in Non-Gaussian Random Fields
Final Report
ONR N00014-89-J-3152

D. H. Johnson
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005-1892
Jun. 22, 1993

DEPARTMENT OF
ELECTRICAL AND COMPUTER ENGINEERING
HOUSTON, TEXAS
Analysis of Temporal Symmetry in Non-Gaussian Random Fields
Final Report
ONR N00014-89-J-3152

Don H. Johnson
Computer and Information Technology Institute
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77251-1392

June 22, 1993

This document has been approved for public release and sale; its distribution is unlimited.
1. Project Goals

The goal of the proposed work was to determine if the temporal asymmetry of signals could be exploited by signal processing algorithms. We specifically intended to specify the kinds of dependence structures having physical basis (rather than those chosen for the modeler's convenience) and to develop detection or estimation algorithms sensitive to these structures that yielded signal processing gains. Over the grant's three-year duration, lasting from 1 September 1989 until 30 September 1992, a total of $148,248 in ONR funds and $7,700 in Rice University matching funds were expended. Because of this support, these research goals were accomplished, students receiving ONR support graduated and have engineering positions, and a significant volume of technical literature appeared in reviewed journals and conference proceedings.

2. Research Results

Fundamentals of temporal symmetry were outlined in a masters thesis [8]. There and in previous conference papers [1, 10], the fundamentals of temporal symmetry analysis techniques for time series were developed. We uncovered for the signal processing community an important result published by another researcher over ten years earlier: The only linear, temporally symmetric random process was the Gaussian. This result means that all linear, non-Gaussian processes were temporally asymmetric, a property theretofore unexplored by the signal processing community. Linear Markov processes comprised the focal point of our work, and they are generated by passing white noise $W_n$ through a first-order digital filter.

\[ X_n = aX_{n-1} + W_n \]

To illustrate temporal asymmetry, we focused on the hyperbolic secant process, a particular linear non-Gaussian process unmentioned in the literature. Superficially, this example greatly resembles a Gaussian one, but has very different properties. Another example, due to Rosenblatt, proved quite insightful. Here, the linear, first-order, process has a uniform amplitude distribution. Through these examples, the following properties were proven valid [11]:

- The forward conditional expected value $E[X_n | X_{n-1}]$ will be linear for all first-order linear Markov processes. The backward conditional expected value $E[X_{n-1} | X_n]$, however, will be linear only in the Gaussian case. Thus, process linearity can be tested by examination of the forward conditional mean. Furthermore, a sensitive test for Gaussianity is to compare these conditional expected values for linearity.

- The backward mean-square prediction error of a non-Gaussian linear Markov process is always less than the forward prediction error. The Rosenblatt example is particularly striking in this regard: The backward mean-square prediction error is zero while the forward prediction error is nonzero. We have further shown that the time-reversed system, which takes $X_n$ and produces $X_{n-1}$, is deterministic, nonlinear, and chaotic. Thus, one set of ordered numbers can both be produced by a stochastic-driven system and a deterministic, iterated one. From another perspective, a signal viewed looking forward in time is random, while viewed looking backward in time is chaotic. We are now pursuing the research question of what truly distinguishes stochastic from chaotic signals.
- We found that unless the process has a class L distribution, it must be generated by a so-called random coefficient system. The hyperbolic secant density is a member of this class, and therefore has some physical basis. We demonstrated a specific test for class L membership.

- We demonstrated a specific algorithm for generating a linear process having a specified correlation coefficient and amplitude distribution. As one theoretical application of this algorithm, we settled, in the affirmative, the technical question as to whether a linear process could have a multimodal amplitude distribution.

Because of the importance of temporal symmetry, accurate measurement of the conditional mean becomes an essential aspect of non-Gaussian signal processing. We developed a novel nonparametric technique of efficiently estimating the conditional mean [5, 6, 7]. Here, we used the theory of nonparametric regression, and showed that it could be applied to both stochastic and chaotic systems analysis. We developed a technique of identifying the input-output relation of the system that generates a set of observations by operating on a white noise input. Because the technique is nonparametric, it makes few assumptions about the generation system; the algorithm does need to have the system's order. These results have been submitted for publication.

In another line of work, we investigated a technique for determining the order of a Markov linear process that did not depend on the ubiquitous Gaussian assumption. Our algorithm is based on the conditional entropy of the process and has been published [4]. This algorithm applies to nonlinear as well as linear Markov processes. Its sole drawback is computational complexity.

Toward the end of the granting period, a new, potentially important result emerged that is based on the notion of temporal symmetry [2, 3]. We showed that all physically obtained time series must result from time-irreversible random processes. Consequently, models that produce time-reversible processes, in particular the Gaussian, have no physical basis. Stationary Gaussian processes cannot serve as models of physical measurements. This important result is being prepared for formal publication.

We developed a specific algorithm for designing optimal detectors for linear, non-Gaussian, continuous-time processes [9]. Here, specification of the random process is only obtained with difficulty. Calculation of the detector requires detailed analysis of the Poisson random measures that underly the observations. These results have been submitted for publication.

3. Students Supported

Over the project's three-year period, three graduate students were supported by ONR funds. An undergraduate worked on aspects of the project, but was not supported by research funds.

Anand R. Kumar received support for his work on model-order estimation. Graduated with a Ph.D. in 1990 and is now working for Motorola in India.

P. Srinivasa Rao received support for his work in temporal symmetry and in robust detection. He was awarded his doctoral degree in 1992 and is now working at the IBM Watson Research Center in New York.

June 22, 1993
Y. Kang Lee received support for his work in nonparametric system identification. He received his masters degree in 1992 and is now pursuing his doctoral degree, studying the fundamental properties of chaotic and non-Gaussian-stochastic signals.

David D. Becker developed a numerical algorithm for calculating the amplitude distribution of $W_n$ that could produce a first-order Markov linear process having a specified amplitude distribution. This work served as the topic of his Senior Honors Project, and was his first research experience. He went on to receive a masters degree from Stanford in 1991 and now works for General Electric Medical Systems.

4. Infrastructure

To complement ONR's award, Rice University provided funds to purchase a SUN (Sparc 1) workstation for use on the project. This computer was used throughout the granting period and is still used today in non-Gaussian signal processing research.
References

ON THE EXISTENCE OF GAUSSIAN NOISE

Don H. Johnson and P. S. Rao

Computer & Information Technology Institute
Department of Electrical & Computer Engineering
Rice University
Houston, TX 77251-1892

ABSTRACT

The dependence structure and the amplitude distribution of stationary random sequences are linked, with specification of one placing constraints on the other. Time-reversible processes can be Gaussian or non-Gaussian, but all Gaussian processes must be time reversible. We examine the thermodynamics of measurement, showing that while "information" can be extracted from a system without altering system entropy, most measurement techniques irreversibly alter thermodynamic state with a consequent entropy increase. Because of the second law of thermodynamics, such entropy changes cannot be undone and measurements reflecting thermodynamic state cannot be time reversible. We conclude that physical measurements are not time reversible, implying that only non-time-reversible processes model physically relevant signals. Consequently, Gaussian processes would seem to be imprecise representations of physical measurements.

I. INTRODUCTION

The Gaussian process is unquestionably the most prevalent model of both signals and noise in communication, control, and signal processing theory. For many reasons, this process yields analytically tractable results for a wide variety of application problems. Several important signal processing tools based on the Gaussian model are the matched filter, the Kalman filter, and the Wiener filter. Despite its theoretical importance, the fundamental equations of physics impose few constraints on the amplitude distribution of noise and, to the authors' knowledge, the stationary Gaussian stochastic process does not emerge as the solution of any physical problem. The Central Limit Theorem (CLT) stands out as a possible exception to this supposition. However, the convergence of independent superimposed processes to the Gaussian is asymptotic; an infinite number of processes does not exist physically and the CLT cannot be used to justify the Gaussian model on physical grounds.

Recently, researchers have realized the prevalence of demonstrably non-Gaussian noise in physical measurements [5,7] and signal processing research has increasingly turned to developing signal processing algorithms that apply to non-Gaussian noise problems. While the equations governing physical phenomena do not directly constrain the probabilistic amplitude distribution of physical variables, they do constrain the statistical spatio-temporal dependence properties (correlation, for example) of signals we might measure. We usually interpret temporal dependence through the power density spectrum; in this view, only a constant power density spectrum would be free of temporal dependence. For example, temporal correlations are induced on propagating ocean acoustic noise by the filtering characteristics of the medium [13]. A somewhat different dependence property of stochastic signals is the notion of temporal symmetry, where time-reversed sample functions are tested for membership in the original process. As we shall see, a process's temporal symmetry cannot be judged from its spectrum. When viewed from the perspective of familiar Gaussian-based random process properties, this sample-function property may seem subtle since power spectrum measurements cannot determine temporal symmetry. However, the process's temporal symmetry restricts what amplitude distributions the process may have. Because physical laws tend to place constraints on admissible signal's temporal properties, we use these to predict what amplitude distributions physically possible signals may have.

II. TEMPORAL SYMMETRY

A stationary process \( \{X_t, t \in T\} \), is temporally symmetric if for every \( t_1, \ldots, t_n \), all \( n \), the random vectors \( \{X_{t_1}, \ldots, X_{t_n}\} \) and \( \{X_{t_1-n}, \ldots, X_{t_n-n}\} \), \( \forall t \), have the same joint probability distributions [4,14]. Thus, for a temporally symmetric process, time-reversed and delayed sample functions are also sample functions of the original process. With this definition, temporal symmetry is a stationary process property having no gradations: a process is temporally symmetric or it's not. For example, consider a zero-mean, Gaussian process; for all such processes, the covariance function completely characterizes the joint amplitude distribution. As a stationary process's covariance function depends only on the magnitude of the difference

\[ \text{We take "non-Gaussian" to express arbitrary amplitude distributions excluding the Gaussian.} \]
between the two sample times—\( \mathcal{C}(X_t, X_{t'}) = f(|t - t'|) \)—be the process temporally symmetric or not, the covariance function is a temporally symmetric quantity. Consequently, all stationary Gaussian processes are temporally symmetric. Weiss [14] showed that the only discrete-time linear process—stationary time series generated by passing white noise through a linear, time-invariant system—that could be temporally symmetric is the Gaussian. Consequently, all non-Gaussian linear processes must be temporally asymmetric. Nonlinear processes may or may not be temporally symmetric [4]. No basic result akin to Weiss’s for categorizing nonlinear processes is known.

Because the covariance function is by definition a temporally symmetric quantity, a process’s temporal symmetry cannot be examined using second-order statistics, such as the power spectrum. To illustrate this point another way, we can manipulate the temporal symmetry (and the amplitude distribution as well) of a non-Gaussian process by linear operations that have no effect on the power spectrum. Pass a non-Gaussian process through an all-pass filter, which by definition only affects its input’s phase. This phase change modifies the dependence structure of the input, resulting in an output having a different dependence structure and a different amplitude distribution. The power spectra of the filter’s input and output are identical since the power spectrum is insensitive to phase distortions. The Gaussian process’s insensitivity to phase may seem “unphysical”, a notion we are about to argue for. Quantities sensitive to temporal symmetry are the bispectrum [8] and the conditional mean [4, 11].

Other aspects of a signal’s dependence structure are affected by the amplitude distribution. Consider all first-order autoregressive processes parameterized by the pole location \( c \). Given a Gaussian amplitude distribution, any value of \( c \) (consistent with stability criteria) is possible; all first-order dependence structures are compatible with the Gaussian amplitude distribution. The Gaussian is not unique in this regard; for discrete-time signals, amplitude distributions in class \( L \) are compatible with all first-order dependence structures [11]. Among these are the stable distributions, the Laplacian, and the hyperbolic secant [10]. For other amplitude distributions, not all values of \( c \) are compatible. Perhaps the most striking is the uniform amplitude distribution; first-order AR processes exist that have a uniform amplitude distribution, but the parameter \( c \) can only equal \( \pm 1/2, \pm 1/3, \ldots \). For this and other \( m \)-class \( L \) distributions, the amplitude distribution’s form has a direct impact on its dependence structure.

For our purposes, we stress the close coupling between a process’s temporal dependence structure and its amplitude distribution. If we can show that a process cannot be temporally symmetric, we must conclude that it cannot be Gaussian. This constraint allows us to explore the amplitude distribution of processes governed by physical laws by considering temporal dependence structures.

\[ \Delta S \geq 0 \quad \text{and} \quad \Delta W = T \Delta S \]

Modern studies in the thermodynamics of computation have clarified this classic, but ill-defined, concept of entropy. One particularly illuminating definition due to Zurek [15] expresses thermodynamic entropy as the sum of two terms. The first is the Shannon entropy \( H = -\sum p \log p \) of the probability distribution of the system’s state; the second is the algorithmic entropy \( K \) defined as the length of shortest possible description for what is known about the system. The algorithmic entropy might be defined as the logarithm of the shortest possible Turing machine program needed to describe what is known about state. Thus, the first term expresses what any device or person does not “know” about a system’s state—the uncertainty—and the second expresses

\[ \text{We have no attempt to make the units of entropy agree among the various entropy definitions. This paper is more concerned with concepts than details. In the end, each must be multiplied by Boltzmann’s constant and the logarithms have a common natural base.} \]

---

**III. THERMODYNAMICS AND MEASUREMENTS**

Virtually all signal processing algorithms are applied to measurements taken from some physical system (figure 1). We presume that the measurement process is not intended to modify the physical system. Instead, the intent is to capture some time-persistent aspect of the system. We thus expect to modulate the measurement’s “random” components as a stationary process. Non-statistical components are often present too. For example, the physical system could be a communication channel where the signal represents digital data and the random component is additive channel noise. The signal reception process should not, in engineering jargon, “load down” the transmission system, continually changing its characteristics. Under these assumptions on the measurement process, we can justify using stationary stochastic models to describe the noise, enabling us to derive appropriate signal processing procedures.

The effects of measurement on a physical system can be quantified by considering thermodynamics. The key concept is thermodynamic entropy. Loosely speaking, a system’s entropy \( S \) is defined to be \( k \log P \), where \( k \) is Boltzmann’s constant and \( P \) is the number of accessible microscopic states. The Second Law of thermodynamics states that a closed system’s thermodynamic entropy can never decrease and that entropy increases are proportional to the work extracted from the system.

\[ \Delta S \geq 0 \quad \text{and} \quad \Delta W = T \Delta S \]
what is known.

\[ S = H + K \]

Note that this definition is "human-free": a person need not be present to make the measurement; devices such as A/D converters are encompassed by these definitions.

This definition for thermodynamic entropy clarifies discussions held over more than one hundred years about Maxwell's demon [1,2,3,12]. In 1871, Maxwell described a "demon" that, by measuring molecule positions, could enable a machine to do work without increasing entropy. Only by recognizing the role of measurement has the demon been reconciled with the Second Law of thermodynamics. Originally, Szilard in 1929 [12] and Brillouin in 1962 [3] argued that measurement of any physical variable must be accompanied by changes in the information content (proportional to the Shannon information) of the variable. This measurement translates uncertainty in system state into certainty and thereby increases the algorithmic entropy. Using modern terminology, they incorrectly argued that balance between state uncertainty and knowledge could not be maintained \((\Delta H > -\Delta K)\), and they concluded that work must be performed in the measurement process with a concomitant increase in thermodynamic entropy. This work would exactly balance the work seemingly provided by Maxwell's demon and thus uphold the Second Law. However, these researchers did not explore whether a more efficient technique existed for performing the required measurement. Based on his work on reversible computation, Bennett in 1982 [1], showed that balance between \(\Delta H\) and \(\Delta K\) could be maintained theoretically and that measurement did not necessarily increase total entropy. Bennett noted that the demon must return to its original state to initiate another cycle. Thus, a detailed analysis of information transfer explains why Maxwell's demon does not satisfy the Second Law.

In most physical cases, performing a measurement does take work, meaning that the measurement system consumes power and that the thermodynamic entropy of the combined physical and measurement systems increases. Ideally, if sufficient care were taken in the measurement process and the detailed information gleaned was never destroyed, overall system entropy would be constant. Since such ideal circumstances rarely exist, measurement in most, if not all, physical systems is not thermodynamically reversible: once the measurement process has increased thermodynamic entropy, the measurement cannot be undone precisely (algorithmic entropy precisely traded for uncertainty). According to this view, a sequence of measurements, which we express by a scalar-valued time series \(X(t)\), are most often obtained by increasing thermodynamic entropy. These increases do not necessarily mean that the entropy of the physical system being measured has increased. Theoretically, a priori uncertainty can be exchanged for measurement certainty without increasing entropy. Consequently, measurement does necessarily not "lose down" the physical system and the resulting measurements can be well-modeled by a stationary process.

Be that as it may, the measurement process cannot be reversed for two very different reasons. First of all, the work expended in making the measurement cannot be returned by undoing the measurement process: entropy has increased and cannot be decreased to provide the necessary work. Secondly, and most importantly, to the degree that measurements are directly associated with certainty about system state, the time-reversed sequence of measurements cannot be equivalent to a measurement sequence from the physical system. Such temporally reversed measurements would seemingly represent undoing the measurements, thereby corresponding to increasing system uncertainty, while maintaining constant knowledge about state (after all, the measurements are in hand). Signals thus obtained by physical measurements cannot be temporally symmetric. This critical fact obviates any stationary stochastic process model for signals or noise that is temporally symmetric.

For these physical reasons, Gaussian random process descriptions of measured signals would seem to be an abstraction without a physical basis. To recap, all stationary Gaussian stochastic processes are temporally reversible; processes modeling measurements cannot be because of the Second Law. Thus, non-Gaussian processes provide the only viable model for physical measurements. However, temporally symmetric non-Gaussian processes are also inappropriate; because of Weiss's theorem, such processes must arise from nonlinear models. Temporally symmetric non-Gaussian processes that describe physical measurements can be produced by both linear and nonlinear models [11]. Our interpretation of thermodynamics has not produced further restrictions on possible random process models for measurements.

IV. DISCUSSION

Because of the Second Law of thermodynamics, measurements convey the conversion from information-theoretic uncertainty to algorithmic (measurement) certainty. Temporally reversing the time series could not represent the same measurement process as the reversed time series would suggest a physically impossible situation: the continual entropy transformation from its algorithmic to its uncertain form without a net entropy increase. Based on these physical restrictions, the most accurate stochastic process models for data are temporally asymmetric ones.

The use of Gaussian processes in signal processing would thus appear to rest on weak ground, justifying consideration of alternate, non-Gaussian signal processing strategies. The structure and properties of non-Gaussian stochastic processes need to be understood before physically relevant subsets of this class can be selected. Once the process class that accurately models measurement has been defined, the signal processor would naturally seek signal processing algorithms that could exploit the structure imposed by the measurement and best glean the information contained in
Unfortunately, optimal non-Gaussian signal processing operations are not equivalent to the Gaussian ones. Detection theory provides one example. The matched filter applies only to Gaussian noise problems; optimal detection for non-Gaussian noise demands alternative structures. Furthermore, the analytic simplicity provided by the Gaussian process rarely carries over to non-Gaussian problems. Few statistical signal processing algorithms have been developed that are tailored to the amplitude distribution as well as to the temporal dependence structure.

We could apply Gaussian-based algorithms to non-Gaussian problems. Taking another example from detection theory, one could use a matched filter (linear) detector for a non-Gaussian noise problem. However, this filter's unit-sample response is not proportional to the signal as it is for Gaussian situations [9]. Furthermore, the performance for the optimal linear detector can greatly surpass that designed for the Gaussian problem. How much the optimal linear detector degrades system performance when compared to the optimal one is not known. We need to specify how to vary Gaussian-based strategies for non-Gaussian problems and to quantify the losses incurred when Gaussian-based systems are used in physical situations instead of those keyed to physically accurate non-Gaussian models.

References

NONPARAMETRIC PREDICTION OF NON-GAUSSIAN TIME SERIES

Y. Kang Lee
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77251-1892

Don H. Johnson

ABSTRACT

In this paper, we apply the nonparametric kernel predictor to the time-series prediction problem. Because nonparametric prediction makes few assumptions about the underlying time series, it is useful when modeling uncertainties are pervasive, as such as when the time series is non-Gaussian. We show that the nonparametric kernel predictor is asymptotically optimal for bounded, mixing time series. Numerical experiments are also performed: For the nonlinear autoregressor: Based on the most recent \( n \) observations, nonparametrically optimal predictor makes only modest assumptions, making it amenable when modeling uncertainties are pervasive.

1. INTRODUCTION

Time-series prediction is a problem frequently encountered in many branches of science and engineering. In this paper, we would like to predict future values of a time series based on its present and previous observations; consider, for example, linear prediction. In practice, this prediction process consists of two steps: estimation of the predictor from all available observations, followed by prediction of a future time-series value by evaluating the estimated predictor using present observations. Classically, the estimation of the predictor has been simplified by the assumption of a parametric model of the time series, so that the optimal predictor can also be described parametrically. As a result of this simplification, the predictor estimation process is greatly reduced to the task of estimating only a finite number of parameters. In the familiar case of linear prediction, we assume the signal, e.g., speech, to be linear auto-regressive (AR). The corresponding predictor is then formed by estimating, often very efficiently, e.g., via Levinson's algorithm, the coefficients of the linear model.

Nonparametric prediction provides an alternative to the classical methods of linear and higher-order prediction when parametric specifications for the time series are either unavailable or dubious. This approach only assumes that the model describing the time series is smooth. Whereas a parametric fit is global and spans all of state space, a nonparametric estimator fits data locally by taking advantage of the smoothness condition. The class of possible relationships in nonparametric estimation is equivalent to the class of smooth functions, which is clearly too large to be parameterized. In a sense, we may contrast parametric and nonparametric estimation by the assumptions they make.

The parametric method requires quantitative specifications as opposed to the qualitative assumptions made in nonparametric estimation. Effectively, nonparametric prediction makes only modest assumptions, making it amenable when modeling uncertainties are pervasive.

2. NONPARAMETRIC PREDICTION

Based on observations \( X_0, X_1, \ldots, X_N \) of stationary time series \( \{X_n\} \), we want to estimate \( X_{N+1} \) using a \( p \)th order predictor. Based on the most recent \( p \) observations, nonparametrically summarized by \( X_N = [X_N, \ldots, X_{N-p+1}] \), estimate \( X_{N+1} \). That is, the predictor of \( X_{N+1} \) can be written as

\[
\hat{X}_{N+1} = r(X_N),
\]

where \( r(\cdot) \) is some function that maps \( \mathbb{R}^p \) to \( \mathbb{R} \). This function is usually unknown and must be estimated from data as well. This paper does not address the order determination problem and assumes that predictor order \( p \) is known. The interested reader can refer to [1] for an order selection technique based on nonparametric regression.

If the mean square error (MSE) criterion is used, the optimal \( p \)th order predictor of \( X_{N+1} \) is the conditional expected value of \( X_{N+1} \) given \( X_N, \ldots, X_{N-p+1} \):

\[
\hat{X}_{N+1} = E[X_{N+1}|X_N],
\]

The conditional expectation above is a random variable measurable with respect to the \( \sigma \)-algebra of \( \{X_N, \ldots, X_{N-p+1}\} \) and can therefore be expressed as

\[
E[X_{N+1}|X_N] = r(X_N),
\]

where \( r(\cdot) \) is a function that maps from \( \mathbb{R}^p \) to \( \mathbb{R} \). We call \( r(\cdot) \) the conditional mean function.

2.1. Kernel Predictor

In practical situations, the conditional mean function is unknown and must be estimated from data. In this estimation process, we search for a mapping that best describes the causal relationship between random vector \( X_n \) and random variable \( X_{N+1} \) based on their observations \( X_0, X_{p+1}, \ldots, X_{N-p+1} \). We have abbreviated vector \( \{X_0, \ldots, X_{N-p+1}\} \) as \( X_n \). We shall refer to each observation vector \( X_n \) as a predictor vector and each scalar observation \( X_{N+1} \) as the response corresponding to predictor vector \( X_n \).

A parametric method such as linear prediction assumes that \( r(\cdot) \) is linear, thus constraining the estimate to be linear. The nonparametric method does not constrain the form of the estimate. By taking advantage of the smoothness of \( r(\cdot) \), the nonparametric kernel regression estimator, the nonparametric estimator of the conditional mean function at a point \( x \in \mathbb{R}^p \) consists of locally averaging the
responses \( X_{n+1} \) corresponding to those predictor vectors within a neighborhood of \( x \):

\[
i(x) = \frac{\sum_{n=p}^{N-1} K \left( \frac{X - X_n}{h_N} \right) X_{n+1}}{\sum_{n=p}^{N-1} K \left( \frac{X - X_n}{h_N} \right)}
\]  

(1)

In the numerator of (1), kernel \( K() \), along with bandwidth \( h_N \), play the role of a weighting function and assigns a weight to each response \( X_{n+1} \) based on the distance between predictor vector \( X_n \) and \( x \). \( K() \) is generally positive and decreases from the origin; one example is the multivariate Gaussian function. Hence, responses corresponding to predictor vectors close to \( x \) are weighted more highly and have more effect than those with predictor vectors farther. The bandwidth parameter \( h_N \) has the role of controlling the extent of the local average about \( x \). A large bandwidth allows more responses (corresponding to predictor vectors around \( x \)) to be averaged, and a small bandwidth has the opposite effect. Choosing an appropriate bandwidth is crucial for a good estimate; this is discussed in the next section. As a general rule, the bandwidth will decrease as \( N \) increases because local volumes will be filled more densely when more data become available. The denominator is (1) simply serves to normalize the weighting of the responses.

As \( x \) varies within its domain, the kernel estimator can be viewed as a moving average in predictor-vector space, as opposed to the usual notion of a moving average in time. Figure 1 shows the scatter plot \( (X_{n+1} \text{ versus } X_n) \) of \( N = 100 \) samples of the first-order nonlinear autoregressive \( (NAR(1)) \) time series

\[
X_{n+1} = 2\text{sin}(X_n) + W_{n+1}; \quad W_n \sim \text{i.i.d. } N(0,1)
\]

The true conditional expectation \( i(x) = 2\text{sin}(x) \) is shown by the dashed curve. Also shown is the kernel regression estimator \( i(x) \) inside the interval \([-3, 3, 0] \); it is computed using a Gaussian kernel with a bandwidth of \( h_{\text{opt}} = 0.23 \). At \( x = -2.0 \), we show how the responses are windowed to produce the corresponding estimate, as indicated by the asterisk. The kernel estimator represents a natural and intuitive estimator of the conditional mean function because the conditional mean is nothing but the local average of the responses within a (infinitesimally) small neighborhood of the conditioned predictor vector.

Completing the prediction process, we evaluate the kernel regression estimator \( i(x) \) at the present predictor vector \( x = X_n \). We can then write the kernel predictor of \( X_{n+1} \) as

\[
\hat{X}_{n+1} = i(X_n).
\]

(2)

2.2. Bandwidth Selection

Selecting an appropriate bandwidth is crucial for good estimates. If the bandwidth is too large, oversmoothing occurs. Likewise, if the bandwidth is too small, the resulting estimate is under smoothed and jaggled. See Figure 2.

To be consistent with the nonparametric nature of the kernel predictor, the selection of the bandwidth should be based on information inherent in the data and not on a priori, possibly inaccurate, assumptions. Such a data-driven technique, called cross validation, has gained wide-range support among statisticians and time-series analysts. Define the "leave-one-out" regression estimator \( \hat{i}_{n,\hat{k}}(x) \) as follows:

\[
\hat{i}_{n,\hat{k}}(x) = \frac{\sum_{n=p}^{N-1} K \left( \frac{X - X_n}{h} \right) X_{n+1}}{\sum_{n=p}^{N-1} K \left( \frac{X - X_n}{h} \right)}
\]

The appropriately named leave-one-out estimator \( \hat{i}_{n,\hat{k}}(x) \) is the kernel regression estimator of (1) computed without using the pair \( (X_n, X_{n+1}) \). The cross validation function \( CV() \) is the sample prediction error using the leave-one-out estimator as the predictor:

\[
CV(\hat{k}) = \frac{1}{N} \sum_{p=1}^{N} (\hat{i}_{n,\hat{k}}(X_n) - X_{n+1})^2
\]

The bandwidth \( \hat{k} \) that minimizes \( CV() \) is selected and used to compute the regression estimator \( i() \). This minimization is performed over all possible \( k \) values, and the constraint set for \( k \) may require some subjectivity in order to reduce the amount of computations. If \( i() \) were used instead of \( \hat{i}_{n,\hat{k}}() \) to compute \( CV(\hat{k}) \), it can be easily shown...
that $CV(h)$ is minimized at $h = 0$, yielding a useless solution. The use of $r_N(.)$ effectively thwarts this singularity. It is important to note that $h$ is derived completely from data and is consistent with the nonparametric nature of our prediction method.

3. CONSISTENCY RESULTS

As we have already mentioned, it is difficult to specify joint distributions of non-Gaussian time series. To analyze properties of estimators, however, certain specifications on the dependence structure of these time series must be made. We therefore select a very general specification, called a mixing condition [2], for the time series we analyze.

Theoretically, the two estimators that we have so far discussed are different. The kernel regression estimator $\hat{r}_N(.)$ is an (pointwise) estimator of the conditional expectation function $r(x)$, whereas the kernel predictor $\hat{r}(X_N)$ is the estimator of the predictor $r(X_N)$. We need to elucidate the distinction at this juncture because consistency of one estimator does not imply consistency of the other. The difference here is akin to the difference between pointwise and norm convergence of a sequence of functions. Because our impetus is the time-series prediction problem, we should concentrate on the analysis of the kernel predictor. Pointwise consistency of the kernel regression estimator for $\phi$-mixing time series (and others) has been shown [3]. See also [4] for consistency results of the nearest-neighbor regression estimator.

Next, we need to determine an appropriate mode of statistical convergence for the kernel predictor. Almost sure consistency can be found in [3]. However, this does not imply that the kernel predictor asymptotically matches the performance of the conditional mean. For this reason, we believe that it is inappropriate to analyze the a.s. or in probability convergence of the kernel predictor. Instead, it would be more appropriate to analyze its $L^2$ convergence. Lee [5] has shown that for bounded $\phi$, $\rho$, and $\alpha$-mixing time series, the kernel predictor is asymptotically optimal in the sense that

$$E[(\hat{r}(X_N) - r(X_N))^2] \rightarrow 0.$$ \hspace{1cm} (2)

The rate at which this convergence occurs depends on the rates at which the mixing coefficients and bandwidth converge to zero. For example, for $\phi$-exponentially $\phi$-mixing time series (i.e., its mixing coefficient $\phi$ is proportional to $a^n$, $a < 1$), the kernel predictor converges at a rate of

$$E[(\hat{r}(X_N) - r(X_N))^2] = O(N^{-1} \log^2 N/(NA^2)) \hspace{1cm} (3)$$

Thus, consistency occurs if $h_N \rightarrow 0$ and $(NH^2)/\log^2 N \rightarrow \infty$. Using (3), the convergence rate can be optimized by taking the bandwidth to be

$$h_N, opt \propto (N/\log^2 N)^{-1/(r+2)} \hspace{1cm} (4)$$

Convergence is faster for time series that have weaker dependence structures (fast converging mixing coefficients). If the dependence is too strong (mixing coefficients converging too slowly to zero), the kernel predictor is not insured to converge. The most rapid convergence occurs in the trivial situation when time series samples are completely independent.

It is important to note that the consistency result above specifies only the bandwidth rates that are admissible. For example, if $h_N$ satisfies (4), so does $H_N$, for some non-zero constant $H$. Clearly, the two sequences will yield different errors, but the rate at which the errors converge will be the same. Thus, bandwidth values that work in theory may not be obtainable in practice. Cross validation can provide somewhat of a bridge between theory and practice. In the i.i.d. setting (the usual setting for regression analysis), cross validation is asymptotically optimal for the kernel estimator, but the rate is slow [6]. Unfortunately, little is known in the time-series setting about cross validation for either the kernel estimator or the kernel predictor. This remains an open area for research.

4. EXAMPLES

**Nonlinear Autoregressive Process**

Consider the following second-order nonlinear autoregressive (NAR(2)) time series.

$$X_{n+1} = 0.9 + 4(X_{n-1} - X_n) + W_{n+1}$$

with $W_n \sim i.i.d. N(0, 1)$. This time series is a generalization of the well known linear autoregressive process. Unfortunately, the nonlinearity makes $X_n$ very difficult to examine analytically. For example, we do not know its (scalar) amplitude distribution, let alone its joint distributions. In fact, we do not even know if it is stationary! Because $(1 + 4(z_1 - z_0))/(1 + (z_1 - z_0)^2)$ is bounded for all $(z_1, z_0) \in \mathbb{R}^2$, $X_n$ has finite moments. The scalar value 0.9 is used to normalize its variance to approximately 3.0. Its observed mean value is zero. See Fig. 3 for a snapshot of 200 sample values. Even though the stationarity of $X_n$ is questionable, its conditional mean is invariant to time index $n$. Because $W_n$ is independent to $X_m$, $m \leq n$, the conditional mean function is simply $r(z_0, z_1) = 0.9(1 + 4(z_1 - z_0))/(1 + (z_1 - z_0)^2)$ for all $n$.

We test the kernel predictor at sample sizes of $N = 200, 500, 1000, 2000$, and 3000. For each sample size, cross validation is performed to select the appropriate bandwidth. The kernel regression estimate at $N = 500$ is shown in Fig. 4; the linear estimate is shown for comparison sake. A Gaussian kernel with a bandwidth of $\hat{h} = 0.40$ is used. The kernel predictor is then tested against the next 1000 samples. It is evident that the linear predictor cannot adequately capture the nonlinear relationship and, consequently, performs poorly when compared to the kernel predictor. At $N = 500$, the kernel predictor has a MSE of about 1.1, compared with 1.9 for the linear predictor and 1.0 for the optimum predictor, or about 95% of optimal (because $X_n$ has a variance of 3.0).

**Chaotic Time Series**

The kernel predictor can be applied to time series produced by deterministic difference equation. It is likely to perform
Figure 4. Performance Analysis for NAR(2) Time Series

well with these time series because no noise is present. A special case of such a time series is one that is chaotic. Consider the Henon time series (Fig. 5):

\[ X_{n+1} = 1 - 1.4X_n^2 + 0.3X_{n-1} \]

with initial conditions \( X_{-1} = X_0 = 0.0 \).

Figure 5. Snapshot of Henon Time Series

The same prediction strategy as before is followed here. A scatter plot of \( X_n \), the true Henon map, and regression estimate at \( N = 500 \) (viewed from two perspectives) are shown in Fig. 6. Prediction errors are negligible, from \( N = 50 \) on, and therefore not shown. Because the kernel estimator performs local averaging of data, estimates are made only at those locations where data aggregates. Although little inference can be made for the map at all locations, the resulting estimate is effective for prediction purposes.

5. CONCLUSIONS

Nonparametric time-series prediction provides an alternative to the classical methods of linear and higher-order prediction. Because it makes only modest, qualitative assumptions, nonparametric prediction may be applicable even when little is known about the time series under study. Such situations arise when the time series is neither linear nor Gaussian. In light of present-day emphasis on non-Gaussian signal processing, it would seem fitting to incorporate nonparametric prediction into the analyst's toolbox.

Open questions for further investigation remain. The most notable is a way to mitigate the effect of the "curse

Figure 6. Henon Time Series: Scatter Plot, Henon Map, and Regression Estimates

of dimensionality," so called because nonparametric methods require a large amount of data for dimensions higher than about three \( (p > 3) \). This shortcoming needs to be overcome before problems like target tracking and speech modeling can reap the benefits provided by nonparametric prediction.

In the case of chaotic time series that are produced by nonlinear iterative equations, nonparametric prediction performs well because no randomness is present in the responses. Because little averaging is needed, dimensionality effects are not as severe as for stochastic time series. Nonparametric prediction seems to have much potential in this area [1, 7, 8].

REFERENCES

Generation and Analysis of Non-Gaussian Markov Time Series

P. Srinivasa Rao, Student Member, IEEE, Don H. Johnson, Fellow, IEEE, and David D. Becker, Member, IEEE

Abstract—Correlated non-Gaussian Markov sequences can be considered as filtered white noise (independent, identically distributed sequences of random variables), the filter being a nonlinear system in general. We discuss the applicability of linear models and nonlinear methods based on the diagonal series expansion of bivariate densities for analyzing this system. Non-Gaussian sequences exhibit different properties in the forward and backward directions of time. We explore the connection to system modeling of this temporal asymmetry and some of its consequences. As an example, we analyze a first-order linear autoregressive model with hyperbolic secant amplitude distribution at its output.

I. INTRODUCTION

The signals and noise encountered in the signal processing environment (e.g., ocean acoustic noise [19]) are often non-Gaussian. Be that as it may, many signal processing algorithms are based on the assumption that the signal, or noise, or both are Gaussian. Even when the appropriate non-Gaussian amplitude distribution is used, the samples are assumed to be independent or at least uncorrelated. The performance of algorithms which ignore the non-Gaussian nature of the input and/or the dependence structure is seriously limited when the algorithms are inappropriately applied. A common way of accounting for the dependency of non-Gaussian data is to model the process as a pointwise transformation of a correlated Gaussian process. Although this method facilitates easy generation of dependent processes, it yields an analytically complex dependency structure which is insufficient to describe the possible range of dependencies [18]. Development of new algorithms which take into account the non-Gaussianity and correlation structure requires an in-depth study of the properties of non-Gaussian time series and how they can be modeled and generated.

The correlation function is inadequate in capturing the dependency structure of a non-Gaussian time series; only the multivariate Gaussian density depends solely on the covariance matrix. Another reason for this inadequacy is the intriguing fact that non-Gaussian processes often exhibit different properties in the forward and backward directions of time quite unlike Gaussian processes. In the sequel, we shall refer to such processes as being temporally asymmetric.1 The inherent symmetry in the definition of correlation function makes it insensitive to temporal asymmetry and reduces its ability to capture the dependence structure of a non-Gaussian process. What aspects of non-Gaussian sequences are then important in specifying their properties? How can they be exploited in signal processing algorithms? A key issue in developing analysis tools that can be used to answer these questions is how to model the generation of non-Gaussian signals. That is, given a specification of a non-Gaussian signal, how can it be produced by a possibly nonlinear system operating on an elementary random sequence? Gaussian signals can be generated by passing independent, identically distributed (i.i.d.) Gaussian time series through the appropriate linear system. For non-Gaussian signals, are nonlinear systems necessary? If so when?

Before attempting to answer these questions, we make two simplifying assumptions. First, we assume that the signals are (strict-sense) stationary. Second, we assume that the signals are Markovian: the generating systems of such signals are characterized by a small number of "states." Making use of the relatively new theoretical notion of temporal symmetry, we will discuss the suitability of linear models for non-Gaussian processes and then propose a technique for developing nonlinear models for a class of these processes.

II. TEMPORAL SYMMETRY

All stationary Gaussian processes are symmetric with respect to the (discrete) time axis: a time-reversed sample function \(X_n\) of a Gaussian process is also a sample function of the same process and is thus statistically indistinguishable from it. Non-Gaussian processes do not necessarily exhibit this symmetry. For example, sunspot number data collected since the year 1750 have been noted to fall more rapidly than they rise [3]. Neural discharge patterns have also been found to be asymmetric with respect to time [14].

Definition: A stationary process \(\{X_n, n = 0, \pm 1, \cdots\}\) is temporally symmetric if the random vectors \(\{X_1, X_{-1}\}\) is temporally symmetric in the literature. See for example [32], [17].
\( \cdots X_m \) and \( \{X_{-m}, X_{-2}, \cdots X_{-1}\} \) have the same joint distribution for all \( k \) and \( n_i, 1 \leq i \leq k \) [32].

Temporal symmetry of Gaussian processes follows from the fact that the joint distribution of the amplitudes of a stationary, zero-mean Gaussian process is completely specified by the covariance function, which is by definition symmetric for all processes, Gaussian or not: \( E[X_{n_1}X_{n_2}] = E[X_{n_2}X_{n_1}] = E[X_{n_2}X_{n_2}] \) follows from stationarity. Weiss [32] showed that all autoregressive backward and "backward" conditional means are not in-

\[ E[X_{-1}X_{-1}] = E[X_{-1}X_{-1}] \]

E[\( X_{-1}X_{-1} \)] are not symmetric with respect to their arguments and do not specify the covariance function, which is by definition expectations \( E[X_{-1}X_{-1}] \).

Higher order spectra have been shown to be suited for parametric linear system (ARMA) characterization [24].

The statistics having possibly more utility are the conditional expectations \( E[X_1 | X_{-1}] \) and \( E[X_{-1} | X_1] \), as they can be used even when the generating system is nonlinear. These quantities, which we shall refer to as the "forward" and "backward" conditional means, are not inherently symmetric with respect to their arguments and do provide information about the dependency structure of the sequence: if these two statistics are different, the sequence cannot be temporally symmetric.

Consider, for example, the joint distribution

\[ p_{x_1,x_2}(x, y) = \frac{1}{2} \delta(x - y + \frac{1}{2}) + \frac{1}{2} \delta(x - y - \frac{1}{2}), \]

\[-\frac{1}{2} \leq x, y \leq \frac{1}{2}.\]

Clearly, this joint density is asymmetric while its marginals are uniform over \( [-\frac{1}{2}, \frac{1}{2}] \). Its conditional means are calculated as the means of its two conditional probability densities \( p_{x_1 | x_2} (x | y) \) and \( p_{x_2 | x_1} (y | x) \). As the marginal is uniform, the conditional densities have the same functional form as the joint distribution. Respectively, the forward and backward conditional means are found to be

\[ E[X_1 | X_{-1}] = \frac{1}{2} X_{-1}; \]

\[ E[X_{-1} | X_1] = 2X_1 \mod \frac{1}{2}. \]

The forward mean is affine while the backward mean is discontinuous; this difference clearly demonstrates the temporal asymmetry of the associated process if it exists.

The other possibility, namely the identicalness of the conditional means is, however, insufficient to prove the temporal symmetry of the process. After developing more insight into the structure of non-Gaussian Markov processes, we shall return to the properties and applications of conditional means.

### III. Non-Gaussian Markov Processes

A random process \( \{X_n\} \) is \( m \)-th order Markov if its conditional probability densities have the property

\[ p_{x_1|x_{-1},x_{-2}, \cdots}(x_1 | y_1, y_2, \cdots) = p_{x_1|x_{-1},(x_{-2}, \cdots)}(x_1 | y_1, \cdots, y_m) \] \hspace{1cm} (1)

for all \( n \). If \( \{X_n\} \) is also stationary, these conditional densities do not depend on \( n \). The process \( \{X_n\} \) is said to be completely specified if all joint densities of amplitudes at different instants are known. It follows easily that stationary Markov processes are completely specified by the conditional density function given above. If the process is first-order Markov, \( m \) conditional densities can be obtained from the "transitional density" \( p_{x_1|x_{-1}}(\cdot | \cdot) \) by

\[ \begin{align*}
E[X_{n+1} | X_n] &= \int p_{x_{n+1}|x_n}(x_{n+1} | x_n) \, dx_{n+1} \\
&= \int p_{x_{n+1}|x_{n-1},(x_{n}, \cdots)}(x_{n+1} | x_n, \cdots) \, dx_{n+1} \\
&= \int p_{x_{n+1}|x_{n-1}}(x_{n+1} | x_n) \, dx_{n+1}
\end{align*} \]
using the Chapman-Kolmogorov equation [7, p. 89]

\[ p_{x_m|x_{m-1}}(x|y) = \int p_{x_m|x_{m-1}}(x|z)p_{x_{m-1}|x_m}(z|y) \, dz, \quad m \geq 2. \]  

(2)

Thus the transitional density or equivalently the bivariate density \( p_{x_m,x_{m-1}}(x, y) \) completely specifies the first-order Markov process.

The definition of Markov processes given in (1) is one-sided and gives the impression that a Markov process has an inherent direction of time, namely, past amplitude values specifying the statistical properties of the present. One might conclude that a time-reversed Markov process is no longer Markov. However, an equivalent definition can be given in terms of the conditional independence of the past and future, given the present. From this symmetric definition, it follows that if \( \{X_n\} \) is Markov, so is \( \{X_{-n}\} \) [7, p. 83]; this result can also be obtained directly from the one-sided definition above [22, p. 386]. However, the two Markov processes, the original and its time reversed version, may have different characteristics and hence be temporally asymmetric. In the case of a first-order Markov process where \( p_{x_n|x_{-1}}(x, y) \) is a symmetric function of \( x \) and \( y \), it follows from Chapman-Kolmogorov equation that \( p_{x_n,x_{-1}}(x, y) \) is also symmetric for all \( m \geq 2 \) and as a result, it is easily seen that the process is temporally symmetric. Hence a first-order Markov process is temporally symmetric if and only if \( p_{x_n,x_{-1}}(x, y) \) is symmetric.

If the Markov process \( \{X_n\} \) is Gaussian, it can be generated by an all-pole, \( M \)-th order linear system described by

\[ X_n = \alpha_1 X_{n-1} + \alpha_2 X_{n-2} + \cdots + \alpha_M X_{n-M} + W_n \]  

(3)

where \( \{W_n\} \) is an i.i.d. Gaussian sequence. Processes generated in this fashion are often referred to as being autoregressive (AR). Autoregressive models are quite commonly used in diverse areas such as geophysics and speech processing [20].

Non-Gaussian Markov sequences may or may not be generated by linear autoregressive systems, but they can be considered as a generalization of AR sequences, which are known to have a simple statistical structure. To obtain the generation model of non-Gaussian Markov sequences, we must begin with the conditional density function. Suppose \( \{U_n\} \) is the output obtained by passing a \( M \)-th order Markov process \( \{X_n\} \) through a system having the input-output relationship given by the conditional distribution function (cumulative)

\[ U_n = P_{X_n|x_{n-1},\ldots,x_{-M}}(X_n|x_{n-1}, \ldots, x_{-M}). \]  

(4)

\( \{U_n\} \) is then i.i.d. and uniformly distributed over \([0, 1]\) [22, p. 181]. Thus, this system is the equivalent of a "whitening" filter for Gaussian time series and yields the innovations sequence \( \{U_n\} \) corresponding to \( \{X_n\} \). Typically, this system is nonlinear with a finite number \( M \) of states. Being a distribution function, the above input-output relationship is monotonic and hence can be inverted to give the generating system of a general \( M \)-th order Markov time series:

\[ X_n = P_{X_n|x_{n-1},\ldots,x_{-M}}^{-1}(U_n|x_{n-1}, \ldots, x_{-M}). \]  

(5)

This generation model is shown in Fig. 1. In the Gaussian case, this generating system takes the form of a memoryless nonlinearity, which transforms the i.i.d. uniform sequence to an i.i.d. Gaussian sequence, followed by an all-pole linear system. In the general case, the memoryless nonlinearity is usually present, but is followed in general by a nonlinear system having memory. In either case, specification of the conditional distribution function leads to the system that generates the process.

A. Non-Gaussian Autoregressive Processes

Very often, inversion of the conditional distribution function is extremely difficult in practice. Sometimes the conditional distribution function itself may not have a closed form. To proceed further, we first explore linear models: from the above description, we assume that the memoryless nonlinearity transforms the i.i.d. uniform sequence into some intermediate non-Gaussian (i.i.d.) sequence which is then passed through a linear filter. We are thus led to AR(\( M \)) models for non-Gaussian sequences. Validity of a linear model can be verified in practice and we illustrate this here. From now on, we will focus attention on first-order Markov processes (\( M = 1 \)).

A stationary AR(1) process \( \{X_n\} \) is defined by

\[ X_n = \rho X_{n-1} + W_n \]  

(6)

where \( \{W_n\} \) is a zero-mean sequence of i.i.d. random variables and \( |\rho| < 1 \). The system constant \( \rho \) is also the normalized correlation coefficient of the output process \( \{X_n\} \). One main issue of concern at this point is what first-order Markov non-Gaussian sequences can be characterized this way (i.e., are linear processes)? We next discuss this issue via i) the characteristics of the forward and backward conditional means (and their relevance to the directionality of the process) and ii) the variety of amplitude distributions for \( \{X_n\} \). We will have more to say about the conditional means and their use in selecting a linear versus nonlinear model for non-Gaussian data in sequel.

1) Conditional Means and Directionality: In the case of linear AR(1) systems, the forward conditional mean is a linear function, the slope of which is the system coefficient:

\[ E[X_n|X_{n-1}] = \rho X_{n-1}. \]

The backward conditional mean depends heavily on the amplitude distribution of the input \( \{W_n\} \). If the input is Gaussian, the backward conditional mean is same as the forward: both are linear. As an example of the non-Gauss-
The dissimilarity of these quantities reveals the temporal asymmetry of the time series. Although a general expression for the backward conditional mean of non-Gaussian AR(1) processes cannot be obtained, Lawrance [17] showed that it is always nonlinear. As this result is useful for us later, we give the proof here.

**Theorem 1:** The backward conditional mean $E[X_{n-1} | X_n]$ of an AR(1) process is linear only in the Gaussian case [17].

**Proof:** Making use of stationarity and the independence of $X_{n-1}$ and $W_n$, we find that

$$
\Phi_W(u) = \frac{\Phi_X(u)}{\Phi_X(pu)} \tag{7}
$$

where $\Phi_X(u) = E[e^{juX}]$ is the characteristic function of the random variable $X$. From (6) and (7), the joint characteristic function of $X_n$ and $X_{n-1}$ is

$$
\Phi_{X_nX_{n-1}}(u, v) = E[\exp \{juX_n + jvX_{n-1}\}] = \Phi_X(pu + v) \Phi_X(u)/\Phi_X(pu). \tag{8}
$$

Differentiating with respect to $v$ and setting $v = 0$, we find that

$$
jE[X_{n-1} e^{juX_n}] = \Phi_X(pu) \Phi_X(u)/\Phi_X(pu).\tag{9}
$$

Using the properties of conditional expectation, the left-hand side could be rewritten as $|E[e^{JuX_n} E[X_{n-1} | X_n]]$. If the backward conditional mean is affine, we must then have

$$
\alpha \Phi_X(u) + j \beta \Phi_X(u) = \Phi_X(pu) \Phi_X(u)/\Phi_X(pu). \tag{10}
$$

Dividing by $\Phi_X(u)$ leads to a functional equation requiring that $\Phi_X(u)/\Phi_X(pu)$ be affine in $u$, which implies a Gaussian marginal distribution.

One of the interesting consequences of temporal asymmetry in autoregressive models is that forward and backward prediction errors need not be equal. Equality of these errors is implicit in signal processing algorithms such as Burg's maximum entropy method [12, p. 22]. In the case of our uniform AR(1) example, it follows from (7) that the input $W_n$ takes the values $-(|k| - 1)/2 |k|, -(|k| - 3)/2 |k|, \ldots, (|k| - 1)/2 |k|$ with equal probability. It can then be shown that $X_{n-1}$ is completely determined by $X_n$: $X_{n-1} = (kX_n) \bmod 1 - 1/2$ for $k$ even and $X_{n-1} = (kX_n + 1/2) \bmod 1 - 1/2$ for $k$ odd. Thus the process has zero prediction error in the backward direction, while the forward mean-squared prediction error equals the mean-squared value of $W_n$ [13].

**Theorem 2:** The backward mean-squared prediction error of a non-Gaussian AR(1) sequence is always less than the forward mean-squared prediction error.

**Proof:** The forward conditional mean of AR(1) models is linear, hence the forward mean-squared prediction error is

$$
E[(X_n - E[X_n | X_{n-1}])^2] = E[(X_n - \rho X_{n-1} - E[W_n])^2]. \tag{11}
$$

Since the conditional expectation is the best mean-square estimator, we have

$$
E[(X_n - E[X_{n-1} | X_n])^2] \leq E[(X_n - g(X_n))^2] \tag{12}
$$

for all $g(.)$ with equality only when $g(X_n) = E[X_{n-1} | X_n]$. Making use of Theorem 1, we find that

$$
E[(X_n - E[X_{n-1} | X_n])^2] < E[(X_n - \rho X_{n-1} - E[W_n])^2]. \tag{13}
$$

The earlier example at the end of Section II corresponds to $\lambda = 2$.  

This curious example was perhaps first pointed out by Rosenblatt [24, p. 52].
variable belonging to class $L$ as [29]
\[
\Psi_x(u) = \log \Phi_x(u) = \int_{R_0} \left( e^{i \alpha u} - 1 - \frac{k(\alpha)}{\alpha} \right) d\alpha \tag{13}
\]
where $R_0 = \mathbb{R} \setminus 0$ and $k(\alpha)$ is an odd symmetric function which is nonnegative, nonincreasing on $(0, \infty)$ and satisfies the condition $\int_{|\alpha| < \infty} |\alpha k(\alpha) \, d\alpha + \int_{|\alpha| > \infty} |\alpha^{-1} k(\alpha) \, d\alpha < \infty$. Taking the derivative on both sides of (13)
\[
-j \frac{d\Psi_x(u)}{du} = \int_{R_0} e^{i \alpha u} k(\alpha) \, d\alpha
\]
which is the Fourier transform of $k(\alpha)$. This formula can be used to verify membership in class $L$ and further characterize its subclasses. To belong to class $L$, the derivative of the logarithm of a candidate distribution's characteristic function (multiplied by $-j$) must have a Fourier transform having the properties of $k(\alpha)$. In the accompanying table, we list the Lévy measure functions $k(\alpha)$ corresponding to some of the known symmetric non-Gaussian distributions.

Consider, for example, the first-order Laplacian autoregressive (LAR(1)) model [17]. If $\{X_n\}$ has a Laplacian density with zero mean and unit variance, $p_x(x) = \exp \left\{ -\sqrt{2} |x|^\frac{3}{2} \right\}$, and $\Psi_x(u) = 2(2 + u^2)$. Substituting this characteristic function into (7), we find that the result is indeed a valid characteristic function with
\[
\Phi_w(u) = \rho^2 + (1 - \rho^2) \frac{2}{2 + u^2}.
\]
Thus, the input $W_n$ is zero with a probability $\rho^2$ and a normalized Laplacian with probability $1 - \rho^2$. In other words, the generating system in (6) could be written for the LAR(1) process as
\[
X_n = \rho X_{n-1} + a_n W_n,
\]
where $\{W_n\}$ is i.i.d. Laplacian (zero mean, unit variance) and $a_n$ is an independent discrete random variable taking the values 0 and 1 with probabilities $\rho^2$ and $1 - \rho^2$, respectively. Generation of the Laplacian model thus requires a random coefficient system. One of the consequences of a random coefficient generating model for the Laplacian case is the appearance of exponential "run downs" (with increasing probability as $\rho$ increases) in the sample functions. This effect is illustrated in Fig. 2(c) for a high correlation of 0.8. This trend may be unsuitable for modeling a given set of data.

In contrast, suppose $\{X_n\}$ has a hyperbolic secant amplitude distribution:
\[
p_x(x) = \frac{1}{2} \operatorname{sech} \frac{x}{2} = \frac{1}{e^{x/2} + e^{-x/2}}, \quad -\infty < x < \infty.
\]
Substituting its characteristic function into (7) and evaluating the inverse Fourier transform [8], [23], we obtain the marginal density of the input $\{W_n\}$:
\[
p_w(w) = \frac{\cos \pi \rho / 2 \cosh \pi w / 2}{\cos \pi \rho + \cosh \pi w}, \quad -\infty < w < \infty.
\]
The system thus required to generate first-order Markov hyperbolic secant, HAR(1), distributed data is not a random coefficient system. Instead, the input to the first-order AR system is a sequence of i.i.d. random variables having the distribution given above. HAR(1) data having a correlation of 0.8 is plotted in Fig. 2(a) along with Gaussian and Laplacian AR(1) data of same correlation for comparison. Although both the hyperbolic secant and Laplacian densities have exponential tails, the IIAR(1) and LAR(1) data differ markedly because of the exponential run downs in the LAR(1) case.

Note from Table I that only in the Laplacian case is $k(\alpha)$ bounded at the origin and that it requires a random coefficient system. In general, we can use this boundedness criterion to determine which class $L$ distributions will necessitate a random coefficient system. For the density function of $\{W_n\}$ not to have an impulse at the origin (which results in a random coefficient system)
\[
\lim_{x \to -\infty} \Phi_x(u) = 0 = \lim_{x \to \infty} \Phi_x(u) = \Psi_x(u) - \Psi_x(\rho u) = -\infty.
\]
From (13), we must then have
\[
\lim_{x \to -\infty} \int_{R_0} \left( e^{i \alpha u} - e^{i \alpha \rho u} \right) \frac{k(\alpha)}{\alpha} \, d\alpha = -\infty.
\]
TABLE I

LEVY MEASURE FUNCTIONS OF SOME SYMMETRIC NON-GAUSSIAN DISTRIBUTIONS

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Characteristic Function $\Phi_X(u)$</th>
<th>Lévy Measure Function $k(\alpha)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian</td>
<td>$1 / (1 + u^2)$</td>
<td>$e^{-</td>
</tr>
<tr>
<td>Hyperbolic secant</td>
<td>sech $u$</td>
<td>$\text{sign} (\alpha)$</td>
</tr>
<tr>
<td>Hyperbolic secant squared</td>
<td>$\frac{\pi u}{2} \coth \left( \frac{\pi u}{2} \right)$</td>
<td>$- \frac{\text{sign} (\alpha) \exp \left( \left</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$e^{-</td>
<td>\alpha</td>
</tr>
<tr>
<td>Symmetric stable</td>
<td>[ \exp \left{ - \frac{2}{\beta(\beta-2) \sin \frac{\pi}{\beta}} \left</td>
<td>u \right</td>
</tr>
</tbody>
</table>

On the other hand, since $\lim_{u \to 0} \int_{-\infty}^{\infty} e^{iuw} k(\alpha)/\alpha \, du \neq \lim_{u \to 0} k(\alpha)$, for the input distribution not to have an impulse at the origin, the Lévy measure must be unbounded at the origin. All $k(\alpha)$ which are bounded at the origin will necessarily demand random coefficient systems for the generation of data having the corresponding amplitude distribution.

Because the input distribution for the HAR(1) model is absolutely continuous, the transitional density of this first-order Markov process can be derived. Using (6)

$$p_{y|\lambda=0}(y) = p_{w}(y) = \frac{\cos \pi \rho/2 \cosh \pi (y - \rho x)/2}{\cos \pi \rho + \cosh \pi (y - \rho x)}/2 . \quad (14)$$

Let us now detail how HAR(1) data can be generated. First the i.i.d. input sequence $\{W_s\}$ is generated from the independent sequence $\{U_s\}$ distributed uniformly between 0 and 1, using the pointwise transformation $W_s = P_w^{-1}(U_s)$ where $P_w^{-1}(\cdot)$ is the inverse of the distribution function of $\{W_s\}$. The correlated data sequence $\{X_s\}$ is then obtained by passing $\{W_s\}$ through the linear system described in the previous section, a general procedure now simplified by the assumption of the linear model. The distribution function of $W_s$ is found to be

$$P_w(w) = P_X \left( \frac{2}{\pi} \sinh^{-1} \left( \frac{\sinh \pi w/2 \cosh \pi \rho/2}{\cos \pi \rho/2} \right) \right) . \quad (15)$$

where $P_X(x) = (2/\pi) \tan^{-1} (\exp \{\pi x/2\})$. Using this distribution function in (15) and evaluating the inverse, we obtain

$$P_{w}^{-1}(u) = \frac{2}{\pi} \sinh^{-1} \left[ \cos \frac{\pi \rho}{2} \sinh \left( \tan^{-1} \frac{\pi u}{2} \right) \right] .$$

If we remove the restriction demanding a model for all $\rho$, output distributions not in class $L$ are possible. In some cases (as when the characteristic function $\Phi_X(u)$ is non-monotonic [11]), there is a critical value $\rho_c$ of the system coefficient beyond which the ratio in (7) exceeds unity, a situation incompatible with the ratio being a characteristic function. Consequently, highly correlated sequences cannot be generated having such amplitude distributions while they can be generated for smaller correlations. If the characteristic function has zeros, the range of $\rho$ is further restricted. Supposing that $u_0$ is a zero of $\Phi_X(\cdot)$, the denominator of (7) becomes zero when $u = u_0/\rho$; for the ratio to be bounded, $u_1 = u_0/\rho$ must also be a zero of $\Phi_X(\cdot)$. This condition then becomes recursive since a zero is required at $u_2 = u_0/\rho^2$, $u_3 = u_0/\rho^3$, etc. Thus $\Phi_X(\cdot)$ must have an infinite number of zeros if it has any. For the uniform AR(1) example, the characteristic function is sin $u/u$ and has an infinite number of equally spaced zeros. First-order Markov uniformly distributed time series are thus defined only for $\rho = 1/k$, $k = \pm 2, \pm 3, \ldots$. Systems with such discrete sets of coefficients seem to be of academic interest only.

Verification of compatibility of a time series' amplitude distribution with the conditions implicit in (7), positive definiteness of the ratio $\Phi_X(u)/\Phi_X(\rho u)$, represents a formidable task if only an analytic approach is used. Success is limited by one's ability to derive the inverse Fourier transform of this ratio and show that the result is non-negative for some range of $\rho$. We used numerical methods to check for the existence of first-order Markovian non-Gaussian time series other than those in class $L$ and variants of the uniform example given above. Our procedure can be used whenever a symmetric histogram estimate of the probability density of $\{X_s\}$ is available: an analytic specification is not necessary. The test consists of the following steps.

1. Given a sampled probability density function, resample it at a lower (rational) rate. Any of several decimation/interpolation strategies can be used here [4].

2. Fourier transforms of the original and downsampled density are computed with care taken that the sum of each density sequence is unity.

3. The point-by-point ratio of these transforms is com-
puted and windowed to eliminate inaccurate division where either of the transforms is small. This window must be chosen so that negative ripples are not introduced in the amplitude domain. Consequently, positive definite windows like the triangular would suffice. We have found that a nondefinite window such as the rectangular one can be used by noting how negative its ripples become and numerically checking that no ripple exceeds that value.

4) The inverse Fourier transform of the windowed ratio is computed and checked for "essential" positivity: negative portions are allowed to exist but should be within numeric inaccuracies.

To illustrate the procedure, we choose the sampled version of the weighted sum of three equal variance Gaussians with means $-1.2, 0,$ and $1.2,$ respectively. The resulting density is multimodal and hence is not in class $L.$ We investigated whether this density could be generated by first-order systems with coefficients $1/5$ and $1/3.$ See Fig. 3. We reduced the sampling rate of the density vector by factors of 5 and 3 by simple downsampling, taking care that aliasing was minimal by computing the Fourier transform. The point-by-point ratio of the two transforms contained numeric noise in the high frequency region due to rounding. We used a rectangular window to remove this noise and obtained inverse transforms shown in the fourth row of Fig. 3. Clearly, the example density seems compatible with $\rho = 1/5$ but not with $\rho = 1/3$ as the latter results contain significant negative values in the tails. This threshold is close to the critical value $\rho_c$ mentioned previously.

While this numeric approach is imprecise, it can be validated via simulation. Assuming a candidate distribution seems viable, the result of the numeric test is the amplitude distribution of the input. By simple calculation of the partial sums, the cumulative distribution of the input can be calculated and used to generate the i.i.d. sequence $\{W_n\}$ predicted by the computations. By passing this sequence through a first-order filter, estimating the amplitude distribution of the output, and comparing the estimate with the candidate distribution, the prediction can be confirmed. We performed this test on the trimodal example just described for $\rho = 1/5.$ The resulting estimate of the output distribution did greatly resemble the candidate distribution and verified that amplitude distributions produced by first-order systems need not be unimodal. We have thus demonstrated the existence of such densities more directly than in [11].

B. Nonlinear Markov Processes and Diagonal Expansions of Bivariate Distributions

Bivariate distributions of stationary random processes have in the past been analyzed using series expansion methods. These expansions find application in the study of Markov processes [33] and of the effect of nonlinearities on random processes, e.g., in the analysis of the output of a cascade of a narrow-band filter and a square law detector [1].

Let $p_{X,Y}(x,y)$ be the joint density of random variables $X$ and $Y$ with marginal densities $p_X(x)$ and $p_Y(y)$, respectively. Suppose $p_{X,Y}(x,y)$ satisfies the condition

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x,y) \, dx \, dy < \infty.$$  \hspace{1cm} (16)

Then, complete orthonormal sets $\{\phi_n(x)\}_{n=0}^{\infty}$ and $\{\psi_n(y)\}_{n=0}^{\infty}$ can be defined in $L^2(p_X \, dx)$ and $L^2(p_Y \, dy)$ such that the series expansion

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \left\{ 1 + \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\psi_i(y) \right\}$$  \hspace{1cm} (17)

commonly referred to as the diagonal expansion because of the single summation, converges in mean-square sense [16]. A well-known example is Mehler's expansion of a bivariate Gaussian density in terms of Hermite poly-
mials [1]. Note that by definition \( \phi_0(\cdot) = \psi_0(\cdot) = 1 \) and that \( \lambda_0 = 1 \). The coefficients of expansion are given by

\[
\lambda_i = \int \int \phi_i(x)\psi_i(y)p_{X,Y}(x, y) \, dx \, dy
\]

and are by convention taken to be nonnegative: the sign is incorporated in the orthonormal sets [30]. The ordering of the basis functions is determined so that the coefficients represent a decreasing sequence: \( 0 \leq \lambda_i \leq 1 \) and \( \lambda_i \geq \lambda_j \) for \( i \geq j \). Furthermore, note that the orthonormal sets are complementary eigenfunctions of the bivariate density:

\[
\begin{align*}
\int p_{X,Y}(x, y)\phi_i(x) \, dx &= \lambda_i\phi_i(x) \\
\int p_{X,Y}(x, y)\psi_i(y) \, dy &= \lambda_i\psi_i(y)
\end{align*}
\]

The terms inside the summation sign of (17) account for the dependency between the two random variables \( X \) and \( Y \); if \( \lambda_i = 0 \) for \( i \geq 1 \), \( X \) and \( Y \) are statistically independent. The coefficient \( \lambda_i \) is referred to as the maximal correlation coefficient because it is the supremum over all (finite-variance) functions \( g_1(\cdot) \) and \( g_2(\cdot) \), of the normalized correlation between \( g_1(X) \) and \( g_2(Y) \) [28]. If \( \phi_i(\cdot) \) and \( \psi_i(\cdot) \) are affine, it follows from orthonormality that \( \phi_i(x) = (x - \mu_x)/\sigma_x \) and \( \psi_i(y) = (y - \mu_y)/\sigma_y \) where \( \mu \) and \( \sigma^2 \) are the mean and variance, respectively. In this case, the maximal correlation coefficient \( \lambda_i \) coincides with the usual correlation coefficient. However, \( \lambda_i \) is in general larger than the correlation coefficient in magnitude and gives a better characterization of the dependence between \( X \) and \( Y \); the random variables are independent if and only if \( \lambda_i = 0 \) [26]. The maximal correlation coefficient and the functions \( \phi_i(\cdot) \) and \( \psi_i(\cdot) \) are important in approximating the series expansion in (17) with a finite number of terms.

Let us apply the foregoing discussion to a stationary Markov random sequence \( \{X_t\} \). Denote its marginal density by \( p_X(\cdot) \) and the joint density of \( X_n \) and \( X_{n-1} \) by \( p_{X_n, X_{n-1}}(\cdot, \cdot) \). We must have

\[
\int p_{X_n, X_{n-1}}(x, y) \, dy = p_X(x)
\]

and

\[
\int p_{X_n, X_{n-1}}(x, y) \, dx = p_X(y)
\]

for all \( n \) because of stationarity. These conditions, however, do not restrict the bivariate density functions of the process to be symmetric; \( p_{X_n, X_{n-1}}(x, y) \) is not necessarily identical to \( p_{X_n, X_{n-1}}(y, x) \). Although asymmetric diagonal expansions of the form given by equation (17) have been studied before [2], [26], [31], they have not been applied to random processes. A special case of (17) is commonly considered where the two sets of orthonormal functions are identical, which yields a symmetric bivariate density and imposes temporal symmetry on the underlying time series. As we have noted earlier, in contrast to the Gaussian case, the bivariate densities of non-Gaussian processes need not be symmetric because of temporal asymmetry. For example, for the HAR(1) model

\[
P_{X_n, X_{n-1}}(x, y) = \frac{1}{2} \text{sech} \frac{\pi x}{2} \frac{\cos \pi r^m/2 \cosh \pi(y - \rho^m x)/2}{\cos \pi r^m + \cosh \pi(y - \rho^m x)}.
\]

The necessity of the general expansion (17) for non-Gaussian processes is thus clear.

Using the diagonal expansion of \( p_{X_n, X_{n-1}}(x, y) \), we can write the conditional distribution function of the Markov process \( \{X_t\} \) as

\[
P_{X_t|X_{t-1}}(x|y) = \int_{-\infty}^{\infty} p_{X}(x) \left\{ 1 + \sum_{a=1}^{\infty} \lambda_i \phi_i(x)\psi_i(y) \right\} \, dz.
\]

The diagonal expansion thus serves as a tool for analyzing the generating system of the Markov process. However, as we have seen in Section II, we need the inverse of the above conditional distribution function to generate the Markov sequence from an i.i.d. uniform sequence. Calculating the inverse of the summation in (20) even in an approximate form is extremely difficult. If, however, we limit the diagonal expansion to a finite number of terms (making sure that the integrand is nonnegative) such that the conditional distribution function can be inverted, we have a method for generating correlated non-Gaussian Markov sequences that are not necessarily linear. Since the additional dependency between \( X_n \) and \( X_{n-1} \) with each added term decreases progressively (\( \lambda_i > \lambda_{i+1} \), these terms could be selected to match the required dependency to a large extent. Sarmanov [27] studied the finite sum, continuous-time version of (20). For continuous-time processes, diagonal expansions with finite number of terms cannot be used when the eigenfunctions and the marginal density function are continuous: the finite sum does not remain nonnegative over the entire domain for all values of separation \( t \) between the samples. Fortunately, this problem does not arise in the discrete-time case.

If the domain of the marginal density is finite, uniform on \([0, 1]\) for example, polynomials can be used in the finite sums. If the marginal density function is one of the classical weight functions, orthogonal polynomials such as Jacobi polynomials can be used; otherwise the moments of the distribution can be used to construct orthogonal polynomials. For example, a uniform \([0, 1]\) distributed, temporally symmetric Markov process can be defined by the joint amplitude distribution

\[
P_{X_n, X_{n-1}}(x, y) = 1 + a(2x - 1)(2y - 1)
\]

\(|a| \leq 1, \quad 0 \leq x, y \leq 1. \]

For distributions defined on the infinite domain, the functions in the expansion have to be chosen depending on the particular case. However, some recipes applicable
to all situations exist. Take for instance, the temporally symmetric process defined by the Morgenstern’s family [5, p. 578] of joint densities

\[ p_{X,Y}(x,y) = p_T(x)p_T(y)(1 + a(2P_T(x) - 1)) \cdot (2P_T(y) - 1) \].

(22)

It can be obtained by passing the uniform Markov sequence \( \{V_\tau\} \) defined in (21) through the nonlinearity \( P_T(\cdot) \) where \( P_T(\cdot) \) is the (cumulative) distribution function of the required output \( \{Y_\tau\} \). Such an operation on \( \{V_\tau\} \) leads to \( p_{X,Y}(x,y) = p_{X,Y}(P_T(x), P_T(y)) \) from which (22) follows by differentiation. Devroye [5, p. 580] gives a simple generation rule for \( \{V_\tau\} \) (and thus for \( \{Y_\tau\} \)). A more direct generation via (4) requires the evaluation of the conditional distribution. This procedure results in a quadratic equation for \( P_T(Y_\tau) \) involving \( Y_{\tau - 1} \) and \( U_\tau \) (recall that this is i.i.d. uniform [0, 1]), the inversion of which gives us the generation formula

\[ Y_\tau = \begin{cases} \frac{P_T^{-1} \left( \frac{3aQ(Y_{\tau - 1}) - 1 + \sqrt{(1 - 3aQ(Y_{\tau - 1}))^2 + 12aU_\tau Q(Y_{\tau - 1})}}{6aQ(Y_{\tau - 1})} \right)}{2P_T(U_\tau)}, \\ P_T^{-1}(U_\tau), \end{cases} \]

where we have set \( Q(Y_{\tau - 1}) = 2P_T(Y_{\tau - 1}) - 1 \) for simplicity.

As an example of a temporally asymmetric case, consider \( \{Z_\tau\} \) defined by

\[ p_{Z,Z}(z_1, z_2) = p_Z(z_1)p_Z(z_2)(1 + a(3P_Z(z_1)

- 2P_Z(z_2))(2P_Z(y) - 1)). \]

(23)

Proceeding as above for the generation of \( \{Z\} \), we find a cubic equation in \( p_Z(Z) \) which makes the generation difficult. It is much simpler to generate the process backwards. In this case, we obtain a quadratic equation for \( p_Z(Z_{\tau - 1}) \) as in the symmetric case above, with the coefficients being different functions of \( Z_\tau \) and \( U_\tau \).

A major drawback in using the diagonal expansion method for generating correlated sequences is that the entire range of correlation coefficients cannot be realized (for the processes \( \{V_\tau\} \) and \( \{Y_\tau\} \) above, \( |\phi| \leq 1/3 \) and for \( \{Z\} \), \( |\phi| \leq \sqrt{2}/3 \sqrt{5} \)). Maximal correlated random variables are important in variance reduction techniques in Monte Carlo simulation. Typically, adding more terms in the finite sum improves the available range of correlations, but it becomes increasingly difficult to ensure the nonnegativity of the sum. The question of the available range of correlation can be linked to the comprehensiveness of the defining bivariate distributions [5]. A family of bivariate distributions is said to be comprehensive if it includes the product of marginals and Frechet’s extremal distributions (which result in extremum positive and negative correlations 1 and -1). Clearly, the first requirement is satisfied for the joint densities defined using diagonal expansion methods while the second is usually not.

For an arbitrary stationary Markov process, the conditional means may or may not be linear. These quantities are given in terms of the components of the diagonal expansion as

\[ E[X_\tau | X_{\tau - 1}] = \sum_{i=1}^{\infty} \lambda_i E[X_\tau \phi_i(X_\tau)] \psi_i(X_\tau - 1) \]

\[ E[X_{\tau - 1} | X_\tau] = \sum_{i=1}^{\infty} \lambda_i E[X_{\tau - 1} \psi_i(X_{\tau - 1})] \phi_i(X_\tau). \]

If any basis function \( \{\psi_i(\cdot)\} \) is linear, under the condition that the functions

\[ \int_\tau \frac{\partial}{\partial x} p_{X,Y}(x, t) \frac{dx}{p_x(x)} \quad \text{and} \quad \int_x \frac{\partial}{\partial y} p_{X,Y}(t, y) \frac{dy}{p_y(y)} \]

do not change sign on their domain \((a, b)\), we can conclude that no other linear term is present and that the linear term must be the first member \( \phi_1(\cdot) \) [21]. The series expansion for the forward conditional mean then truncates with the result

\[ E[X_\tau | X_{\tau - 1}] = \lambda_1 E[X_\tau \phi_1(X_\tau)] \psi_1(X_{\tau - 1}) \]

indicating that the forward conditional mean is proportional to the eigenfunction \( \psi_1(X_{\tau - 1}) \). Similarly, when one of the members of \( \{\psi_i(\cdot)\} \) is linear, the expansion for the backward conditional mean reduces to a single term. In these cases, the conditional means yield direct information about the components of the diagonal expansion, thereby leading to the conditional distribution and the generation system.

We tested the analysis procedures described here as well as the more common, Gaussian based ones on two sets of data: the linear HAR(1) time series \( \{X_\tau\} \) and the nonlinear time series \( \{Y_\tau\} \) defined by (22) also having a hyperbolic secant marginal distribution. The nonlinear model is thus defined by

\[ p_{X,Y}(x,y) = \frac{1}{2} \text{sech} \frac{x}{2} \cdot \frac{1}{2} \text{sech} \frac{y}{2} \left[ 1 + 3a \cdot \left\{ 4 \tan^{-1} \left( e^{x/2} \right) - 1 \right\} \right] \]

\[ \cdot \left\{ 4 \tan^{-1} \left( e^{y/2} \right) - 1 \right\} \].

The correlation coefficient of the adjacent samples of this process is \( \rho_T = 3b^2/a = 0.88375 \) where \( b = \int_{-\infty}^{\infty} (4/x) \tan^{-1} e^{x/2} \left( 1/2 \right) \text{sech} x/2 \cdot dz = 14/\pi^3 \cdot \frac{1}{3} \) and \( f(\cdot) \) is Riemann’s zeta function. Also, from (2) we
we get

\[ p_{x_{t}, x_{t-1}}(x, y) = \frac{1}{2} \text{sech} \frac{\pi x}{2} - \frac{1}{2} \text{sech} \frac{\pi y}{2} \left( 1 + 3a^2 \right) \]

\[ \cdot \left[ \frac{4}{\pi} \tan^{-1} \left( e^{\frac{\pi}{2}} \right) - 1 \right] \]

\[ \cdot \left[ \frac{4}{\pi} \tan^{-1} \left( e^{\frac{\pi}{2}} \right) - 1 \right] \]

and it follows that the correlation function of this temporally symmetric model is \( R_x(m) = (1 - 3b^2)5(m) + 3b^2a^2 \). Note that this temporally symmetric first-order Markov time series can only be produced by a nonlinear system. The correlation function of the corresponding linear time series, while having the same marginal distribution, is simply \( R_x(m) = \rho^m \).

We generated these two time series so that they have the same correlation coefficient of 0.25 between adjacent samples. The power spectral density estimates of the two data plotted in Fig. 4 are quite similar. The forward conditional mean of the HAR(1) data is linear while the backward mean is not, thus confirming its temporal asymmetry. For the nonlinear, temporally symmetric data, both conditional means are nonlinear but identical. See Fig. 5. The conditional means thus identify temporal asymmetry well where power spectral estimation based techniques fail.

IV. CONCLUSIONS

Non-Gaussian processes present new challenges to the statistical signal processor attempting to develop analysis techniques. We have shown that correlation analysis cannot be expected to suffice, which immediately distinguishes non-Gaussian data from Gaussian. Temporal symmetry can be assessed with conditional mean analysis. While not shown here, the statistical characteristics of conditional mean estimates are identical to those of histogram-based probability density estimators [25]. Consequently, hypothesis tests for determining the temporal symmetry of a time series can be established. However, first-order conditional mean analysis does not capture all of a time series' temporal symmetry properties: similar forward and backward means can belie an asymmetric process.
Successful analysis can be measured by one's ability to generate a statistically identical version via simulation. This yardstick has implicitly formed the basis for our system-based modeling of non-Gaussian Markov processes. Linearity of the system can be tested by considering the forward and backward conditional means. If the forward conditional mean is linear and the backward mean is non-linear (the time series must be temporally asymmetric if linear and non-Gaussian), then a linear model may suffice. If a first-order model is appropriate, the amplitude distribution of the driving "white noise" process can be determined with our numerical method. If the method fails however, for the reasons given above, they are limited in scope. New techniques based on a fundamental understanding of non-Gaussian processes are clearly needed.

**References**


P. Srinivasa Rao (S’89) was born in Gunur, AP, India, on September 23, 1964. He received the B.Tech. degree in 1983 from the Indian Institute of Technology, Madras, and the M.S. degree in 1988 from Rice University, Houston, TX. He is currently working towards the Ph.D. degree at Rice University. His research interests include statistical signal processing, communication theory, and the theory of random processes.

Don H. Johnson (M’78–SM’83–F’90) was born in Mt. Pleasant, TX, on July 9, 1946. He received the S.B. and S.M. degrees in 1970 and the Ph.D. degree in 1974, all in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He joined the M.I.T. Lincoln Laboratory as a Staff Member in 1974 and worked on digital speech systems; he is currently a consultant there. In 1977, he joined the faculty of the Department of Electrical and Computer Engineering at Rice University, Houston, TX, where he is currently Professor. He received an American Society of Engineering Education Summer Study Fellowship in 1980 and studied passive sonar systems at the Naval Ocean Systems Center. His present research activities focus on statistical signal processing and the transmission of information by neural signals.

Dr. Johnson is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and is Chair of the Signal Processing Society’s Digital Signal Processing Technical Committee. He is a member of the Acoustical Society of America and the American Academy for the Advancement of Science, as well as Eta Kappa Nu and Tau Beta Pi.

David D. Becker (S’88–M’89) was born in Milwaukee, WI, on July 18, 1967. He received the B.S. degree in electrical engineering and the B.A. degree in mathematical sciences (summa cum laude) from Rice University, Houston, TX, in 1989 and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1991.

Since February 1991, he has been with General Electrical Medical Systems in Waukesha, WI, where he works in hardware design for diagnostic ultrasound imaging equipment.

Mr. Becker is a member of Phi Beta Kappa, Tau Beta Pi, and Eta Kappa Nu.