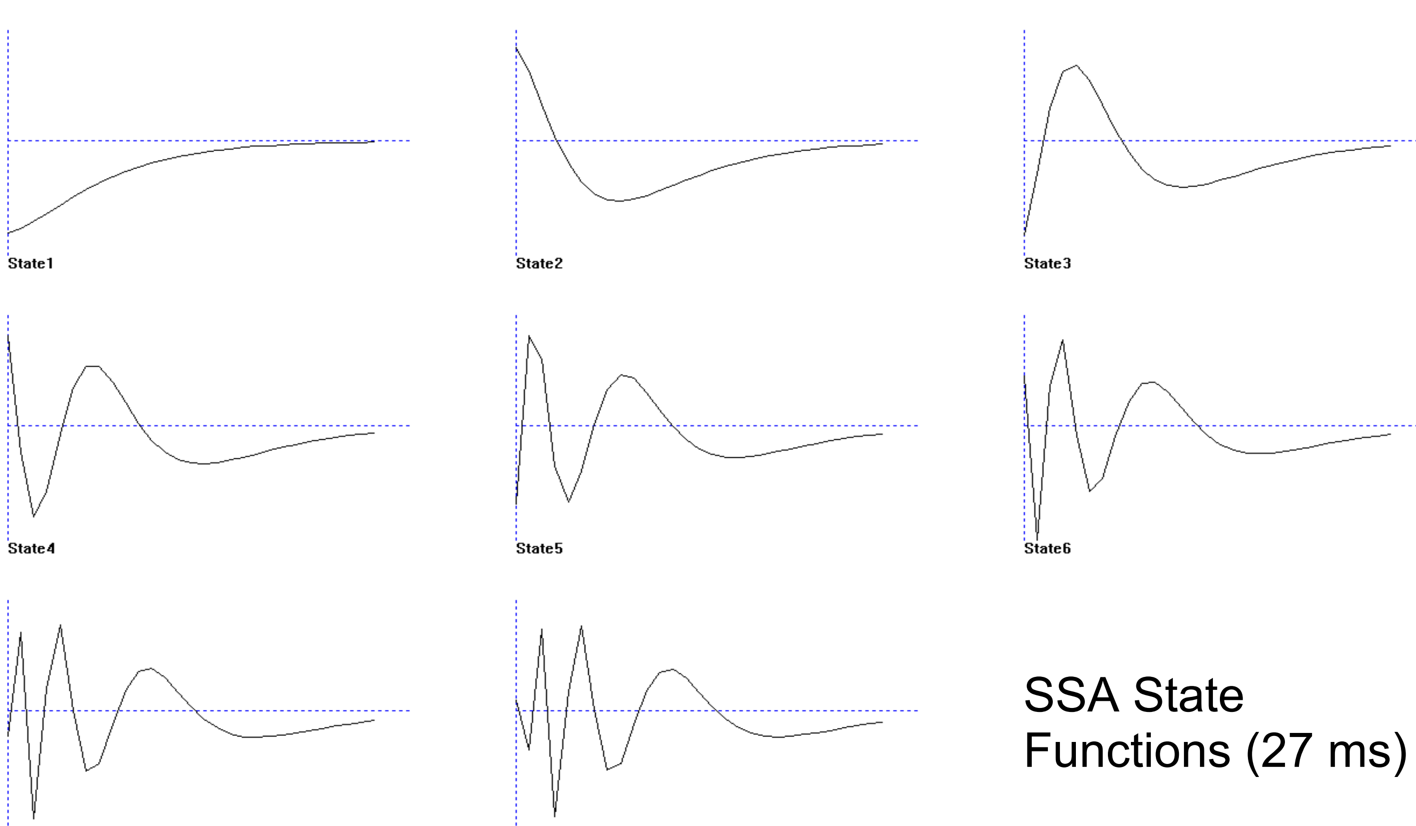
SSA State Functions (21 ms)



ARMA(3,1) Processes with **Nonlinear** Coupling (500 ms)

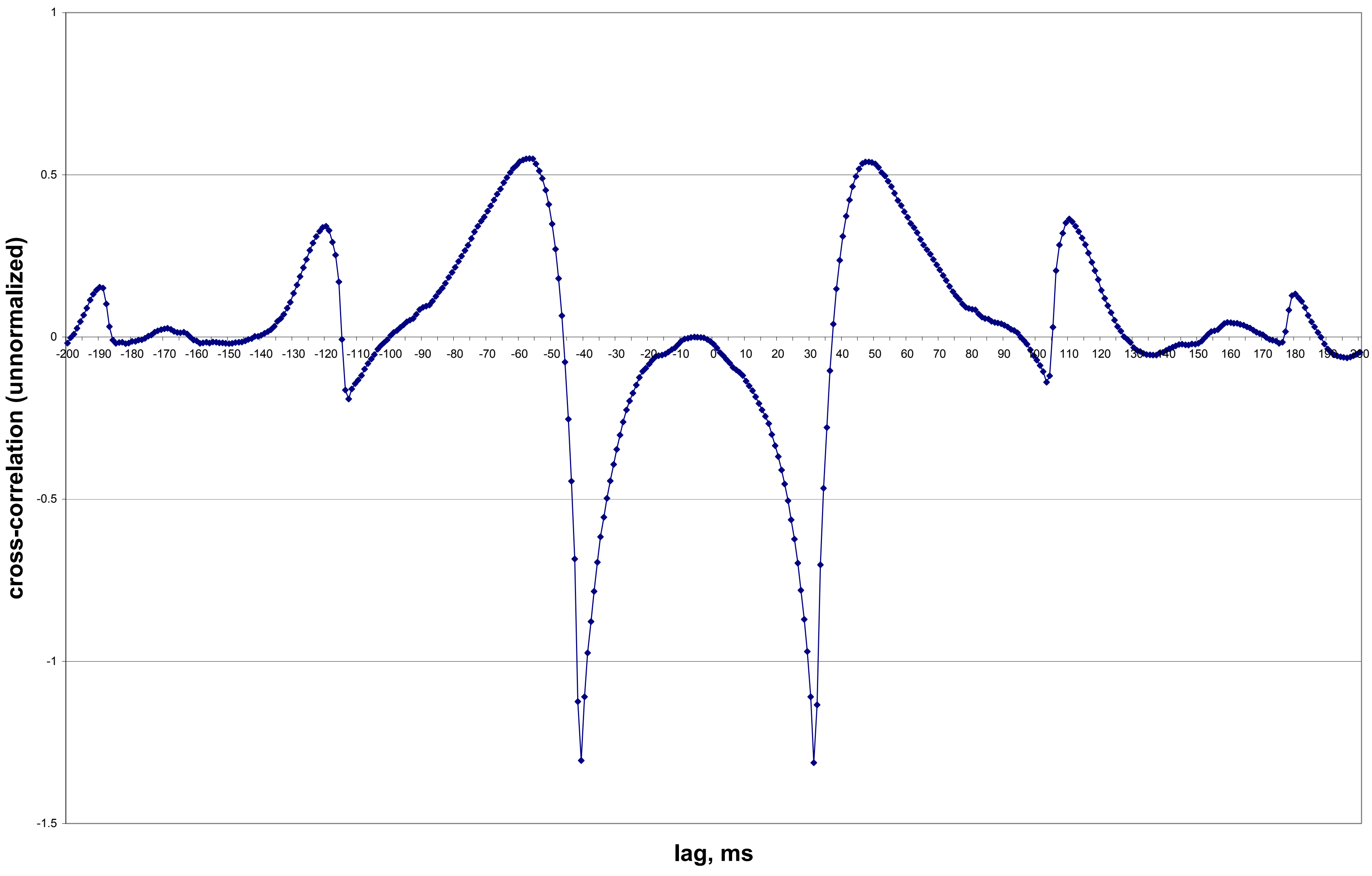




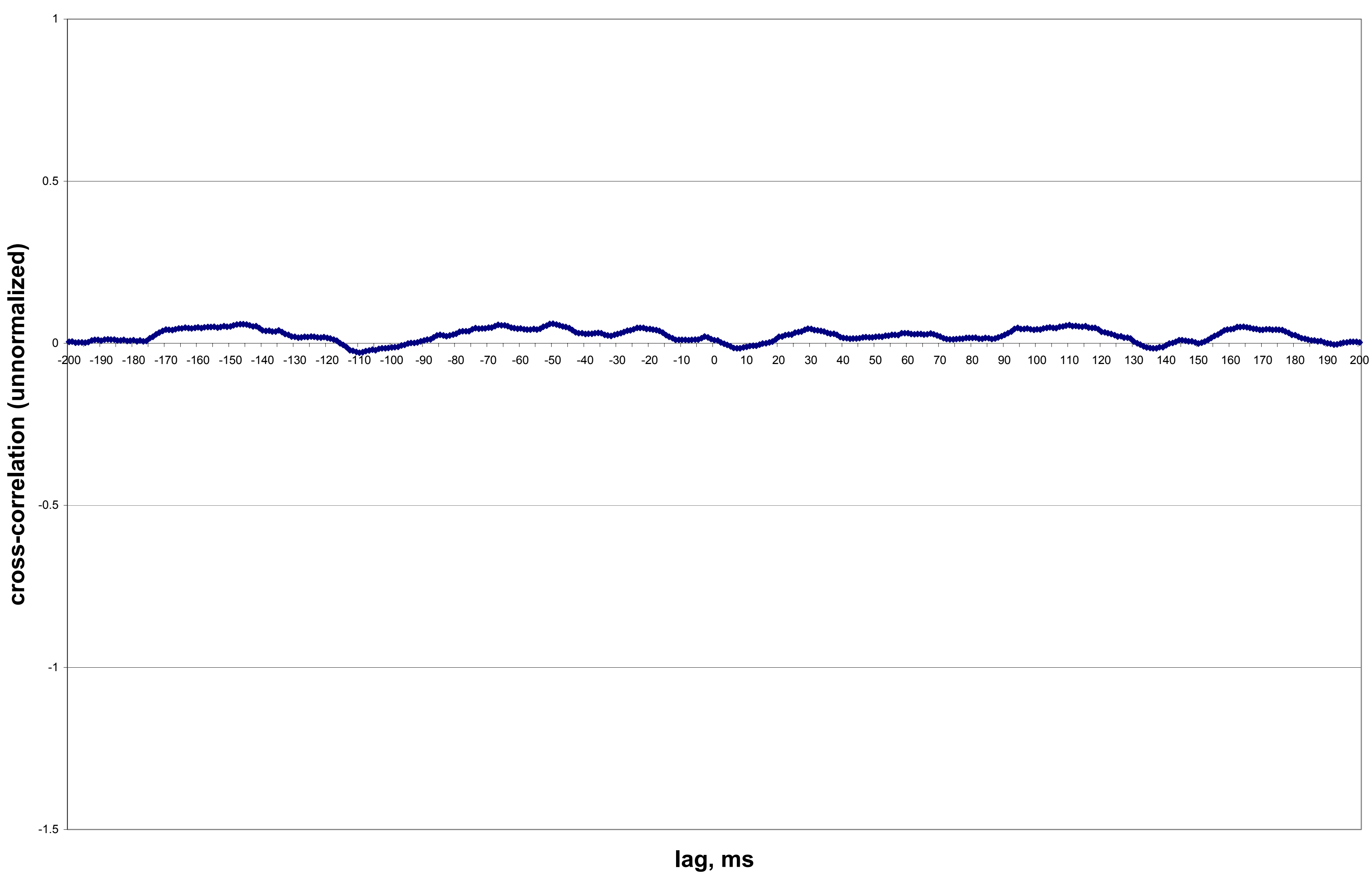


SSA State Functions (27 ms)

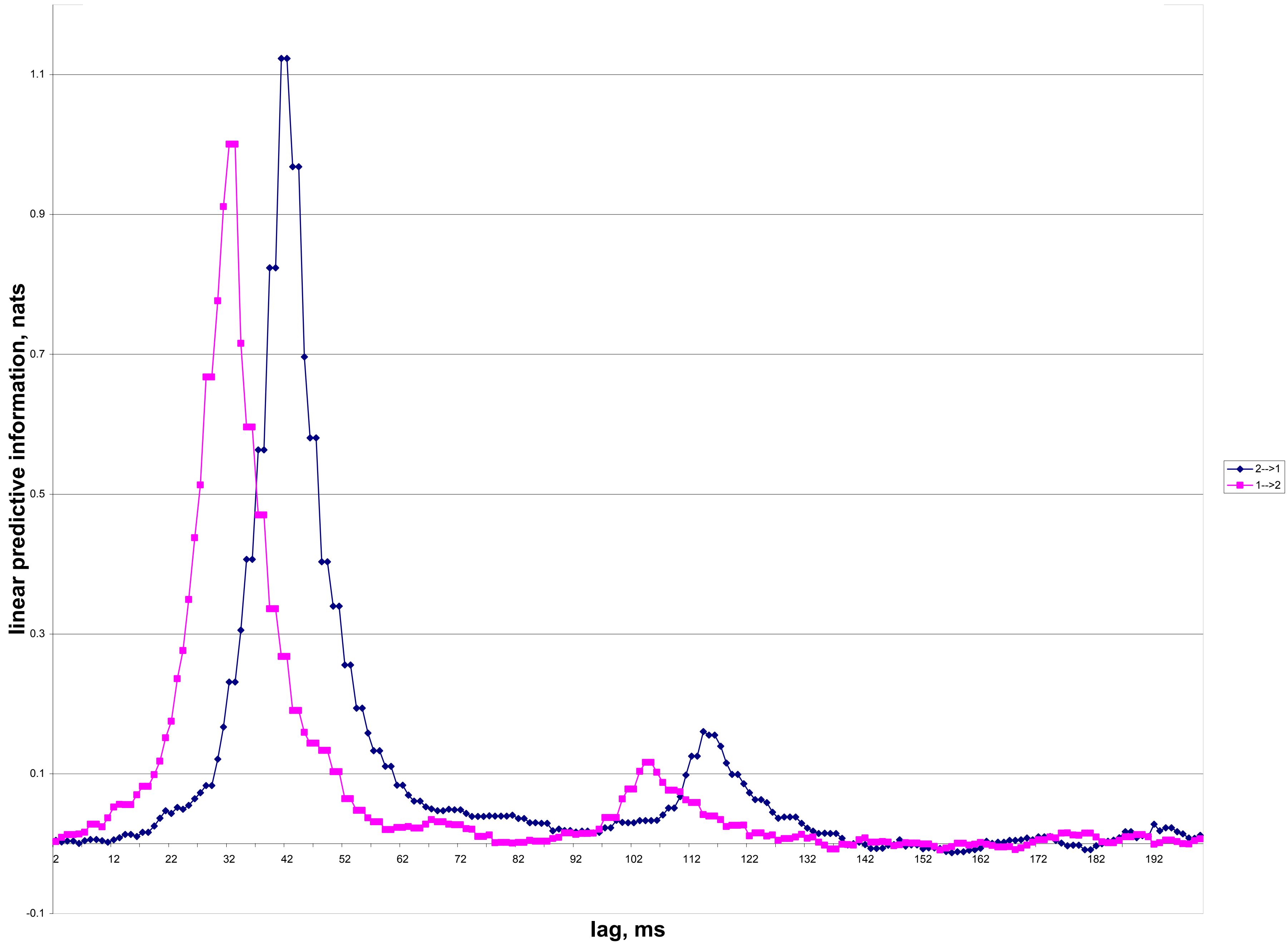
Cross-correlation Function for **Linearly** Coupled Processes



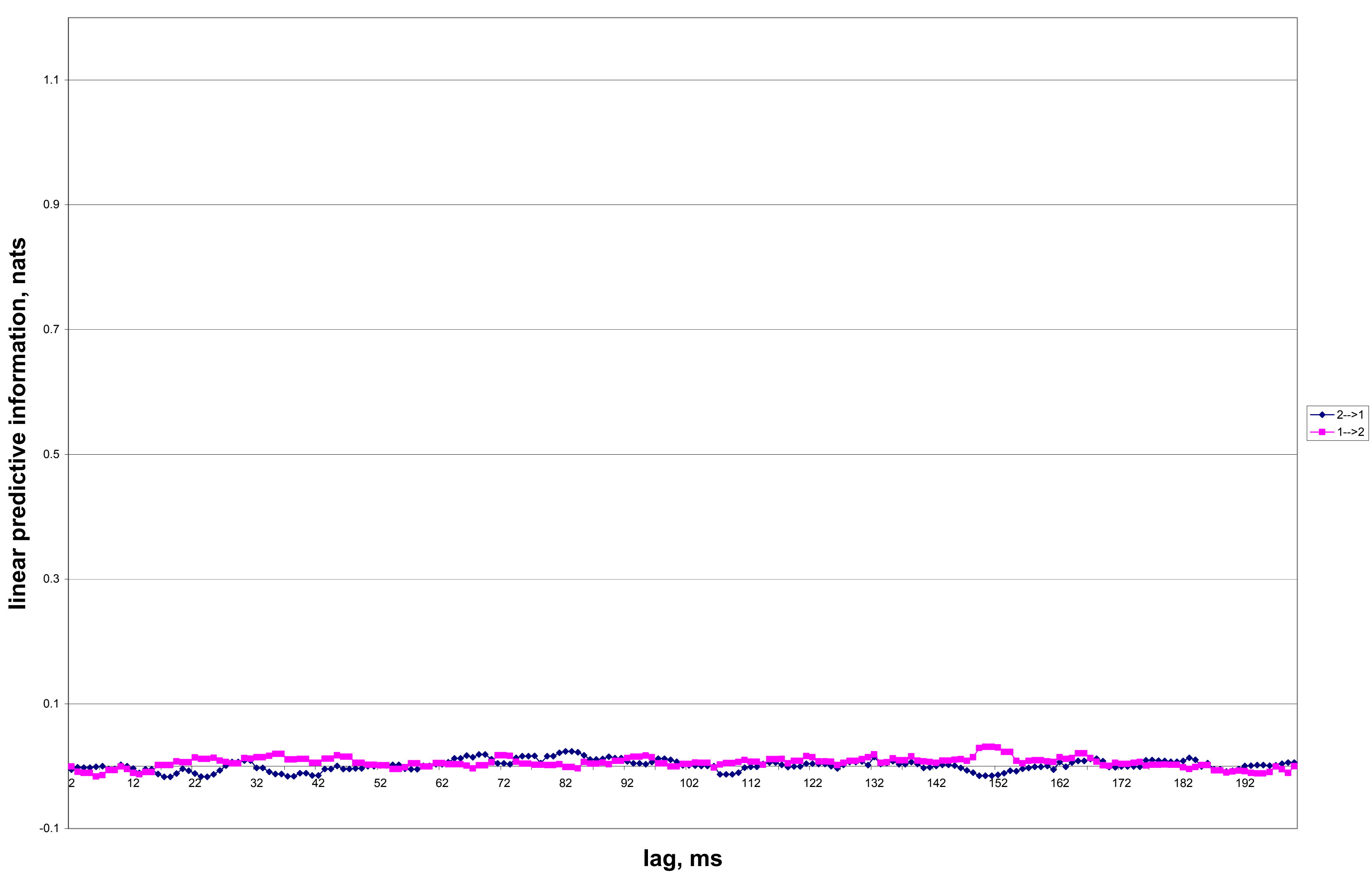
Cross-correlation Function for **Nonlinearly** Coupled Processes



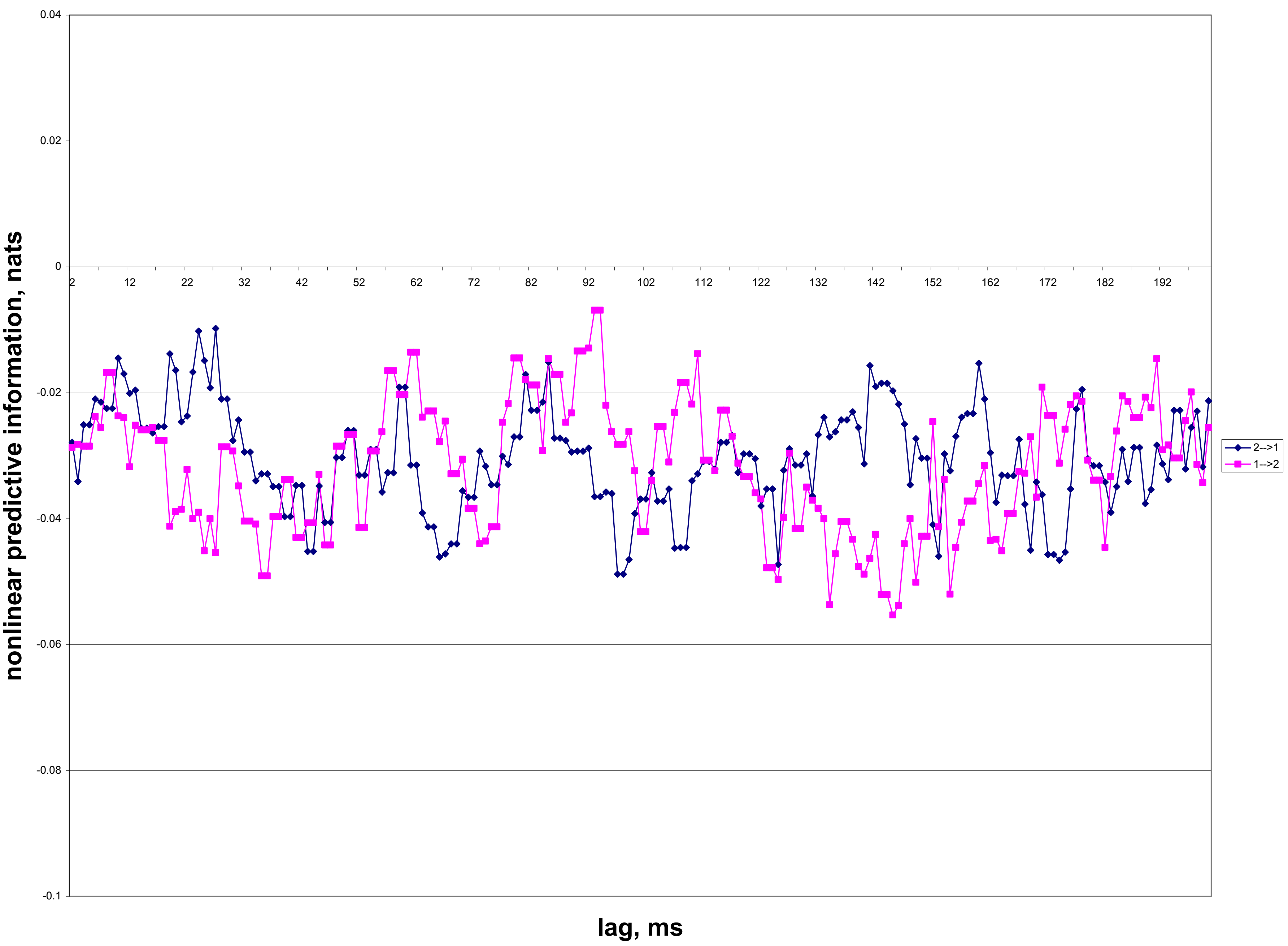
Time-lagged **Linear** Predictive Information for **Linearly** Coupled Processes



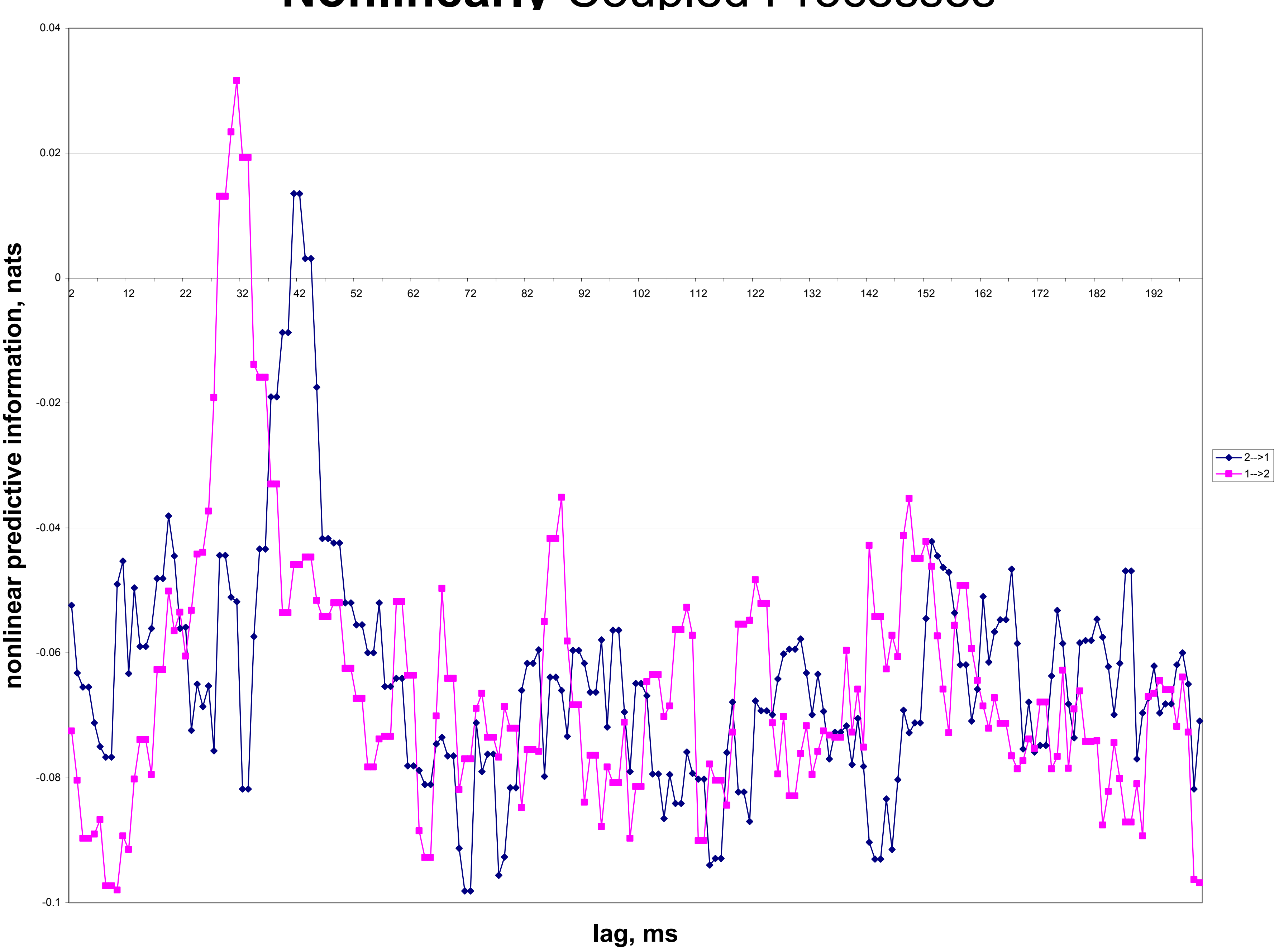
Time-lagged **Linear** Predictive Information for **Nonlinearly** Coupled Processes



Time-lagged **Nonlinear** Predictive Information for **Linearly** Coupled Processes



Time-lagged **Nonlinear** Predictive Information for **Nonlinearly** Coupled Processes



Results

For the *linear coupling process*, both linear methods—cross-correlation and linear predictive information—exhibit prominent peaks at lags 30 ms and 40 ms for 1→2 coupling and 2→1 coupling, respectively. Nonlinear predictive information, on the other hand, has no features that are clearly related to the known dynamics.

By contrast for the *nonlinear coupling process*, the linear measures are relatively flat, and (after rescaling) have no peaks at the known lags. Only the nonlinear predictive information measure is able to detect the 30 ms and 40 ms lags.

Secondary (and tertiary) peaks about 70 ms from the primary peaks are visible in the two linear measures for the linear process. These are "recycled" effects (70 ms being the time required to complete the circuit).

Discussion

The most important result is that the nonlinear predictive information measure is able to extract the lags in spite of no utilization of particular knowledge about the nonlinearity.

Linear behavior is all of one general sort, which can be handled by well-established tools (such as the SVD). Thus, linear information—governed by Gaussian processes—may be extracted first; then all that remains, by definition, is *non*linear.

The leave-one-out entropy estimate that utilizes a Gaussian kernel density estimator requires one key parameter, namely, the scaling parameter (sigma). The entropy estimate depends critically on the choice of this parameter. Our experience so far suggests that a scaling parameter of 1 may be optimal, or nearly optimal, *after sphering the data sample.* Thus, it may turn out to be important to extract all linear information content first, before attempting to extract the nonlinear information content.

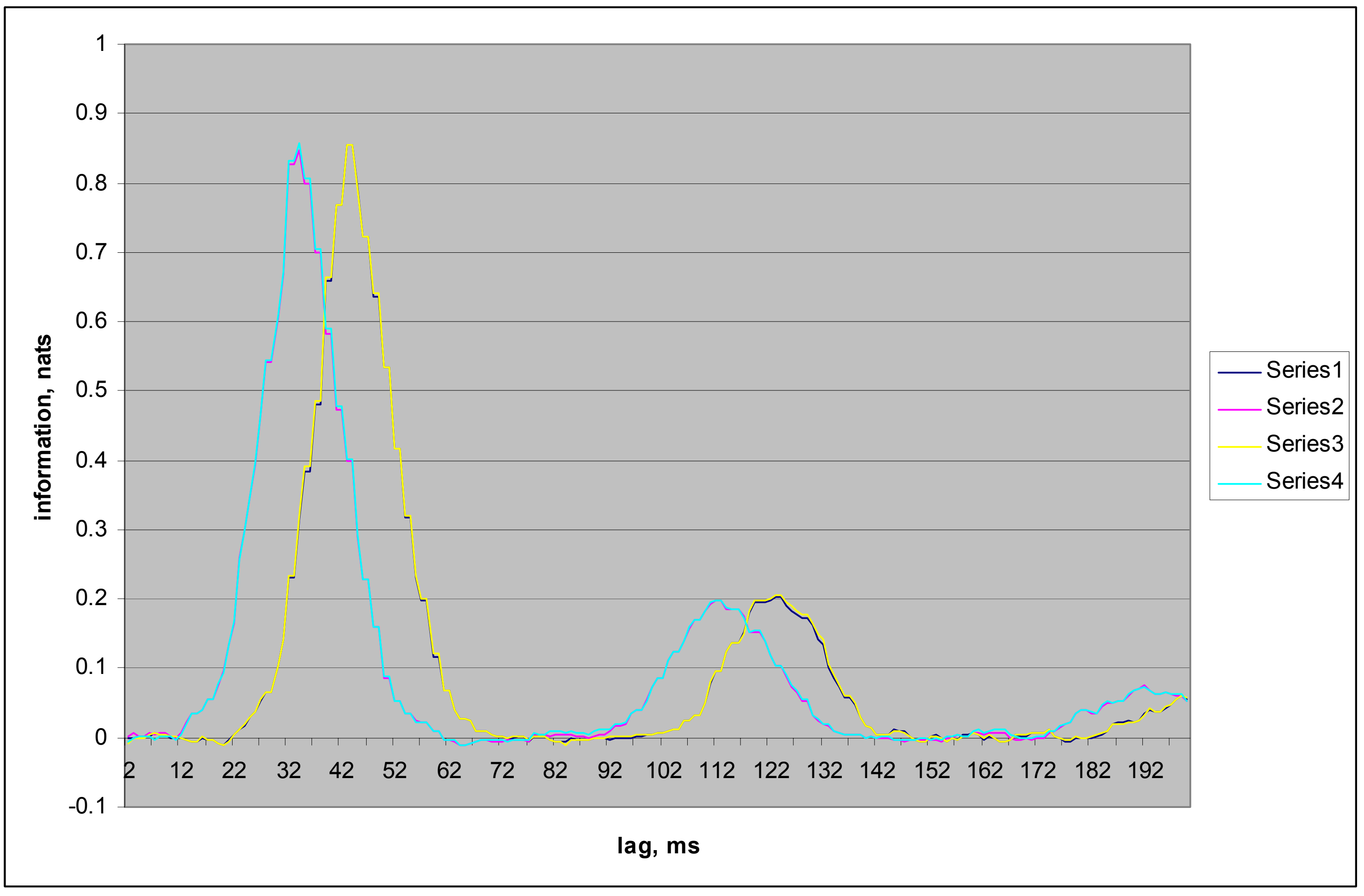
In any case, it is convenient to decompose the information content into linear and nonlinear components. In general, note that many processes will have *both* linear and nonlinear components.

From Predictive Information to Causal Information

As noted earlier, predictive information is necessary, but not sufficient, for causality. For example, a network may consist of three nodes, *A→B→C*, where *A* has a predictive, but not *directly* causative effect on *C*. In this case, if we take *B* into account, *A* provides no further information about *C.* In linear regression models, this may be accomplished by "partialization" (Kalitzin et al., 1997). An analogous operation may be performed using entropies, whether linear or nonlinear.

More subtle examples of the insufficiency of predictive information for causality can occur when there are zero-lag correlations between two processes. For neuroelectric time series, zero-lag correlations could be due to passive volume conduction, or to physiologically interesting synchronization. These motivate a concept of causal information currently under investigation that takes into account the concurrent states of the two processes. That is, *causal information* is defined as the extra amount of information on the state of process *A* contained in the earlier state of process *B* given that we already know the earlier concurrent state of process *A* and the later concurrent state of process *B*.

Causal information in this sense may under some circumstances reduce to predictive information, as shown below for a simple ARMA(2,1) process: causal information series 3 & 4 coincides with the predictive information series 1 & 2.



References

Beirlant J, Dudewicz EJ, Györfi L, van der Meulen EC (1997): Nonparametric entropy estimation: an overview. *Intern. J. Math. Stat. Sci.* 6(1): 17-39.

Bressler SL (1995): Large-scale cortical networks and cognition. *Brain Research Reviews* 20: 288-304.

Bressler SL, Kelso JA (2001): Cortical coordination dynamics and cognition. *Trends in Cognitive Sciences* 5(1): 26-36.

Broomhead DS, King GP (1986): Extracting qualitative dynamics from experimental data. *Physica D* 20: 217-36.

Diks C, DeGoede J (2001): A general nonparametric bootstrap test for Granger causality. In: HW broer, B Krauskopf, G Vegter (Eds.), *Global Analysis of Dynamical Systems* (Institute of Physics Publishing, Bristol, UK), pp. 391-403.

Friston K (1994): Functional and effective connectivity: a synthesis. *Human Brain Mapping* 2: 56-78.

Granger CWJ (1969): Investigating causal relations by econometric methods and cross-spectral methods. *Econometrica* 37: 424-38.

Granger C, Lin JL (1994): Using the mutual information coefficient to identify lags in nonlinear models. *The Journal of Time Series Analysis* 15(4): 371-84.

Granger CW, Maasoumi E, Racine J (Jan 20, 2000): A dependence metric for nonlinear time series. Unpublished manuscript.

Horwitz B, Tagamets M-A, McIntosh AR (1999): Neural modeling, functional brain imaging, and cognition. *Trends in Cognitive Sciences* 3(3): 91-8.

Ivanov AV, Rozhkova (1981): Properties of the statistical estimate of the entropy of a random vector with a probability density. *Problems of Information Transmission* 17: 171-8.

Kalitzin S, van Dijk BW, Spekreijse H, van Leeuwen WA (1997): Coherency and connectivity in oscillating neural networks: linear partialization analysis. *Biological Cybernetics* 76(1): 73-83.

Kaminski M, Ding M, Truccolo WA, Bressler SL (2001): Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics* 85(2): 145-57.

Mesulam MM (1990): Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology* 28: 597-613.

Shannon CE, Weaver W (1949): *The Mathematical Theory of Communication.* Chicago: University of Illinois Press.

Acknowledgement

This work is supported by the NIH/NIBIB (1 R43 EB000614-01).