Chaos and detection

Andrew M. Fraser
Systems Science Ph.D. Program, Portland State University, Portland, Oregon 97207-0751

(Received 1 December 1995)

I report on numerical experiments in which a detector reliably found chaotic signals at signal to noise ratios as low as −15 dB. The detector was based on a variant of the hidden Markov models used in speech research. The task was particularly difficult because the Fourier power spectrum of the noise was constructed to match the spectrum of the signal. I review likelihood ratio detectors, limitations on the performance of linear models implied by the broad Fourier power spectra of chaotic signals, and the upper limit that the *Kolmogorov-Sinai (KS) entropy* of a chaotic system places on the expected log likelihood attainable by any model. I find that KS entropy estimates indicate that even better detection performance is possible. [S1063-651X(96)01705-9]

PACS number(s): 05.45.+b, 84.40.Ua, 02.50.−r

I. INTRODUCTION

The multitude of experiments revealing chaotic physical phenomena that have been reported in the literature of the past decade and a half suggest that chaos is ubiquitous. These measurements have firmly established the notion that erratic time series may be explained by low dimensional deterministic dynamics. Many investigators are now transferring insight gained from studying chaos to work on practical tasks such as forecasting, control, and communication. This paper reports on numerical experiments in which I used nonlinear models to solve a detection problem that linear models could never solve. While much of the chaos and applications literature emphasizes the deterministic aspects of chaotic systems, this paper focuses on probabilistic models and stochastic properties of chaotic systems.

If trajectories of a chaotic system are projected on a coarse grained or discrete observable, determinism is lost. It is impossible to determine the value of a future observation on the basis of past observations. The sequences of measurements constitute a *stochastic processes.* Suppose, for example, that the function F operating on a continuous state space with elements z has a chaotic attractor with a stable asymptotic probability density. Given a discrete partition of the state space $\alpha=\{a_1,a_2,\ldots,a_N\}$, one can map sequences of states $(\ldots,z(-2),z(-1),z(0),z(1),\ldots)$ to sequences of observations $(\ldots,a(-2),a(-1),a(0),a(1),\ldots)$ by assigning $a(t)$ the value a_k when $z(t)\in a_k$. In the original state space one has $z(t+1)=F(z(t))$, but in the space of observations one is left with a stochastic process, i.e., a set of probability functions for sequences of all lengths $\{P_{a_1^t}:t\geqslant 1\}$. [Notation: I use subscripts on probability functions to indicate a function itself rather than the value of a function at a point or when it is not clear from the argument which function I intend. I use a subscript and superscript to denote a sequence, i.e., $a_1^t\equiv(a(1),a(2),\ldots,a(t))$.]

In the theory of signal processing and communication, signal sources are treated as stochastic processes. Thus in filtering, one is interested in $P(x|y)$ the conditional distribution of source signals x that could have caused on observed signal y. In forecasting, one is interested in $P(y(t+\tau|y_1^t))$, the conditional distribution of future values given past values. In a detection problem, the detector is given a measured sequence u_1^T and asked to choose between two hypotheses. (Of the many references on detection, I have used the work of Van Trees [1] and of Fukunaga [2].) Hypothesis H_0 is that no target is present, and hypothesis H_1 is that a target is present. The hypotheses correspond to two different stochastic processes that could have generated the measured sequence. The costs of the four possible outcomes are denoted $C_{0,0}$, the cost of choosing H_0 when H_0 is true; $C_{0,1}$, the cost of choosing H_0 when H_1 is true; $C_{1,0}$, the cost of choosing H_1 when H_0 is true; $C_{1,1}$, the cost of choosing H_1 when H_1 is true. The decision rule that minimizes the expected cost is as follows: Choose H_1 if and only if

$$\frac{P_{H_1}(u_1^T)}{P_{H_0}(u_1^T)} > \frac{(C_{10}-C_{00})P(H_0)}{(C_{01}-C_{11})P(H_1)}, \tag{1.1}$$

where $P(H_0)$ and $P(H_1)$ are the prior probabilities that the target is present or not present, respectively. When measured data are used as the arguments of a probability function, the value of the function is called a *likelihood*. Thus the left-hand side of inequality (1.1) is a ratio of likelihoods and the decision rule is called a *likelihood ratio test.* Using η to denote the right-hand side and taking logs, inequality (1.1) takes the form

$$\sum_{t=1}^{T} \log\frac{P_{H_1}(u_t|u_1^{t-1})}{P_{H_0}(u_t|u_1^{t-1})} > \log\eta. \tag{1.2}$$

This form suggests a recursive evaluation of the log likelihood ratio function.

To build intuition on the use of the likelihood of models for detection, consider Fig. 1. The figure represents numerical data from the double scroll system that is described in Sec. II. Figure 1(a) is a histogram of 5000 samples at a signal to noise ratio of 50 dB, and Fig. 1(b) is a histogram of 5000 noise samples. If the ten test values u_1^{10} depicted in Fig. 1(c) are observed and one must guess whether they came from the source characterized by Fig. 1(a) or the source characterized by Fig. 1(b), it seems more plausible to claim that they are drawn from the latter process. Figures 1(d)-1(f) depict the case when the signal to noise ratio drops to 5 dB. Two-

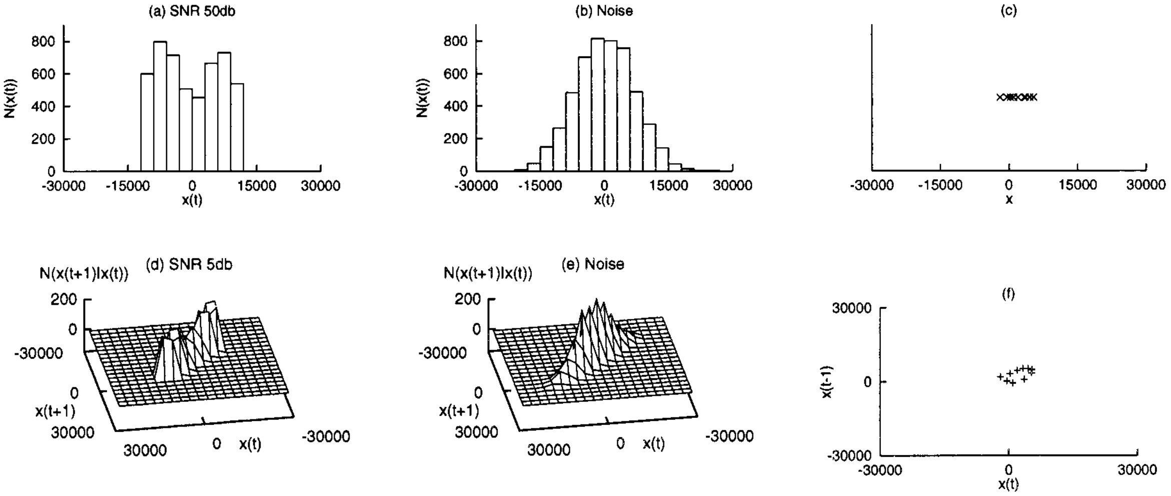

FIG. 1. Histograms for detection. (a) A 20 bin histogram of a 5000 point sample of the target signal at 50 dB SNR. (b) A 20 bin histogram of a 5000 point sample of the background noise. (c) Test sequence of ten observations. (d) A 20×20 bin histogram of a 5000 point sample of the target signal at 5 dB SNR. (e) A 20×20 bin histogram of a 5000 point sample of the background noise. (f) Test sequence of ten observations.

dimensional histograms provide more discriminating characterizations for the more difficult task. As the dimension is increased, the number of cells in a histogram grows exponentially, as does the number of samples required to estimate the probability of falling in any particular cell. In other words, the models have too many free parameters. To reduce the number of parameters, one might fit multivariate Gaussians to the data. (In fact, such models are the basis for most common signal processing techniques.) However, for the data considered here, the best Gaussians that can be fit to the two source processes are *identical by construction* and thus are of no value for detection.

Although a likelihood ratio test is optimal, implementing one requires knowledge of the two likelihood functions. Likelihood functions are difficult to estimate and it is often better to estimate the distribution of a simple function of a measurement $F(u_1^t)$ called a *feature*. I did not use features in the experiments described in this paper; the detectors were built on direct estimates of likelihood functions for entire measurement sequences.

Given a sample sequence $\overline{x_1^T}$ from a stochastic process, one would like to build a model P_θ that could be used to evaluate the likelihood $P_\theta(u_1^t)$ that the same process produced some other sequence. For the experiments reported in this paper, I used *maximum likelihood estimation*, i.e., for a class of models with free parameters θ one selects the parameters that maximize the likelihood of the sample sequence $P_\theta(\overline{x_1^T})$.

I use the expected log likelihood per sample $1/T\langle \log P_\theta(x_1^T)\rangle$ as a figure of merit for models. In Sec. III A I explain that the entropy of a stochastic process gives an upper bound on this figure of merit and the upper bound is only attained when a model gives the right probability for each possible sequence. Log likelihood per sample can be interpreted as bits per sample. Given a model P_θ, an arithmetic code can represent a sequence in less than $\log_2 P_\theta(x_1^T)+2$ bits. Rissanen [3], who invented arithmetic coding, has used this observation to cast estimation as an aspect of coding.

Fitting a complex model to a particular sequence $\overline{x_1^T}$, one often encounters ‘‘overfitting,’’ i.e.,

$$\log P_\theta(\overline{x_1^T}) > \log P_{x_1^T}(\overline{x_1^T}) > \langle \log P_\theta(x_1^T)\rangle.$$

There are several refinements to maximum likelihood estimation that address overfitting, but as Sec. III D suggests, overfitting was not a serious problem in the experiments.

A standard class of models assumes that signals are produced by stable linear systems excited by Gaussian noise. Given the Fourier power spectral density (PSD) of a signal source, one can calculate an upper bound to the expected log likelihood that this approach can obtain. The bound is described in Sec. III B. On the other hand, an estimate of Kolmogorov and Sinai’s (KS) *entropy* of a chaotic source provides a similar bound for *any* approach. That bound is described in Sec. III C. For chaotic sources the difference between these two bounds indicates that the performance of signal processing systems that are based on linear models is much less than optimal.

II. NUMERICAL DATA

I used the routine ODEINT from Press *et al.* (see [4], p. 721) to integrate the double scroll system as described in Chua, Komuro, and Matsumoto [5]:

$$\dot{x}_1 = \alpha(x_2 - h(x_1)),$$

$$\dot{x}_2 = x_1 - x_2 + x_3,$$

$$\dot{x}_3 = -\beta x_2,$$

where $h(y) = m_1 y + \frac{1}{2}(m_0 - m_1)[|y+1| - |y-1|]$, and I used the parameters $\alpha = 9.0$, $\beta = 100/7$, $m_0 = -1/7$, and $m_1 = 2/7$. Figure 2 characterizes the system. For the examples in this paper, I generated a sequence of 10^6 x_1 values sampled at $\tau_s = 0.3$. I multiplied each sample by 5000 and recorded 16 bit integers to simulate digitized measurements and enable meaningful comparisons to the bounds described

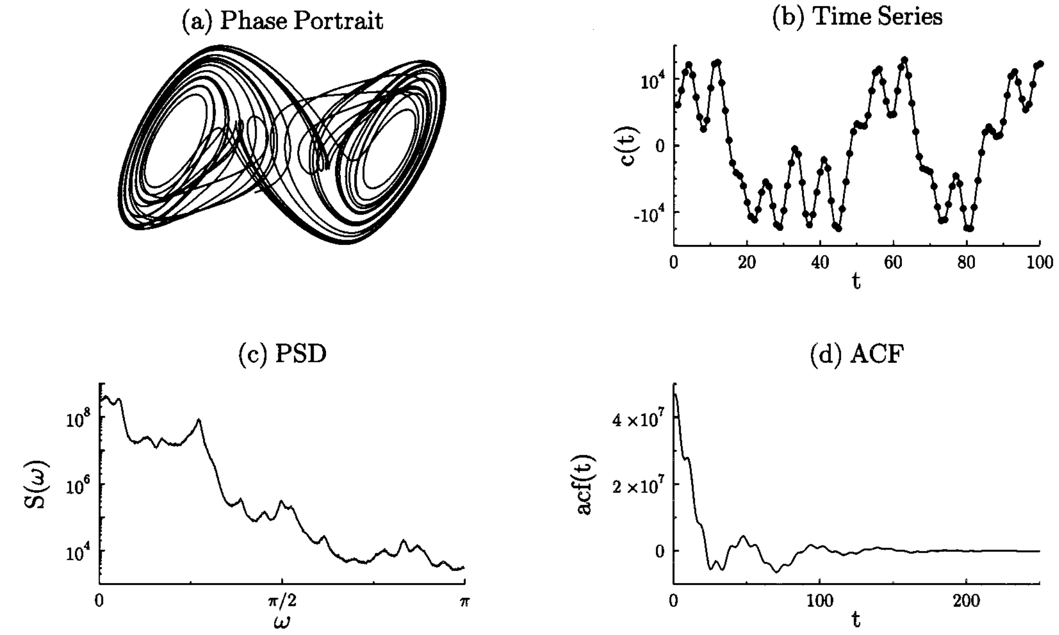

FIG. 2. Strange (or chaotic) attractor. (a) Phase portrait of the double scroll system. (b) Scalar time series of the observable *c(t)*. (c) Fourier power spectrum of the observable. (d) Autocorrelation function.

in Secs. III B and III C. In the remainder of this paper I will refer to subsequences of these data as

$$c_1^T \equiv (c(1), c(2), \ldots, c(T)),$$

changing the name of the measured variable and rescaling to a unit sampling interval for simplicity.

I designed the background noise process to make linear models useless for distinguishing the signal from the noise. I used the Levinson-Durbin algorithm (see [6] for a simple description) to fit a sequence of autoregressive (AR) models [Eq. (2.1) is an AR model] to the data and then used the models to generate the background noise. A smaller model order would have been sufficient to ensure that the difference between the Fourier spectra of the signal source and the noise source would be insignificant, but the calculations are fast, so I used a model order of 200. With the notation

$$b(t) = \sum_{k=1}^{N} b(t-k) a_{N,k} + \sigma_N \epsilon(t), \quad \epsilon(t) \sim \mathcal{N}(0,1) \tag{2.1}$$

for the AR model of order N, the procedure I used to generate sequences of length $T+1$ can be written as

$$b(0) = \sigma_0 \epsilon(0),$$

$$b(1) = b(0) a_{1,1} + \sigma_1 \epsilon(1),$$

$$b(2) = b(1) a_{2,1} + b(0) a_{2,2} + \sigma_2 \epsilon(2),$$

$$\vdots$$

$$b(200) = \sum_{k=1}^{200} b(200-k) a_{200,k} + \sigma_{200} \epsilon(200),$$

$$b(201) = \sum_{k=1}^{200} b(201-k) a_{200,k} + \sigma_{200} \epsilon(201),$$

$$\vdots$$

$$b(T) = \sum_{k=1}^{200} b(T-k) a_{200,k} + \sigma_{200} \epsilon(T),$$

where the noise terms are independently identically distributed (IID) with $\epsilon(t) \sim \mathcal{N}(0,1)$. Using 200 different models to avoid start up transients in the generated noise samples is also overkill. I could have simply discarded the start up transients. But the equations that describe how to generate the noise also describe how to evaluate the likelihood of a measured signal. Thus they are important for building a detector that can work on short measurement sequences.

III. BOUNDS ON LIKELIHOOD

Chaotic time series are useful test cases for nonlinear signal processing techniques because one can estimate bounds on the likelihood that the *best* models could achieve. Thus one can compare the performance of a proposed technique against an absolute reference. The performance bound is given by the KS entropy. A similar bound on the performance of linear models that can be calculated from the Fourier power spectrum is also due to Kolmogorov. I will refer to these bounds as the *KS bound* and the *linear bound.* The KS bound is defined in terms of *discrete sources*, i.e., sources of sequences that take values from a discrete set at each time step. On the other hand, the linear bound is concerned with continuous sources. It describes how well a linear system driven by (IID) Gaussian noise can approximate a source.

A. Entropy and likelihood

Given a source of discretely valued sequences with probabilities that are actually given by P_c, consider models of the source that approximate the probabilities of sequences with parametrized functions P_θ. An essential characterization of the performance of a model is the expected value of the log of the conditional likelihood

$$\lim_{T\to\infty}\langle \log P_\theta(c(T)|c_1^{T-1})\rangle_{P_c}$$

(the subscript on the angular brackets indicates that the expected value is with respect to the true probability).

The Gibbs inequality [Cover and Thomas (see [7], p. 76) call it the *information inequality*] says

$$\langle \log P_\theta(c_1^T)\rangle_{P_c} \leq \langle \log P_c(c_1^T)\rangle_{P_c}$$

and

$$\langle \log P_\theta(c(T)|c_1^{T-1})\rangle_{P_c} \leq \langle \log P_c(c(T)|c_1^{T-1})\rangle_{P_c}$$

with equality only when $P_\theta(c_1^T)=P_c(c_1^T)$ almost everywhere. The *entropy rate* is defined by

$$-H(\mathscr{C}) \equiv \lim_{T\to\infty}\frac{1}{T}\langle \log P_c(c_1^T)\rangle_{P_c} = \lim_{T\to\infty}\langle \log P_c(c(T)|c_1^{T-1})\rangle_{P_c}.$$

(These limits exist if the stochastic process is stationary and has a finite alphabet.) Thus

$$\lim_{T\to\infty}\langle \log P_\theta(c(T)|c_1^{T-1})\rangle_{P_c} \leq -H(\mathscr{C}).$$

In other words, *the average performance of any model is bounded by the entropy rate.*

The McMillan theorem [sometimes called the Shannon-McMillan-Breiman theorem (see any text on ergodic theory or information theory, e.g., [7], p. 474, or [8], p. 131)], which is the linchpin of information theory, says that for an ergodic process

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\log P_c(c(t)|c_1^{t-1}) = -H(\mathscr{C}) \tag{3.1}$$

in probability. By analogy, I conjecture that in probability

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\log P_\theta(c(t)|c_1^{t-1}) = \lim_{T\to\infty}\langle \log P_\theta(c(T)|c_1^{T-1})\rangle_{P_c}. \tag{3.2}$$

If one had a subroutine to evaluate a function $P_\theta(u(t)|u_1^{t-1})$ for which Eq. (3.2) held, then for any δ, $0<\delta<1$, and any $\epsilon>0$, a single string of sufficient length would provide an estimate of $\lim_{t\to\infty}\langle \log P_\theta(c(t)|c_1^{t-1})\rangle_{P_c}$ within ϵ with probability $1-\delta$.

The linear bound is given in terms of the *differential entropy* of a *continuous* source that has the same autocorrelation function as the target. If its probability density is smooth, the differential entropy of a continuous (IID) variable is

$$h(X) \equiv \int -p(x)\log p(x)\,dx.$$

This is a weaker characterization of a random variable or process than the simple discrete entropy because, for a given probability density, the differential entropy can be forced to have any specified value by changing the coordinate system, e.g., if $y=\alpha x$, then

$$h(Y)=h(X)+\log\alpha.$$

If x is quantized with bins of size Δ to yield the discrete variable z^Δ, then

$$h(X)=\lim_{\Delta\to 0}[H(Z^\Delta)+\log\Delta]$$

and the limit is approached as $p(x)$ becomes constant over entire bins. The numerical data were constructed with $\Delta=1.0$ and the data were multiplied by 5000 so that probability densities from a linear process fit to the chaotic data would be almost constant over entire bins. Thus, in the chosen coordinates, the entropy of the discretized linear process closely approximates the differential entropy of the continuous process.

B. Linear models

The canonical models for time series are linear systems driven by (IID) Gaussian noise, i.e., the convolution

$$\mathbf{u}=\mathbf{h}\otimes\boldsymbol{\epsilon},$$

$$u(t)=\sum_{\tau=-\infty}^{\infty}h(\tau)\epsilon(t-\tau).$$

One needs the *impulse response function* $\mathbf{h}$ for applications such as filtering, in which one wishes to extract an unobserved driving signal $\boldsymbol{\epsilon}$ from an observed output $\mathbf{u}$. But for many applications (including detection), the only thing that matters is the set of probability density functions for sequences of all possible lengths. Each of these densities is a multivariate Gaussian and is entirely specified by a covariance matrix C, i.e., for $\mathbf{u}\equiv u_1^T$

$$P(\mathbf{u})=\frac{1}{\sqrt{(2\pi)^T|C|}}e^{(-1/2)\mathbf{u}^\dagger\cdot C^{-1}\cdot\mathbf{u}}.$$

If the process is stationary and has mean zero, the covariance matrix is determined by the autocovariance function R:

$$C_{i,j}=R(i-j)=R(j-i)=\langle u(i)u(j)\rangle.$$

Note that $C=\mathbf{h}\cdot\mathbf{h}^\dagger$, but the covariance C does not uniquely specify the impulse response $\mathbf{h}$.

The covariance matrix C has the Toeplitz form. If the autocovariance $R(t)$ decays to zero quickly enough, the covariance matrix will become *asymptotically equivalent* [9] to a circulant matrix as the length T of the sequences u_1^T considered goes to infinity. The discrete Fourier transform diagonalizes circulant matrices. Hence, for large T, operations involving C, C^{-1}, or $|C|$, e.g., the evaluation of $P(u_1^T)$ or the entropy $H(U_1^T)$, can be well approximated quickly using fast Fourier transforms. As $T\rightarrow\infty$ the principal axes of $P(u_1^T)$ approach Fourier basis functions with eigenvalues given by the Fourier power spectrum. Thus, for large T, Shannon's formula for the differential entropy of a multivariate Gaussian in terms of the eigenvalues λ of the covariance matrix C

$$h(U_1^T)=\frac{T\ \log(2\pi e)+\log|C|}{2}=\frac{T\ \log(2\pi e)+\Sigma_k\log\lambda_k}{2}$$

can be approximated using $S(\omega)$, the Fourier PSD. In the limit $T\rightarrow\infty$ one obtains Kolmogrov's expression for differential entropy rate and mean square prediction error based on infinite history (see [7], p. 274)

$$h(\mathscr{U})=\frac{1}{2}\log 2\pi e+\frac{1}{4\pi}\int_{-\pi}^{+\pi}\log S(\omega)d\omega. \qquad (3.3)$$

This number characterizes the best performance possible using linear models. For the numerical data set, it is 7.74 nats = 11.16 bits, where the unit nat indicates base e for the logs and bit indicates base 2. The interpretation is that using a linear model, the 16 bit samples in the numerical data set could be losslessly compressed to 11.16 bits per sample.

C. KS entropy and Lyapunov exponents

For several decades, ergodic theorists worked to determine if a change of coordinates could transform the function $f(x)=2x \bmod 1$ into the function $g(x)=3x \bmod 1$. In a sequence of papers in 1958 and 1959 Kolmogorov and Sinai used a carefully defined entropy rate that is coordinate independent and has different values for the two systems to prove that no such isomorphism exists. Their *KS entropy* is

$$h_\mu(\phi)\equiv\sup_\alpha\lim_{T\rightarrow\infty}H(A(T)|A_1^{T-1}).$$

Here ϕ is a dynamical system, μ is a measure (probability) that is invariant under ϕ, and α is a partition. The partition reduces trajectories in the underlying space to symbol sequences $\ldots,a(t-1),a(t),a(t+1),\ldots$ by recording which element of the partition is occupied at each time. The conditional entropy for a partition α is

$$H(A(T)|A_1^{T-1})=-\sum_{a_1^T}P(a_1^T)\log P(a(T)|a_1^{T-1}).$$

The ideas are summarized in Sinai's lecture notes [8].

I assume that there is a unique *natural measure* μ for the double scroll system and that it is approximated by long trajectories such as the data I have generated. It is difficult to apply the definition of KS entropy directly, but the Pesin identity

$$h_\mu(\phi)=\sum_{\lambda_k(\lambda_k>0)}\lambda_k \qquad (3.4)$$

relates h_μ to the Lyapunov exponents λ, which in turn can be numerically estimated accurately and easily. (The numerical procedures are easy for low-dimensional systems such as the double scroll, but there are technical questions about the existence of certain limits and the fidelity of numerical simulations of chaotic systems, which I have ignored.) The Pesin identity and the notion of natural measure are reviewed by Eckmann and Ruelle in [10].

I have estimated the KS entropy for the numerical source to be $h_\mu=0.0951$ nats $=0.137$ bits per sample interval. In other words, using an optimal nonlinear model one could losslessly compress the source down from 16 bits per sample to an average of 0.137 bits per sample (a factor of 117).

D. Hidden Markov models

Linear models are not adequate to detect a target signal against background noise with a similar spectrum. For the examples in this paper, I have used what Poritz [11] calls *hidden filter hidden Markov models* (HFHMMs). They are variants of the standard hidden Markov models (HMMs) used in speech research. Although a comparison to the KS bound indicates that HFHMM performance is not even close to ideal for noise free data, they seem to degrade gracefully as signal complexity increases, and it is easy to combine a HFHMM and an AR model that describe signal and noise, respectively, to create a model for the sum of signal and noise.

A HFHMM is concerned with two kinds of random variables at discrete times, an unobserved discrete state $s(t)$ and a continuous observable $u(t)$. The assumptions are (i) given the current state, the next state is conditionally independent of previous states and outputs

$$P(s(t+1)|s_1^t,u_1^t)=P(s(t+1)|s(t));$$

(ii) given the current state and D previous outputs, the current output is conditionally independent of previous states and outputs

$$P(u(t)|s_1^t,u_1^{t-1})=P(u(t)|s(t),u_{t-D}^{t-1});$$

(iii) the output model is linear autoregressive with Gaussian residuals

$$P(u(t)|s(t),u_{t-D}^{t-1})=\frac{1}{\sqrt{2\pi\sigma_{s(t)}^2}}\exp\left(-\frac{[u(t)-\hat{u}]^2}{2\sigma_{s(t)}^2}\right),$$

where $\hat{u}$ depends on the state $s(t)$ and D previous outputs

$$\hat{u}=\overline{u_{s(t)}}+\mathbf{a}_{s(t)}\cdot u_{t-D}^{t-1}\ .$$

Thus the model parameters θ are the discrete conditional transition probabilities $P_{s(t+1)|s(t)}$ and the parameters of the output distribution associated with each state s, i.e., $\overline{u_s}$, σ_s, and the vector of autoregressive coefficients $\mathbf{a}_s$. Given a training sequence u_1^T, one adjusts the model parameters to maximize the likelihood $P_\theta(u_1^T)$.

The computer programs that I used for the present paper are minor modifications of the programs used by Fraser and

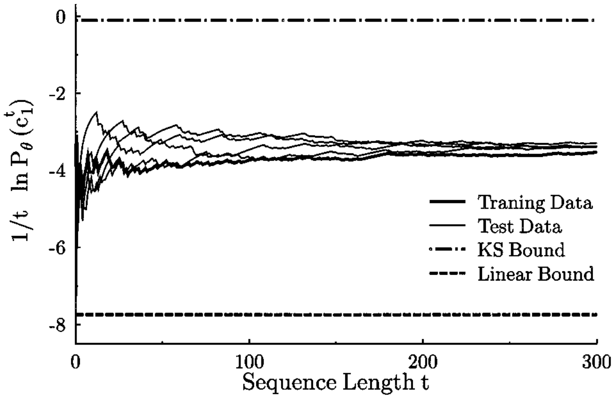

FIG. 3. Average log likelihood. A curve for the training data used to fit the 30 state model with fifth-order AR outputs and curves for four randomly selected realizations of the noise free signal c_1^{300} are plotted.

Dimitriadis [12] for work on time series forecasting. Readers interested in more than this cursory description of the methods should refer to [12].

I was not careful to avoid overfitting the training data, but Fig. 3 suggests overfitting was not a problem. If the model had too many free parameters, the curve for the training data would separate dramatically from the others, indicating severe overfitting.

Figure 4 illustrates the decay with time of the importance of past observations for forecasts of subsequent observations. It is noteworthy that memory needs to be longer when there is more noise. The intuition for this general effect is that there is less information in each measurement, so more measurements are necessary to specify the ''state of the system.''

E. Likelihood summary

Table I contains the KS bound, linear bound, and the actual log likelihood performance of a HFHMM. It also reports the *perplexity* and root mean square error σ that would correspond to the given log likelihood for a Gaussian distribution. In the literature of different fields the likelihood of models are reported in a variety of ways. For comparisons, the following relations between *perplexity* $\mathscr{P}$, *variance* σ^2, and *entropy* h for a Gaussian are helpful:

$$h=\log\sigma+\tfrac{1}{2}\log(2\pi e)\approx\log\sigma+1.419,$$

$$\mathscr{P}=e^h,$$

$$\sigma=\frac{\mathscr{P}}{\sqrt{2\pi e}}\approx\frac{\mathscr{P}}{4.133}.$$

Each model or probability function in Table I could be used for a compression scheme. The last column of the table reports the average number of bits per sample that such a scheme would use. The 4.28 nat separation between the linear bound and the log likelihood per step of the HFHMM is a key component for building a detector. Aspects of detection that are not captured in the table are the variation of the log likelihood over different data sequences and the manner in which the log likelihood gap between the target signal and the linear bound shrinks as the signal to noise ratio is decreased. I touch on these issues in the next section.

IV. DETECTION

Reiterating the introduction, optimal detectors implement likelihood ratio tests. Given a test sequence of T observations u_1^T, a detector must guess whether u_1^T is simply background noise b or a mixture of the target signal and the background noise $c+b$. The ratio of the likelihoods or its log, i.e., $\log P_{c+b}(u_1^T)/P_b(u_1^T)$, summarizes all of the information about the measured signal that is relevant for making the decision. An optimal detector will decide that the target signal is present if $\log P_{c+b}(u_1^T)/P_b(u_1^T)>\log\eta$ where the threshold η is chosen on the basis of the costs of making errors and the *a priori* probability that the target is present.

In practice, true likelihood functions are not available. For the numerical experiments, I used HFHMMs P_θ to approximate $P_{c+b}(u_1^T)$. Consequently, the performance of my detectors was determined by the distribution of

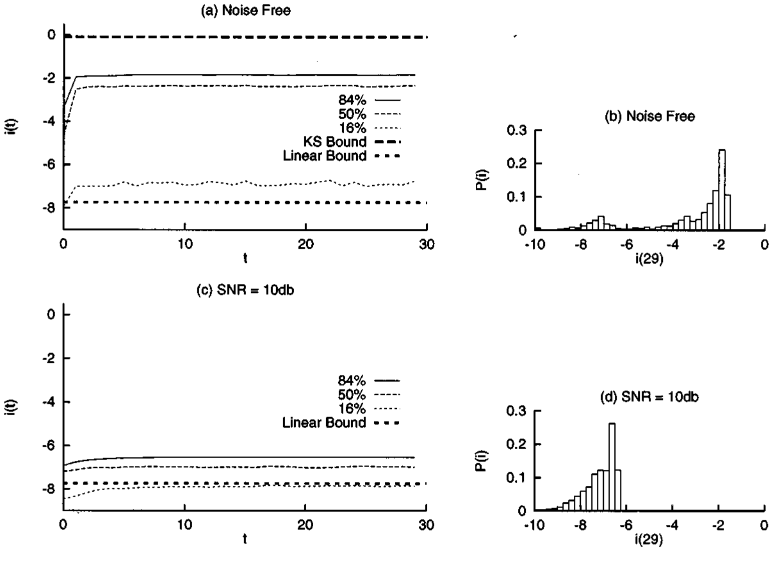

FIG. 4. Incremental log likelihood Each plot represents calculations on 5000 different realizations of u_1^{30}. As labeled, the curves indicate 16%, 50%, and 84% points on the cumulative distribution. Histograms of $i(29)\equiv\log P_\theta(u(30)|u_1^{29})$ are plotted to the left. (a) and (b) Noise free signal. The plots describe the performance of a 30 state model with fifth-order AR outputs. The linear bound and KS bound are derived from Eqs. (3.3) and (3.4), respectively. (c) and (d) Noisy signal at a SNR of 10 dB. The plots describe the performance of a model that was derived from the model in (a) and (b) by using tenth-order AR outputs with parameters appropriate for a SNR of 10 dB. I know of no bound for noisy data that corresponds to the KS bound of the noise free case.

TABLE I. Expected log likelihood per sample interval. The HFHMM was a 30 state model with fifth-order AR outputs.

Model	Perplexity	σ	$-\langle \log[P_\theta(y(t)\vert y_1^{t-1})]\rangle$	
uniform	65536	15858	11.09 nats	16 bits
best linear	2294	555.0	7.74 nats	11.16 bits
HFHMM	31.94	7.73	3.46 nats	5.00 bits
best nonlinear	1.100	0.266	0.0951 nats	0.137 bits

$$q(u_1^T) \equiv \log \frac{P_\theta(u_1^T)}{P_b(u_1^T)}. \tag{4.1}$$

In order for a detector to work reliably, noise signals b_1^T must produce much smaller values of q than signals consisting of the target mixed with noise y_1^T. In other words, for typical realizations b_1^T and

$$y_1^T = (\sqrt{\alpha}) b_1^T + (\sqrt{1-\alpha}) c_1^T,$$

one wants $q(b_1^T) \ll q(y_1^T)$. Note that the signal to noise ratio (SNR) here is $10 \log_{10}[(1-\alpha)/\alpha]$dB. A comparison of the distributions $P_{q(b_1^T)}$ and $P_{q(y_1^T)}$ determines how well a detector will work. Figure 5 plots the dependence of these two distributions on the signal to noise ratio. The intuition that the signal should become undetectable as the SNR decreases is confirmed by the manner in which the distributions become similar as the SNR decreases. The figure represents the performance on signals 500 samples long, the noise model P_b was a tenth-order AR model, and the signal plus noise model P_θ was a HFHMM with 30 states and tenth-order AR outputs.

Consider the choice of the threshold value $\log\eta$. If targets are escaping detection, one can always lower the threshold (at the expense of increasing the false alarm probability). This trade-off is captured in the receiver operating characteristic (ROC), which is a plot of the probability of detection (PD), given that there is a target present against the probability of a false alarm (PF), given that there is no target present. For a perfect detector, some value of the threshold would yield PD equal to 100% and PF equal to 0. On the other hand, PD equal to PF is the characteristic of a worthless detector. The 50%-50% point on the worthless ROC can be implemented by a tossing a fair coin: if the coin comes up heads (tails) declare the target present (declare no target), respectively. Other points on the worthless ROC can be implemented with biased coins. Perfect, worthless, and realistic ROCs appear in Fig. 6.

Figure 7 depicts the decay of the ROCs as the signal to noise ratio decreases.

In many cases the distribution of q will be almost Gaussian. Define the statistic

$$q'(u_1^T) \equiv \log \frac{P_\theta(u(T)\vert u_1^{T-1})}{P_b(u(T)\vert u_1^{T-1})}$$

and note

$$q(u_1^T) = \sum_{t=1}^{T} q'(u_1^t).$$

If the noise and target processes are stationary and have ''short'' memories, then q' will inherit these properties and the central limit theorem says that the distribution of q will become Gaussian as the sequence lengths increase. I have used this Gaussian approximation for all of the plots in this section except Fig. 7(a).

V. CONCLUSION

Probabilistic time series models can be used for a number of applications including forecasting, detection, classification, and compression. The numerical experiments described in this paper illustrate detection. For most applications, the performance of a model is closely related to its expected log likelihood per time step, i.e., $\lim_{T\to\infty}\langle \log P_\theta(y(T)\vert y_1^{T-1})\rangle_{P_y}$. If one uses only canonical linear Gaussian models, the best log likelihood that can be obtained is described in terms of the Fourier power spectral density by Kolmogorov's expression [Eq. (3.3)]. On the other hand, if the signal source is a chaotic system, the log likelihood of even the best model is bound by the KS entropy [Eq. (3.4)]. The HFHMMs in the numerical experiments yielded log likelihoods that fall midway between these bounds (see Fig. 3). Thus they are both substantially better than canonical linear Gaussian models and substantially worse than optimal. The availability of known performance bounds make chaotic time series useful test cases for assessing modeling techniques.

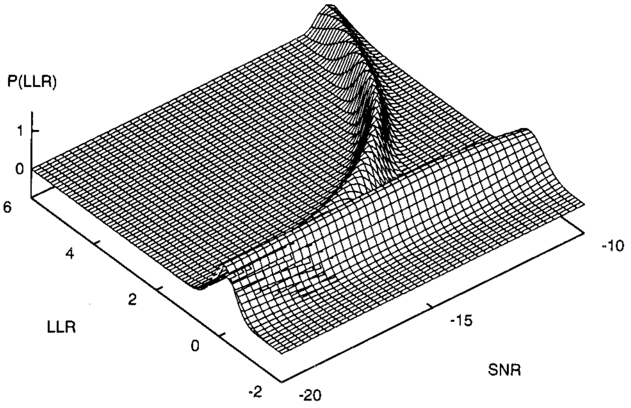

FIG. 5. Distribution of the log likelihood ratio $P(\log P_\theta(u_1^{500})/P_b(u_1^{500}))$ vs the SNR. Plots of two families of probability densities are superimposed: The lower right branch depicts the distribution when u_1^{500} is from the background noise source and the upper left branch depicts the distribution when u_1^{500} is from the target signal added to noise.

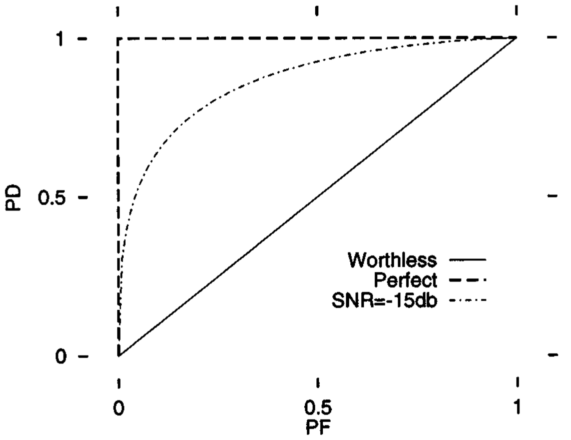

FIG. 6. Three receiver operating characteristics (ROCs): The ''perfect'' characteristic and the ''worthless'' characteristic are hypothetical and the SNR equal to -15 dB characteristic represents the performance of a HFHMM with 30 states and tenth-order outputs applied to sample sequences u_1^{250} with 250 points.

Although the KS entropy provides a bound on the average log likelihood for a model of the noiseless target, i.e.,

$$h_\mu \leqslant \left\langle -\frac{1}{T}\log P_\theta(c_1^T)\right\rangle,$$

detection performance depends on the mean *and the variance* of the log likelihood *ratio* for both the noise and the sum of the target and the noise. Thus more is required than the KS entropy to calculate theoretical bounds on detection performance. I would like to see a theory that gave bounds on the detection of chaotic signals in additive noise and characterized the distribution of likelihood of the best possible models.

One might claim that the techniques described in this paper are of limited relevance because the numerical experiments concerned chaotic signals. Beyond citing the many papers describing chaotic signals in nature, the claim can be refuted by examining the signal characteristics that escape a linear Gaussian model but can be exploited by a HFHMM. Since a global linear model must be stable to be bounded, canonical models are generally forced to be stable. Chaotic systems are nonlinear, locally unstable, deterministic, and bounded. Of these features HFHMMs can reflect local instability and nonlinearity, canonical models reflect boundedness, and neither are deterministic. Thus HFHMMS may be useful whenever nonlinearity or local instability is an important signal characteristic.

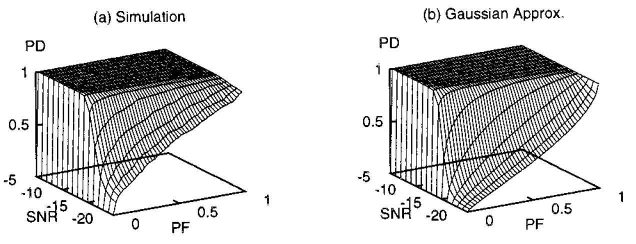

FIG. 7. ROC vs SNR (a) reports results of numerical detection experiments on 500 sequences u_1^{500} at each SNR. (b) is a plot of theoretical ROCs based on estimates of the mean and variance of $q(u_1^{500}) \equiv \log P_\theta(u_1^{500})/P_b(u_1^{500})$.

ACKNOWLEDGMENTS

This paper builds on work with Qin Cai [13,14] and Alexis Dimitriadis [15,12], which was supported by NSF Grant No. MIP-9113460. The idea of using HMMs to detect chaotic signals emerged from conversations with Ron Hughes. I used the Gnu tools from The Free Software Foundation and other free software to do the numerical experiments. (With the exception of the NUMERICAL RECIPES [4] routines, all of the software I used is free and available on the Internet, most of it at http://sunsite.unc.edu/pub/Linux.) This publication was supported by a grant from Intel Corporation.

APPENDIX A: RELATIVE ENTROPY IN TERMS OF PSD

In this appendix, I sketch a derivation of Kolmogorov's relation between the Fourier power spectral density and entropy. In parallel, I obtain the relative entropy rate, which is something like a distance between processes. The key idea is that as one considers covariance matrices in ever higher dimensions, the spectrum of eigenvalues approaches the PSD. I begin with the definition of the *relative entropy* of two probability density functions p and q,

$$D(p||q) \equiv \left\langle \log\frac{p}{q}\right\rangle_p = -\langle \log q\rangle_p - H(p).$$

I assume that p and q are characterized by covariance matrices $C_p = \langle \mathbf{x}\mathbf{x}^\dagger\rangle_p$ and $C_q = \langle \mathbf{x}\mathbf{x}^\dagger\rangle_q$ with

$$q(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^T |C_q|}} e^{-\mathbf{x}^\dagger C_q^{-1}\mathbf{x}/2},$$

where $\mathbf{x} \equiv x_1^T$, and that C_p and C_q are well enough behaved that they are asymptotically equivalent to their circulant approximations [9]. Now

$$\langle \log q\rangle_p = \left\langle \frac{-T}{2}\log 2\pi - \frac{1}{2}\log|C_q| - \frac{\mathbf{x}^\dagger C_q^{-1}\mathbf{x}}{2}\right\rangle_p .$$

Asymptotically, the discrete Fourier transform simultaneously diagonalizes C_q^{-1} and C_p, so asymptotically

$$\langle \mathbf{x}^\dagger C_q^{-1}\mathbf{x}\rangle_p = \sum_{k=0}^{T} \frac{S_p\left(\frac{2\pi k}{T}\right)}{S_q\left(\frac{2\pi k}{T}\right)}$$

and

$$\log|C_q| = \sum_{k=0}^{T} \log S_q\left(\frac{2\pi k}{T}\right),$$

where $S(\omega)$ is the Fourier spectrum. Hence

$$\langle \log q\rangle_p = \frac{-T}{2}\log 2\pi - \frac{1}{2}\sum_{k=0}^{T}\left[\log S_q\left(\frac{2\pi k}{T}\right) + \frac{S_p\left(\frac{2\pi k}{T}\right)}{S_q\left(\frac{2\pi k}{T}\right)}\right]$$

$$\approx \frac{-T}{2}\log 2\pi - \frac{T}{4\pi}\int_0^{2\pi}\log S_q(\omega) + \frac{S_p(\omega)}{S_q(\omega)}d\omega,$$

from which the expressions for differential entropy rate and relative entropy rates follow:

$$h(\mathscr{X}) \equiv \lim_{T\to\infty} -\frac{1}{T}\langle \log p\rangle_p$$

$$= \frac{1}{2}\log 2\pi + \frac{1}{2} + \frac{1}{4\pi}\int_0^{2\pi}\log S_p(\omega)d\omega,$$

$$\lim_{T\to\infty}[D(p_T||q_T) - D(p_{T-1}||q_{T-1})]$$

$$= \frac{1}{4\pi}\int_0^{2\pi}\frac{S_p(\omega)}{S_q(\omega)} - 1 - \log\frac{S_p(\omega)}{S_q(\omega)}d\omega$$

$$= -\frac{1}{2} + \frac{1}{4\pi}\int_0^{2\pi}\left[\frac{S_p(\omega)}{S_q(\omega)} + \log\frac{S_q(\omega)}{S_p(\omega)}\right]d\omega.$$

APPENDIX B: THE SUM OF AN AR PROCESS AND A HFHMM

For the numerical detection experiments, I needed models for the target signal added to the background noise at various signal to noise ratios. Deriving such a model is a special case of fitting a HFHMM to the sum of two independent sources, each of which is described by its own HFHMM. If the first source has M_α states $\{s_{\alpha,1}, s_{\alpha,2}, \ldots, s_{\alpha,M_\alpha}\}$ and the second source has M_β states $\{s_{\beta,1}, s_{\beta,2}, \ldots, s_{\beta,M_\beta}\}$, then the model of the sum will have $M = M_\alpha M_\beta$ states and the transition probabilities for the product states will be given by multiplying the transition probabilities of the component states.

The output model for a product state depends on second-order moments that cannot be derived from the output models of the component states. Suppose that for a particular product state the output model for the first process is

$$y_\alpha(t) = a_{\alpha,0} + \sum_{k=1}^{N} a_{\alpha,k}x_{\alpha,k}(t) + \sigma_\alpha\epsilon(t) = \mathbf{a}_\alpha^\dagger\mathbf{x}_\alpha + \sigma_\alpha\epsilon(t),$$

where N is the order of the output model, $\epsilon(t)$ is (IID) Gaussian with zero mean and unit variance $x_{\alpha,0}=1$ and $x_{\alpha,k} = y_\alpha(t-k)\forall k,\ 1\leqslant k\leqslant N$. Similarly, suppose that the output model for the second process is $y_\beta(t) = \mathbf{a}_\beta^\dagger\mathbf{x}_\beta + \sigma_\beta\epsilon(t)$. Further, suppose that the two processes are summed using weights α and β, i.e.,

$$y(t) = \alpha y_\alpha(t) + \beta y_\beta(t).$$

Values for $\mathbf{a}$ and σ in the equation

$$y(t) = \mathbf{a}^\dagger\mathbf{x} + \sigma\epsilon(t)$$

are required to complete the model for the sum of the processes. $\mathbf{a}$ is determined by minimizing

$$\chi^2 = \langle[(\alpha\mathbf{x}_\alpha + \beta\mathbf{x}_\beta)^\dagger\mathbf{a} - (\alpha y_\alpha + \beta y_\beta)]^2\rangle \tag{B1a}$$

$$= \langle\mathbf{a}^\dagger(\alpha\mathbf{x}_\alpha + \beta\mathbf{x}_\beta)(\alpha\mathbf{x}_\alpha + \beta\mathbf{x}_\beta)^\dagger\mathbf{a} - \mathbf{a}^\dagger(\alpha\mathbf{x}_\alpha + \beta\mathbf{x}_\beta)(\alpha y_\alpha + \beta y_\beta) - (\alpha y_\alpha + \beta y_\beta)^\dagger(\alpha\mathbf{x}_\alpha + \beta\mathbf{x}_\beta)^\dagger\mathbf{a} + (\alpha y_\alpha + \beta y_\beta)^\dagger(\alpha y_\alpha + \beta y_\beta)\rangle. \tag{B1b}$$

Because the α and β processes are independent, $\langle\mathbf{x}_\alpha\mathbf{x}_\beta{}^\dagger\rangle = \langle\mathbf{x}_\alpha\rangle\langle\mathbf{x}_\beta{}^\dagger\rangle$, $\langle y_\alpha\mathbf{x}_\beta\rangle = \langle y_\alpha\rangle\langle\mathbf{x}_\beta\rangle$, and $\langle y_\beta\mathbf{x}_\alpha\rangle = \langle y_\beta\rangle\langle\mathbf{x}_\alpha\rangle$. Differentiating Eq. (B1) yields

$$\frac{1}{2}\frac{\partial\chi^2}{\partial\mathbf{a}} = (\alpha^2\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\dagger\rangle + \alpha\beta\langle\mathbf{x}_\alpha\rangle\langle\mathbf{x}_\beta^\dagger\rangle + \alpha\beta\langle\mathbf{x}_\beta\rangle\langle\mathbf{x}_\alpha^\dagger\rangle + \beta^2\langle\mathbf{x}_\beta\mathbf{x}_\beta^\dagger\rangle)\mathbf{a} - \alpha^2\langle\mathbf{x}_\alpha y_\alpha\rangle - \alpha\beta\langle\mathbf{x}_\alpha\rangle\langle y_\beta\rangle - \alpha\beta\langle\mathbf{x}_\beta\rangle\langle y_\alpha\rangle - \beta^2\langle\mathbf{x}_\beta y_\beta\rangle. \tag{B2}$$

Maximum likelihood estimates are obtained by solving Eq. (B2) for $\partial\chi^2/\partial\mathbf{a} = 0$ to determine the value of $\mathbf{a}$ and setting $\sigma^2 = \chi^2$ as determined by substituting $\mathbf{a}$ into Eq. (B1). The necessary first- and second-order moments should be estimated and saved as part of the process of training the models of the component processes.

For the detection experiments in this paper, the second process had only one state and the output was a zero mean linear AR process that modeled the background noise. Hence $\langle\mathbf{x}_\beta\rangle = 0$, $\langle y_\beta\rangle = 0$,

$$\chi^2 = \langle\mathbf{a}^\dagger(\alpha^2\mathbf{x}_\alpha\mathbf{x}_\alpha^\dagger + \beta^2\mathbf{x}_\beta\mathbf{x}_\beta^\dagger)\mathbf{a} - 2\mathbf{a}^\dagger(\alpha^2\mathbf{x}_\alpha y_\alpha + \beta^2\mathbf{x}_\beta y_\beta) + \alpha^2 y_\alpha^2 + \beta^2 y_\beta^2\rangle,$$

and

$$\frac{1}{2}\frac{\partial\chi^2}{\partial\mathbf{a}} = (\alpha^2\langle\mathbf{x}_\alpha\mathbf{x}_\alpha^\dagger\rangle + \beta^2\langle\mathbf{x}_\beta\mathbf{x}_\beta^\dagger\rangle)\mathbf{a} - \alpha^2\langle\mathbf{x}_\alpha y_\alpha\rangle - \beta^2\langle\mathbf{x}_\beta y_\beta\rangle.$$

[1] H. V. Trees, *Detection, Estimation, and Modulation Theory* (Wiley, New York, 1968).

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, San Diego, 1990).

[3] J. Rissanen, IEEE Trans. Inf. Theory **IT-30**, 629 (1984).

[4] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, Cambridge, 1992).

[5] L. O. Chua, M. Komuro, and T. Matsumoto, IEEE Trans. Circuits Syst. **CAS33**, 1072 (1986).

[6] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression* (Kluwer, Norwell, MA, 1992).

[7] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[8] Y. Sinai, *Introduction to Ergodic Theory* (Princeton University Press, Princeton, 1976).

[9] R. M. Gray, Stanford ISL Report No. 6504-1, 1971 (unpublished), available by anonymous ftp from decaf.stanford.edu in pub/toeplitz.

[10] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).

[11] A. Poritz, in *Proceedings of the IEEE International Conference on Acoustic Speech and Signal Processing,* edited by Benjamin Monderer (IEEE, Piscataway, NJ, 1988).

[12] A. Fraser and A. Dimitriadis, in *Time Series Prediction: Forecasting the Future and Understanding the Past*, edited by A. Weigend and N. Gershenfeld (Addison-Wesley, Reading, MA, 1994), pp. 265–282.

[13] Q. Cai, Master's thesis, Portland State University, 1993 (unpublished).

[14] A. M. Fraser and Q. Cai, in *Proceedings of the IEEE SSAP Workshop, Victoria, British Columbia, 1992,* edited by Dale Jshpak (IEEE, Piscataway, NJ, 1992).

[15] A. Dimitriadis and A. Fraser, IEEE Trans. Circuits Syst. **CAS40**, 683 (1993).