# Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization

Ignacio Santamaría, *Senior Member, IEEE*, Puskal P. Pokharel, *Student Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

*Abstract*—With an abundance of tools based on kernel methods and information theoretic learning, a void still exists in incorporating both the time structure and the statistical distribution of the time series in the same functional measure. In this paper, a new generalized correlation measure is developed that includes the information of both the distribution and that of the time structure of a stochastic process. It is shown how this measure can be interpreted from a kernel method as well as from an information theoretic learning points of view, demonstrating some relevant properties. To underscore the effectiveness of the new measure, a simple blind equalization problem is considered using a coded signal.

*Index Terms*—Blind equalization, entropy, generalized correlation kernel, information theoretic learning, reproducing kernel Hilbert space (RKHS).

## I. INTRODUCTION

**N**ATURAL processes of interest for engineering are composed of two basic characteristics: statistical distribution of amplitudes and time structure. Time in itself is very fundamental and is crucial to many real-world problems, and the instantaneous random variables are hardly ever independently distributed, i.e., stochastic processes possess a time structure. For this reason, there are widely used measures that quantify the time structure like the autocorrelation function. On the other hand, there are a number of methods that are solely based on the statistical distribution, ignoring the time structure. A single measure that includes both of these important characteristics could greatly enhance the theory of stochastic random processes.

The fact that reproducing kernels are covariance functions as described by Aronszajn [1] and Parzen [2] explains their early role in inference problems. More recently, numerous algorithms using kernel methods, including support vector machines [3], kernel principal component analysis [4], kernel Fisher discriminant analysis [5], and kernel canonical correlation analysis [6],

[7] have been proposed. Likewise, advances in information theoretic learning (ITL), have brought out a number of applications, where entropy and divergence employ Parzen's nonparametric estimation [8]–[11]. Many of these algorithms have given very elegant solutions to complicated nonlinear problems. Most of all these contemporary algorithms are based on assumptions of independent distribution of data, which in many cases is not realistic. Obviously, an accurate description of a stochastic process requires both the information of the distribution and that of its time structure. A void still exists in incorporating both in the same functional measure, providing the insight and usefulness of similarity over time.

Our new function is a step forward in this direction. Specifically, we define a generalized correlation function (GCF) in terms of inner products of vectors in a kernel feature space. Since inner products are a measure of similarity, this function in effect measures the pairwise interaction of the feature vectors, separated by a certain time delay in input space. On the other hand and from an ITL point of view, this new measure quantifies the shape and size of the group of points in feature space, which gives the information of the statistical distribution in the input space.

Interestingly, it can be deduced that the GCF is directly related to Renyi's quadratic entropy estimate of data using Parzen windowing: For this reason, we will denote this new function as *correntropy*. Moreover, we show that it exhibits some properties that make it structurally similar to the correlation function.

The organization of the paper is as follows. After a brief presentation of some background on kernel methods and information theoretic learning, the definition of the new GCF, its main properties and some examples illustrating its behavior are presented in Section III. In Section IV, we present a simple application of the GCF to blind equalization of coded signals. As we shall discuss in more detail, our approach is based on the fact that in some communication systems the correlation resulting from a precoder is known in advance and this can be exploited to our advantage. Since we use more information of the time structure by employing correntropy, one can see a significant improvement over contemporary methods. Finally, Section V summarizes the main conclusions and points out some lines for further research.

## II. SOME BACKGROUND ON KERNEL-BASED ALGORITHMS AND ITL

In recent years, a number of kernel methods, including support vector machines [3], kernel principal component analysis [4], kernel Fisher discriminant analysis [5], and kernel

canonical correlation analysis [6], [7], have been proposed and successfully applied to several problems. The basic idea of kernel algorithms is to transform the data $\mathbf{x}_i$ from the input space to a high dimensional feature space of vectors $\Phi(\mathbf{x}_i)$, where the inner products can be computed using a positive definite kernel function satisfying Mercer's conditions [3]: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. This simple and elegant idea allows us to obtain nonlinear versions of any linear algorithm expressed in terms of inner products, without even knowing the exact mapping $\Phi$.

A particularly interesting characteristic of the feature space is that it is a *reproducing kernel Hilbert space* (RKHS), i.e., the span of functions $\{\kappa(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ defines a unique functional Hilbert space [1], [2], [12], [13]. The crucial property of these spaces is the *reproducing property* of the kernel

$$f(\mathbf{x}) = \langle \kappa(\cdot, \mathbf{x}), f \rangle, \quad \forall f \in \mathcal{F}.$$

In particular, we can define our nonlinear mapping from the input space to a RKHS as $\Phi(\mathbf{x}) = \kappa(\cdot, \mathbf{x})$, then we have

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = \langle \kappa(\cdot, \mathbf{x}), \kappa(\cdot, \mathbf{y}) \rangle = \kappa(\mathbf{x}, \mathbf{y})$$

and thus $\Phi(\mathbf{x}) = \kappa(\cdot, \mathbf{x})$ defines the Hilbert space associated with the kernel.

Without loss of generality, in this paper, we will only consider the translation-invariant Gaussian kernel, which is the most widely used Mercer kernel.[1]

$$\kappa(\mathbf{x} - \mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \left( \frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right). \tag{1}$$

On the other hand, ITL addresses the issue of extracting information directly from data in a nonparametric manner [8]. Typically, Renyi's entropy or some approximation to the Kullback–Leibler distance have been used as ITL cost functions and they have achieved excellent results on a number of problems, e.g., time-series prediction [9], blind-source separation [10], or equalization [11].

It has been recently shown that ITL cost functions, when estimated using the Parzen method, can also be expressed using inner products in a kernel feature space which is defined by the Parzen kernel, thus suggesting a close relationship between ITL and kernel methods [14], [15]. For instance, if we have a data set $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{R}^d$, and the corresponding set of transformed data points $\Phi(\mathbf{x}_i), \cdots, \Phi(\mathbf{x}_N)$, then it turns out that the squared mean of the transformed vectors, i.e.,

$$\|\mathbf{m}_\Phi\|^2 = \left\langle \frac{1}{N} \sum_{i=1}^{N} \Phi(\mathbf{x}_i), \frac{1}{N} \sum_{j=1}^{N} \Phi(\mathbf{x}_j) \right\rangle$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa(\mathbf{x}_i - \mathbf{x}_j)$$

is the information potential $V(\mathbf{x})$ as defined in [8].[2]

---

[1]Notice that in most of the kernel-based methods, the Gaussian kernel is defined without the scale factor $1/\sqrt{2\pi}\sigma$ On the other hand, in information theoretic measures derived from Parzen windowing with Gaussian kernels, the scale factor is considered in order to keep a constant unit volume of the kernel. Here, we follow this probability density function (pdf)-based kernel definition.

[2]The quadratic Renyi's entropy is defined as $H_R = -\log(V(\mathbf{x}))$.



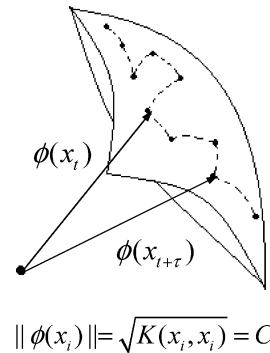$$\| \phi(x_i) \| = \sqrt{K(x_i, x_i)} = C$$

Fig. 1.   Mapping of a time series with the reproducing Gaussian kernel.

Similarly, the equivalence between kernel independent component analysis and a Cauchy–Schwartz independence measure has been pointed out in [16]. In fact, all learning algorithms that use nonparametric pdf estimates in the input space admit an alternative formulation as kernel methods expressed in terms of dot products. This interesting link allows us to gain some geometrical understanding of kernel methods, as well as to determine the optimal kernel parameters by looking at the pdf estimates in the input space.

Since the cost functions optimized by ITL algorithms (or, equivalently, by kernel methods) involve pdf estimates, these techniques are able to extract the higher order statistics of the data and that explains to some extent the improvement over their linear counterparts observed in a number of problems. Despite its evident success, a major limitation of all these techniques is that they assume independent and identically distributed (i.i.d.) input data. However, in practice, most of the signals in engineering have some correlation or temporal structure. Moreover, this temporal structure can be known in advance for some problems (for instance in digital communications working with coded source signals). Therefore, it seems that most of the conventional ITL measures are not using all the available information in the case of temporally correlated (nonwhite) input signals. The main goal of this paper is to present a new function that, unlike conventional ITL measures, effectively exploits both the statistical and the time-domain information about the input signal. This new function, which we refer to as GCF, will be presented in the next section.

## III. GENERALIZED CORRELATION FUNCTION

### A. Motivation and Definition

Let $\{\mathbf{x}_t, t \in T\}$ be a stochastic process with $T$ being an index set and $\mathbf{x}_t \in \mathcal{R}^d$. The nonlinear transformation induced by the reproducing kernel (1), maps the input samples over a sphere in the feature space, since $\|\Phi(\mathbf{x}_t)\|^2 = \kappa(0) = 1/(\sqrt{2\pi}\sigma)$. Therefore, the distance between $\Phi(\mathbf{x}_{t_1})$ and $\Phi(\mathbf{x}_{t_2})$ on that sphere (i.e., the geodesic distance) is proportional to the angle between the vectors from the origin to those points (see Fig. 1)

$$d(\Phi(\mathbf{x}_{t_1}), \Phi(\mathbf{x}_{t_2})) \propto \cos^{-1} \left( \frac{\langle \Phi(\mathbf{x}_{t_1}), \Phi(\mathbf{x}_{t_2}) \rangle}{\|\Phi(\mathbf{x}_{t_1})\| \|\Phi(\mathbf{x}_{t_2})\|} \right)$$

$$= \cos^{-1} \left( \sqrt{2\pi}\sigma\kappa(\mathbf{x}_{t_1} - \mathbf{x}_{t_2}) \right).$$

In other words, the kernel function is in fact computing the cosine of the angle between two points over the sphere (i.e., a distance). In addition, we notice from the previous discussion and also from Fig. 1 that the transformed data $\Phi(\mathbf{x}_t)$ must lie on some embedded manifold over the first quadrant on the sphere (since the kernel takes only positive values, the angle belongs to the interval $[0, \pi/2]$). Finally, it is also known from the link between ITL and kernel methods that the cloud of feature vectors also conveys information about the quadratic Renyi's entropy of the input data set (remember that $\log(\|\mathbf{m}_\Phi\|^2)$ is the quadratic Renyi's entropy).

With these remarks, we can define the new GCF as follows.

*Definition:* Let $\{\mathbf{x}_t, t \in T\}$ be a stochastic process with $T$ being an index set and $\mathbf{x}_t \in \mathcal{R}^d$. The generalized correlation function $V(t_1, t_2)$ is defined as a function from $T \times T$ into $\mathcal{R}^+$ given by

$$V(t_1, t_2) = E\left[\kappa\left(\mathbf{x}_{t_1} - \mathbf{x}_{t_2}\right)\right] \tag{2}$$

where $E[\cdot]$ denotes mathematical expectation over the stochastic process $\mathbf{x}_t$.

Using a series expansion for the Gaussian kernel, the GCF can be rewritten as

$$V(t_1, t_2) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} E\left[\|\mathbf{x}_{t_1} - \mathbf{x}_{t_2}\|^{2n}\right] \tag{3}$$

which involves all the even-order moments[3] of the random variable $\|\mathbf{x}_{t_1} - \mathbf{x}_{t_2}\|$. Specifically, the term corresponding to $n = 1$ in (3) is proportional to

$$E\left[\|\mathbf{x}_{t_1}\|^2\right] + E\left[\|\mathbf{x}_{t_2}\|^2\right] - 2E\left[\langle \mathbf{x}_{t_1}, \mathbf{x}_{t_2}\rangle\right]$$
$$= \sigma_{x_{t_1}}^2 + \sigma_{x_{t_2}}^2 - 2R_x(t_1, t_2)$$

where $R_x(t_1, t_2)$ is the covariance function of the random process; this shows that the information provided by the conventional covariance function (the autocorrelation for zero-mean processes) is included within the new function.

From (3), we can see that in order to have a univariate GCF, all the even-order moments must be invariant to a time shift. This is a stronger condition than wide sense stationarity, which involves only second-order moments. More precisely, a sufficient condition to have $V(t, t - \tau) = V(\tau)$ is that the input stochastic process must be strictly stationary on the even moments [17]; this means that the joint pdf $p(\mathbf{x}_t, \mathbf{x}_{t+\tau})$, must be unaffected by a change of time origin. We will assume this condition in the rest of the paper when using $V(\tau)$.

For a discrete-time strictly stationary stochastic process, we define the generalized correlation function as $V[m] = E[\kappa(\mathbf{x}_n - \mathbf{x}_{n-m})]$, which can be easily estimated through the sample mean

$$\hat{V}[m] = \frac{1}{N - m + 1} \sum_{n=m}^{N} \kappa(\mathbf{x}_n - \mathbf{x}_{n-m}). \tag{4}$$

[3]A different kernel would yield a different expansion, for instance the sigmoidal kernel $\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2}) = \tanh(\langle \mathbf{x}_{t_1}, \mathbf{x}_{t_2}\rangle + \theta)$ admits an expansion in terms of the odd moments of its argument.

### B. Properties

Some important properties of the GCF can be listed as follows.

*Property 1:* For any symmetric positive definite kernel (i.e., Mercer kernel) $\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ defined on $\mathcal{R} \times \mathcal{R}$, the GCF defined as $V(t_1, t_2) = E[\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})]$ is a reproducing kernel.

*Proof:* Since $\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ is symmetrical, it is obvious that $V(t_1, t_2)$ is also symmetrical. Now, since $\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ is positive definite, for any set of $n$ points $\{x_1, \cdots, x_n\}$ and any set of real numbers $\{a_1, \cdots, a_n\}$, not all zero

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(\mathbf{x}_i, \mathbf{x}_j) > 0. \tag{5}$$

It is also true that for any strictly positive function $g(\cdot, \cdot)$ of two random variables $x$ and $y$, $E[g(x, y)] > 0$. Then

$$E\left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(\mathbf{x}_i, \mathbf{x}_j)\right] > 0 \quad \Rightarrow$$
$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j E\left[\kappa(\mathbf{x}_i, \mathbf{x}_j)\right] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j V(i, j) > 0.$$

Thus, $V(t_1, t_2)$ is both symmetric and positive definite. Now, the Moore–Aronszajn theorem [1] proves that for every real symmetric positive definite function $\kappa$ of two real variables, there exists a unique RKHS with $\kappa$ as its reproducing kernel. Hence, $V(t_1, t_2) = E[\kappa(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})]$ is a reproducing kernel. This concludes the demonstration. ∎

The following properties consider a discrete-time stochastic process; obviously, the properties are also satisfied for continuous-time processes.

*Property 2:* $V[m]$ is a symmetric function $V[-m] = V[m]$.

*Property 3:* $V[m]$ reaches its maximum at the origin, i.e., $V[m] \leq V[0], \quad \forall m$.

*Property 4:* $V[m] \geq 0$ and $V[0] = 1/\sqrt{2\pi}\sigma$.

All these properties can be easily proved. Properties 2 and 3 are also satisfied by the conventional autocorrelation function, whereas Property 4 is a direct consequence of the positiveness of the Gaussian kernel.

*Property 5:* Let $\{\mathbf{x}_n, n = 0, \cdots, N-1\}$ be a set of i.i.d. data drawn according to some distribution $p(\mathbf{x})$. The mean value of the GCF estimator (4) converges asymptotically to the estimate of information potential obtained through Parzen windowing with Gaussian kernels.

*Proof:* The Parzen pdf estimate is given by $\hat{p}(\mathbf{x}) = (1/N) \sum_{n=0}^{N-1} \kappa(\mathbf{x} - \mathbf{x}_n)$, and the estimate of the information potential is

$$V = \int_{-\infty}^{\infty} \left(\frac{1}{N} \sum_{n=0}^{N-1} \kappa(\mathbf{x} - \mathbf{x}_n)\right)^2 d\mathbf{x}$$
$$= \frac{1}{N^2} \sum_{n'=0}^{N-1} \sum_{n=0}^{N-1} \kappa'(\mathbf{x}_n - \mathbf{x}_{n'}) \tag{6}$$

where $\kappa'$ denotes a Gaussian kernel with twice the kernel size of $\kappa$. On the other hand, the GCF estimate is $\hat{V}[m] = (1/N -$

$|m|) \sum_{n=m}^{N-1} \kappa(\mathbf{x}_n - \mathbf{x}_{n-m})$, for $-(N-1) \le m \le (N-1)$, and therefore its mean value is

$$\left\langle \hat{V}[m] \right\rangle = \frac{1}{2N-1} \sum_{m=-N+1}^{N-1} \frac{1}{N-|m|} \sum_{n=m}^{N-1} \kappa(\mathbf{x}_n - \mathbf{x}_{n-m}). \quad (7)$$

It is trivial to check that all the terms in (6) are also in (7); however, each term is weighted differently in both expressions. Therefore, the mean value of the GCF estimator is a biased (although asymptotically unbiased) estimator of the information potential. This concludes the proof. ∎

Property 5 clearly demonstrates that this generalization includes information about the pdf. On the other hand, we also showed that it also conveys information about the correlation. For these reasons, in the sequel we will refer to $V[m]$ as *correntropy*.

*Property 6:* Given $V[m]$ for $m = 0, \cdots, P-1$, then the following Toeplitz correntropy matrix of dimensions $P \times P$

$$\mathbf{V} = \begin{pmatrix} V[0] & V[1] & \cdots & V[P-1] \\ V[1] & V[0] & \cdots & V[P-2] \\ \vdots & \ddots & \ddots & \vdots \\ V[P-1] & V[P-2] & \cdots & V[0] \end{pmatrix}$$

is positive definite.

*Proof:* Matrix $\mathbf{V}$ can be decomposed as $\mathbf{V} = \sum_{n=m}^{N} \mathbf{A}_n$, where $\mathbf{A}_n$ is given by the equation shown at the bottom of the page, if $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel satisfying Mercer's conditions, then $\mathbf{A}_n$ is a positive definite matrix $\forall n$. On the other hand, the sum of positive definite matrices is also positive definite [18]; this proves that $\mathbf{V}$ is a positive definite matrix. ∎

We would like to point out that this important property opens the possibility of applying this new function to all the signal processing methods using conventional correlation matrices: signal and noise subspace decompositions, projections, etc. In particular, we can define a generalized power spectral density (PSD) function as

$$P(\omega) = \sum_{m=-\infty}^{\infty} V[m] e^{-j\omega m} \ge 0$$

which satisfies all the properties of a conventional PSD function and could be used to analyze time series in the frequency domain. Since it is the Fourier transform of correntropy, we denote this new function as *correntropy spectral density (CSD)*. The potential applications and usefulness of these ideas are beyond the scope of this paper.

*Property 7:* Let $\{x_n \in \mathcal{R}, n \in T\}$ be a discrete-time wide-sense stationary zero-mean Gaussian process with autocorrela-

tion function $r[m] = E[x_n x_{n-m}]$. The correntropy function for this process is given by

$$V[m] = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}, & m = 0 \\ \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma^2[m])}}, & m \ne 0 \end{cases} \quad (8)$$

where $\sigma$ is the kernel size and $\sigma^2[m] = 2(r[0] - r[m])$.

*Proof:* The correntropy function is defined as $V[m] = E[\kappa(x_n - x_{n-m})]$. Since $x_n$ is a zero-mean Gaussian random process, for $m \ne 0$, $z_m = x_n - x_{n-m}$ is also a zero-mean Gaussian random variable with variance $\sigma^2[m] = 2(r[0] - r[m])$. Therefore

$$V[m] = \int_{-\infty}^{\infty} \kappa(z_m) \frac{1}{\sqrt{2\pi}\sigma[m]} \exp - \left( \frac{z_m^2}{2\sigma^2[m]} \right) dz_m. \quad (9)$$

Since we are considering a Gaussian kernel with variance $\sigma^2$, (9) is the convolution of two zero-mean Gaussians of variances $\sigma^2$ and $\sigma[m]^2$ evaluated at the origin; this yields (8) immediately. ∎

Property 7 clearly reflects that correntropy conveys information about the time structure of the process and also about its pdf via quadratic Renyi's entropy. As a consequence of Property 7, if $\{x_n \in \mathcal{R}, n \in T\}$ is a white zero-mean Gaussian process with variance $\sigma_x^2$, we have that $V[m] = 1/(\sqrt{2\pi(\sigma^2 + \sigma_x^2)})$, $\forall m \ne 0$, which coincides with the mean value of the function and, of course, is the information potential[4] of a Gaussian random variable of variance $\sigma_x^2$, when its pdf has been estimated via Parzen windowing with a Gaussian kernel of size $\sigma^2$.

*Property 8:* The correntropy estimator (4) is unbiased and asymptotically consistent.

The properties of the estimator can be derived following the same lines used for the conventional correlation function [19].

### C. Some Examples of Correntropy

In this section, we show some examples that provide us a better understanding of the behavior and characteristics of correntropy. We consider the generation model shown in Fig. 2, where different source distributions and different linear time-invariant filters can be used. First, we switch off the filter and generate zero-mean unit variance white processes with different distributions: Gaussian, impulsive, and exponential. Fig. 3 depicts $V[m]$ and the conventional autocorrelation function $R[m]$. The correntropy function has been estimated from registers of $N = 10\,000$ samples, and the used kernel size was $\sigma = 1$. We can see that the mean value of correntropy changes for different

[4]$\int (\hat{p}(x))^2 dx = 1/\sqrt{2\pi(\sigma^2 + \sigma_x^2)}$.

$$\mathbf{A}_n = \begin{pmatrix} \kappa(\mathbf{x}_n - \mathbf{x}_n) & \kappa(\mathbf{x}_n - \mathbf{x}_{n-1}) & \cdots & \kappa(\mathbf{x}_n - \mathbf{x}_{n-P-1}) \\ \kappa(\mathbf{x}_n - \mathbf{x}_{n-1}) & \kappa(\mathbf{x}_n - \mathbf{x}_n) & \cdots & \kappa(\mathbf{x}_n - \mathbf{x}_{n-P-2}) \\ \vdots & \ddots & \ddots & \vdots \\ \kappa(\mathbf{x}_n - \mathbf{x}_{n-P-1}) & \kappa(\mathbf{x}_n - \mathbf{x}_{n-P-2}) & \cdots & \kappa(\mathbf{x}_n - \mathbf{x}_n) \end{pmatrix}$$
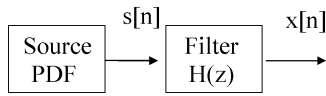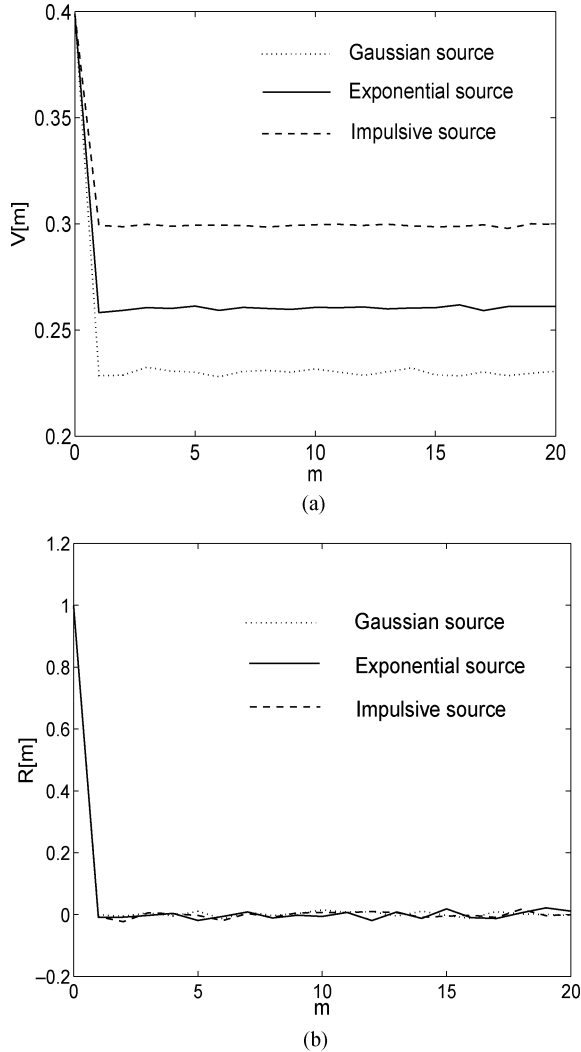
Fig. 2. Generation model.



(a)



(b)

Fig. 3. White processes with different input source distributions: (a) Correntropy and (b) autocorrelation.



(a)



(b)

Fig. 4. Filtered processes with different input source distributions: (a) Correntropy and (b) autocorrelation.

source distributions, whereas the autocorrelation function is basically the same, independent of the source distribution. Specifically, we can see from Fig. 3 that the value of correntropy at any nonzero lag corresponds the true value of the information potential, which can be theoretically obtained for Gaussian (0.23), impulsive (0.30), and exponential sources (0.26). This observation may yield a method to eliminate the bias on the estimation of entropy from finite data sets using the Parzen window method.

From this example, we can also see that in the case of i.i.d data, the correntropy at any lag (except zero) is all the information that we need to estimate quadratic Renyi's entropy. This idea can be related to the stochastic information gradient algorithm [20], which uses a pair of consecutive samples to estimate the entropy at the output of a linear system.
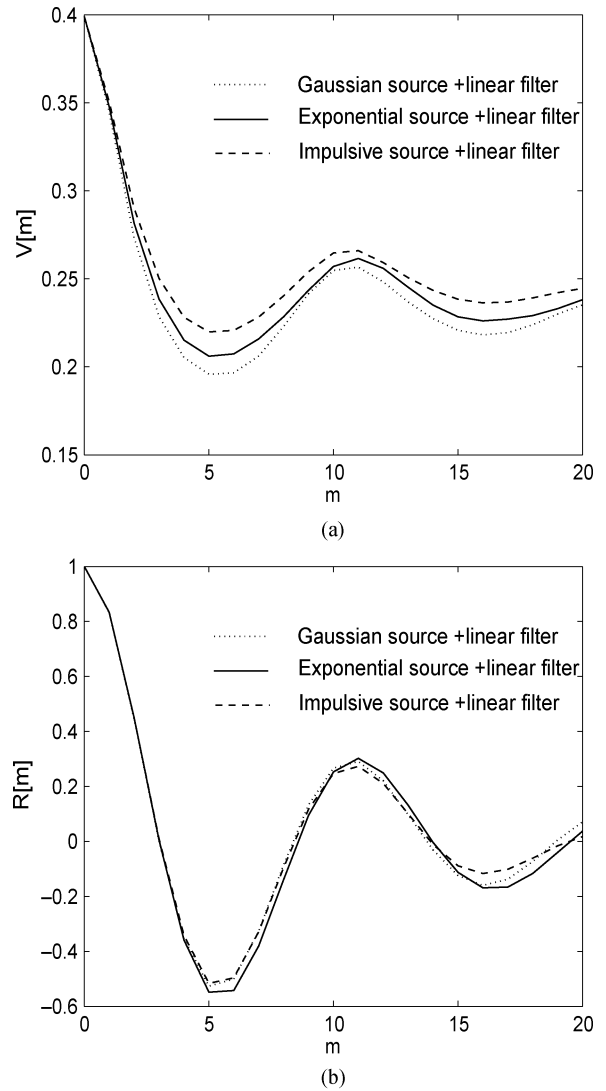
In the second experiment, the source distribution changes among Gaussian, exponential, and impulsive, but now we introduce the following infinite-impulse-response (IIR) filter[5]: $H(z) = 1/(1 - 1.5z^{-1} + 0.8z^{-2})$. Fig. 4 shows that, in this case, the overall shape of $V[m]$ is very similar to $R[m]$, thus reflecting the time structure of the process (which only has linear relationships among the variables). Since after linear filtering the pdf of the output becomes more Gaussian, the differences among the correntropy function for different source distributions are less evident in Fig. 4(a). Obviously, the differences would be smaller by using a longer filter. Finally, Fig. 5 shows the CSD estimated using 20 lags of the correntropy with mean removed and the true PSD: The main periodicity of the PSD is also evident in CSD, but other features are also present that require further investigation.

---

[5]For readers not familiar with the notation typically used in the signal processing literature $H(z)$ denotes the $z$ transform of the impulse response of the filter.
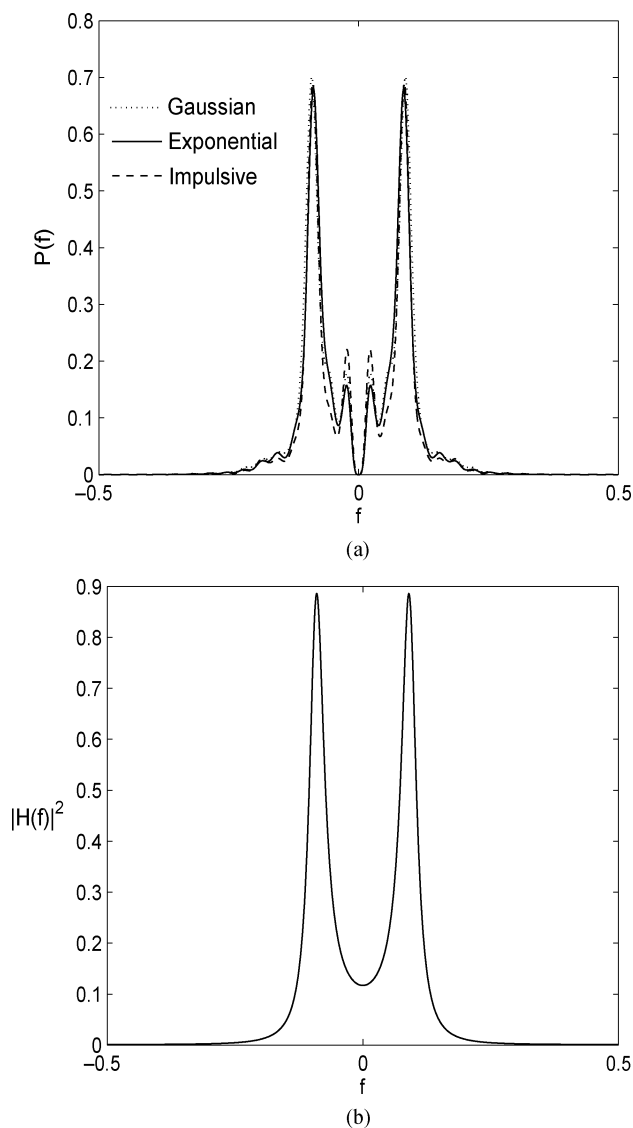
Fig. 5.   Comparison in the frequency domain: (a) Correntropy spectral density and (b) true PSD: $|H(f)|^2$.



Fig. 6.   (a) Correntropy and (b) autocorrelation for the Lorenz chaotic system.

In the final example, we consider the chaotic Lorenz system [21] described by the following equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = -y - xz + Rx$$
$$\dot{z} = xy - bz \tag{10}$$

where $R = 28$, $\sigma = 10$, and $b = 8/3$. We generate 9000 samples of the Lorenz system by solving the equations with fourth-order Runge–Kutta method with integral step 0.05. Fig. 6(a) shows that $V[m]$ peaks up at the same lag for all the state variables: since $x$ is coupled nonlinearly to $y$ and $z$ the periodic similarities in one state variable affect the other states and then any state of the system could be used to gain information about the process. This does not happen with the conventional auto-correlation, which for the state variables $y$ and $z$ do not show any (linear) correlation around the fifteenth lag. This example
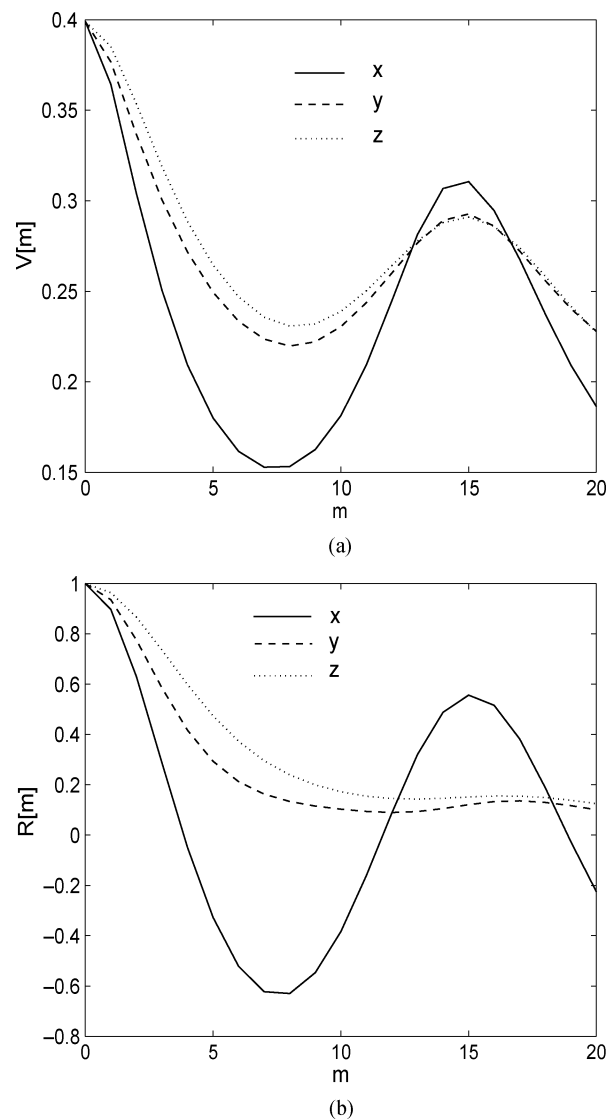
clearly shows that correntropy is able to extract non linear coupling information embedded in the time structure of time series, unlike the conventional autocorrelation.

## IV. APPLICATION TO BLIND EQUALIZATION

### A. Motivation to Use Correntropy

In digital communication systems, the transmitted signals are often distorted through a bandlimited channel which introduces intersymbol interference (ISI). Blind equalization refers to the problem of restoring the original digital signal when only the channel's output is available. To solve the problem, blind techniques exploit some knowledge about the statistical properties of the input signal or the structure of the channel [22].

Benveniste *et al.* [23] were the first to prove that a sufficient condition for perfect equalization is that the pdf of the recovered symbols be equal to the source pdf. According to this idea some blind algorithms aiming at forcing a given pdf at the output of the equalizer have been proposed [24]–[26]. Later, Shalvi and Weinstein relaxed this condition by showing that it is enough
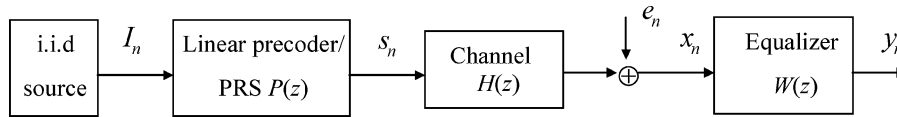
Fig. 7. Baud-rate digital communications system with linear precoding.

to match a few cumulants between the output signal and the source in order to achieve perfect equalization [27]. For instance, the so-called "super-exponential" algorithm [28] maximizes the kurtosis at the output of the channel subject to a constraint on the output variance. Other algorithms based on this result are the cumulant matching methods proposed by Tugnait [29], and also by Hatzinakos and Nikias [30].

To summarize, the theoretical results in [23] and [27] demonstrate that the higher order statistics are enough to solve the problem when the source signal is i.i.d. In some practical systems, however, the source symbol sequence presents some correlation (for instance in systems that use some type of precoding). In this situation, the performance of most of the existing blind techniques degrades drastically or they even fail to converge. To solve this problem, the proposed correntropy function enables us to derive new cost functions for blind equalization that exploits simultaneously our knowledge about both the pdf of the source and its time structure.

### B. Problem Setting

We consider a baud-rate sampled baseband representation of a linearly precoded digital communication system (see Fig. 7). An i.i.d. source $I_n$ is linearly precoded at the transmitter to form a sequence of correlated symbols $s_n$. The linear precoder can be a partial response system (PRS), typically introduced to make an efficient use of the available bandwidth [31]–[33] or any other type of line coding introducing correlation such as the alternate mark inversion (AMI), in which the logical "0" is represented by no line signal and the logical "1" is represented by positive and negative alternate pulses [31].

Now, let us suppose that the correlated source signal $s_n$ is sent through a linear time-invariant channel with coefficients $h_n$. The resulting channel output can be expressed as

$$x_n = \sum_i h_i s_{n-i} + e_n \qquad (11)$$

where $e_n$ is a zero-mean white Gaussian noise with variance $\sigma_e^2$. The objective of a blind linear equalizer is to remove the ISI at its output without using any training sequence. Typically, the equalizer is designed as a finite-impulse-response (FIR) filter with $M$ coefficients $\mathbf{w}$; then, its output is given by

$$y_n = \sum_{i=0}^{M-1} w_i x_{n-i} = \mathbf{w}^T \mathbf{x}_n.$$

The "super-exponential" algorithm proposed by Shalvi and Weinstein [27], [28] maximizes $|K_y|$, subject to the constraint $E[|y_n|^2] = E[|s_n|^2]$, where $K_y$ is the kurtosis at the output of the equalizer, which is defined as

$$K_y = E\left[|y_n|^4\right] - 2\left(E\left[|y_n|^2\right]\right)^2 - \left|E\left[y_n^2\right]\right|^2. \qquad (12)$$

The Shalvi–Weinstein algorithm is typically implemented as a batch technique, i.e., (12) is iteratively maximized using a block of collected data. Other blind techniques update the equalizer coefficients on a sample-by-sample basis using stochastic gradient descent algorithms. To this category belongs the family of Godard algorithms [34], which minimize the cost function

$$J_G(\mathbf{w}) = E\left[\left(|y_n|^p - R_p\right)^2\right], \quad p = 1, 2, \cdots \qquad (13)$$

where $R_p = E[|s_n|^{2p}]/E[|s_n|^p]$ depends on the input constellation. For the particular case $p = 2$, (13) is the cost function of the constant modulus algorithm (CMA), [34], [35]. Using a stochastic gradient descent approach, the CMA can be written as

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \mu\left(|y_n|^2 - R_2\right) y_n \mathbf{x}_n. \qquad (14)$$

As a measure of equalization performance we use the ISI defined as

$$\text{ISI} = 10\log_{10}\frac{\sum_n |\theta_n|^2 - \max_n |\theta_n|^2}{\max_n |\theta_n|^2}$$

where $\theta = \mathbf{h} * \mathbf{w}$ is the combined channel-equalizer impulse response, which is a delta function for a zero-forcing equalizer.

### C. Proposed Cost Function and Iterative Algorithm

Exploiting our knowledge about the linear precoder, first we must obtain the theoretical correntropy function for a given source signal. For instance, consider that the linear precoder used in Fig. 7 is given by $P(z) = 1 + z^{-1}$, which is the filter used in a duobinary modulation. If the input of this linear filter is a binary i.i.d. signal ($\pm 1$), then at its output we have a correlated symbol sequence drawn from the alphabet $\{-2, 0, +2\}$ with probabilities 1/4, 1/2, and 1/4, respectively, whose theoretical correntropy function is given by

$$V_s[m] = \begin{cases} \kappa(0), & m = 0 \\ \frac{1}{2}\kappa(0) + \frac{1}{2}\kappa(2), & m = \pm 1 \\ \frac{3}{8}\kappa(0) + \frac{1}{2}\kappa(2) + \frac{1}{8}\kappa(4), & |m| > 1 \end{cases}$$

where $\kappa(x)$ is the Gaussian kernel defined in (1).

Similarly, the theoretical correntropy function for the AMI line code (whose correlated symbol sequence is drawn from the alphabet $\{-1, 0, +1\}$) is given by

$$V_s[m] = \begin{cases} \kappa(0), & m = 0 \\ \frac{1}{4}\kappa(0) + \frac{1}{2}\kappa(1) + \frac{1}{4}\kappa(2), & m = \pm 1 \\ \frac{3}{8}\kappa(0) + \frac{1}{2}\kappa(1) + \frac{1}{8}\kappa(2), & |m| > 1 \end{cases}$$

For both examples, the theoretical correntropy function conveys information about the pdf and the correlation of the non-

white source signal. Therefore, we propose to use the following cost function for equalization:

$$J_{ce}(\mathbf{w}) = \sum_{m=1}^{P} (V_s[m] - V_y[m])^2 \qquad (15)$$

where $V_y[m]$ is the correntropy function estimated at the output of the equalizer and $P$ is the number of lags (notice that we have removed the zero lag since it is always equal to $\kappa(0)$).

The proposed cost function can be minimized either in a batch mode or in an adaptive setting by using a gradient descent approach. In a batch mode for all the iterations, we use the same set of collected data, while in an online mode we use a sliding window of a given length and update once per input sample. For both modes, the basic iteration of the algorithm is given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu \sum_{m=1}^{P} (V_s[m] - V_y[m]) \frac{\partial V_y[m]}{\partial \mathbf{w}}$$

where

$$\frac{\partial V_y[m]}{\partial \mathbf{w}} = -\frac{1}{N-m} \sum_{i=n-N-m+1}^{n} \kappa(y_i - y_{i-m}) \\ \times (y_i - y_{i-m})(\mathbf{x}_i - \mathbf{x}_{i-m})$$

and $N$ denotes the number of samples used to estimate the correntropy function.

### D. Simulation Results

*1) Batch Algorithm:* In the first example, we test the correntropy algorithm in a batch mode and compare its performance with the "super-exponential" algorithm proposed by Shalvi and Weinstein [28] (in the sequel, we will refer to this algorithm as SW). $N = 2000$ samples of a duobinary signal were generated with the linear precoder $P(z) = 1 + z^{-1}$ and then distorted by an FIR channel with 11 taps taken from [31] and representative of a telephone channel: $h_i = (0.04, -0.05, 0.07, -0.21, -0.5, 0.72, 0.36, 0, 0.21, 0.03, 0.07)$. White Gaussian noise is added to the channel output to get a given SNR.

A 21-tap equalizer was used and initialized with the center coefficient set to unity and the rest to zero. The correntropy cost function was estimated at $P = 10$ lags and iteratively minimized by applying 1000 iterations with a fixed step size $\mu = 0.25$. The kernel size to estimate the correntropy function was calculated after each iteration as the standard deviation of the equalizer's output $\sigma = \text{std}(y)$. After the blind stage, the algorithm switches to a decision-directed mode and 2000 additional iterations are carried out with a step size $\mu = 0.004$. Finally, the bit error rate (BER) is estimated by counting errors after transmitting $10^5$ or $10^6$ symbols, depending on the SNR. We run 50 independent simulations.

On the other hand, the SW algorithm was tested in a similar setting: first a pure blind mode and then a decision-directed mode. The SW algorithm requires a white source signal (strictly, the source must be fourth-order white), therefore when applied to the distorted duobinary signal it is unable to open the eye diagram. Since the linear precoder $P(z)$ is known, a reasonable
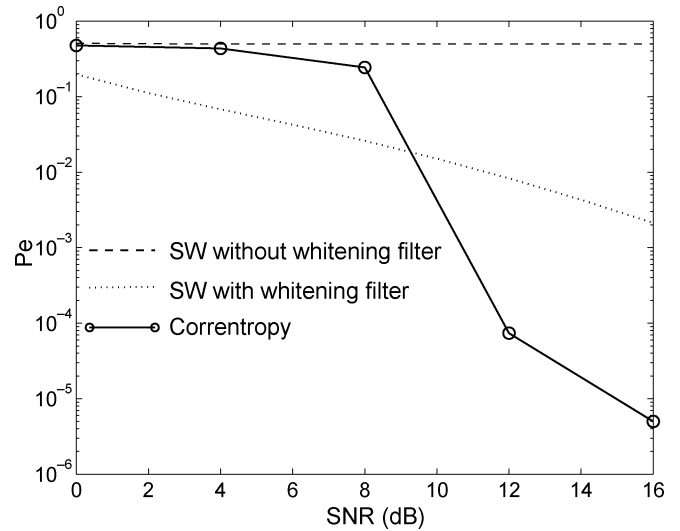


Fig. 8. Bit error rate versus SNR for correntropy, SW algorithm and SW algorithm with a whitening filter.

alternative would be to whiten the source signal by using the inverse of the linear precoder $P(z)^{-1}$ before the equalizer. The problem with this approach is that the precoders used in partial response systems such as $P(z) = 1 + z^{-1}$ cannot be inverted since they have zeros on the unit circle. To avoid this problem, we will use an approximate inverse FIR filter with 13 coefficients given by

$$G(z) = \frac{P^*(\omega)}{|P(\omega)|^2 + \Delta}$$

where $\Delta = 0.05$ was used. This type of "regularized" inverse has been used in [33] as a whitening step before the equalizer that restores the i.i.d. assumption of the source signal.

Fig. 8 compares the BER of the correntropy algorithm, the SW algorithm and the SW with a whitening filter. As it was discussed previously, the SW algorithm without whitening step fails to converge since the duobinary source is not i.i.d. On the other hand, for low SNRs, the SW with whitening filter achieves a lower BER than correntropy, whereas for moderate and high SNR's the correntropy algorithm provides the best solution.

*2) Online Algorithm:* For the second example, we test the correntropy algorithm in an on-line setting and compare its performance with the CMA algorithm. We consider an alternate mark inversion (AMI) source signal distorted by an IIR channel: $H_1(z) = 1/(1 - 0.5z^{-1})$, then white Gaussian noise is added for a final SNR = 20 dB. A 3-tap equalizer was used and initialized with the center coefficient set to unity and the rest to zero.

The correntropy function was estimated using a window of $N = 100$ samples and $P = 4$ lags. The kernel size was iteratively updated like in the previous example as the standard deviation of the equalizer's output for the current window. Fig. 9 shows the ISI curves for correntropy and CMA obtained by averaging 25 independent Monte Carlo simulations. The step-size for correntropy was $\mu = 0.6$, whereas for CMA was $\mu_{cma} = 0.04$: for both algorithms, these are the largest step-sizes for which all the trials converged while providing a similar converge speed as can be seen in Fig. 9.
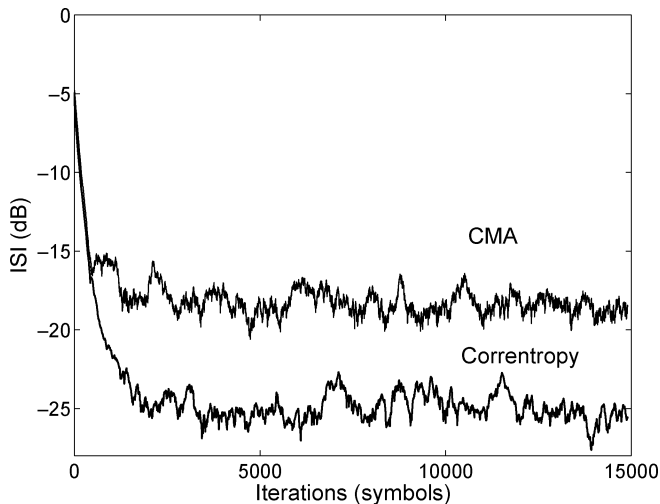
Fig. 9. ISI convergence curves for correntropy and CMA under Gaussian noise.



Fig. 10. ISI convergence curves for the correntropy and CMA under impulsive noise.

We see that for this example the correntropy function effectively extracts more information than the CMA about the source signals and it is able to provide a better solution. In order to explain the poor behavior of CMA for this particular case, we should remember that the use of a nonuniform symbol distribution has the effect of raising the kurtosis, thus making the pdf of the source distribution more Gaussian. Specifically, the kurtosis for the AMI signal of our example is $E[|s_n|^4]/(E[|s_n|^2])^2 = 2$, whereas for a uniform binary phase shift keying the kurtosis is 1, for a pulse amplitude modulation with 32 levels is 1.798 and for a Gaussian is 3. Although the source remains sub-Gaussian, the increase in kurtosis has the effect of lifting the CMA cost function (thus increasing the excess mean squared error) and flattening its surface (thus reducing the convergence speed) [36]. Moreover, the use of a correlated source can also cause major problems in the CMA convergence as it was also pointed out in [36].

Another interesting property of the correntropy function is its robustness against impulsive noise. This additional advantage is due to the fact that when an outlier is present, the inner product in the feature space computed via the Gaussian kernel tends to be zero (i.e., $\kappa(y_i - y_{i-k}) \approx 0$ when either $y_i$ or $y_{i-k}$ have a large value). To illustrate this point, we have used the same simulation example, but this time the channel output is distorted with a zero-mean white impulsive noise, $e_n$, generated according to the following Gaussian mixture model:

$$f(e_n) = \frac{\epsilon}{\sqrt{2\pi}\sigma_1} \exp - \left(\frac{e_n^2}{2\sigma_1^2}\right) + \frac{(1-\epsilon)}{\sqrt{2\pi}\sigma_2} \exp - \left(\frac{e_n^2}{2\sigma_2^2}\right)$$

where typically $\epsilon \ll 1$ and $\sigma_1^2 \gg \sigma_2^2$. Specifically, in our simulations, we used $\epsilon = 0.15$, $\sigma_1^2 = 50\sigma_2^2$ and $\sigma_2 = 0.12$. Fig. 10 shows the ISI curves obtained in this case. Now the step-size for correntropy was $\mu = 0.4$ and $\mu_{\text{cma}} = 0.001$. We can see that even for this small step-size, the CMA is not able to converge due to the large noise spikes at the channel output. On the other hand, with the correntropy function we obtain practically the same convergence behavior as in the Gaussian case.
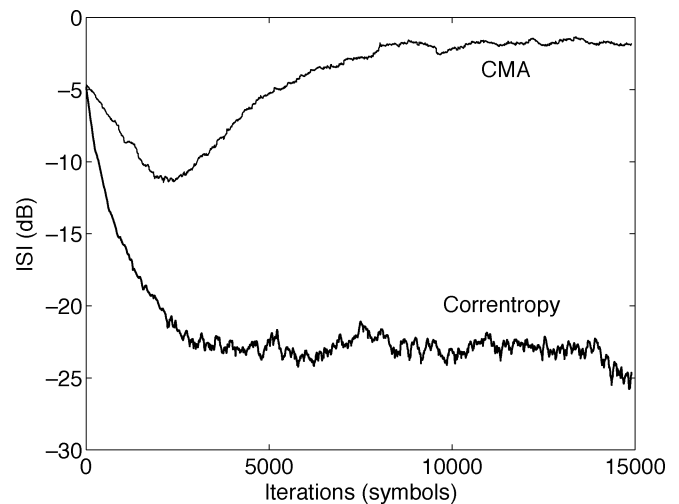
## V. CONCLUSION

In this paper, we have proposed a new reproducing kernel for stochastic processes that extracts information about its time structure as well as about its statistical distribution. Unlike the autocorrelation function, the correntropy function conveys information about the quadratic Renyi's entropy of the generating source (i.e., about its higher order moments). On the other hand, unlike most kernel methods and related ITL algorithms, the correntropy function is sensitive to the time structure of the process.

The intention of this paper is to present the definition of correntropy and outline some of its properties and possible applications. We were able to show that the correntropy satisfies a number of interesting properties: for instance it is a symmetrical positive definite function with a maximum at zero lag, and its mean value is the information potential. Therefore, a spectral density can be defined, and the correntropy matrix is positive definite. In perspective, the correntropy extends the autocorrelation function to nonlinear systems analysis and non-Gaussian sources. This is significant because the autocorrelation function, the workhorse of stochastic process analysis, is unable to quantify statistical properties beyond the second order moments, and cumulants are very difficult to estimate from finite data. Likewise, when dealing with nonlinear systems, the autocorrelation function is obviously not an appropriate measure, but as far as we know, there are no reasonable alternatives in the literature. The examples presented in the paper are simple, and for illustrative purposes only, but they convey the possible usefulness of correntropy. For instance, the Lorenz example shows that correntropy is sensitive to nonlinear coupling between the states of the system, while the autocorrelation function is not.

Much more work is necessary to fully characterize this new measure of similarity and fully exploit its use in model based signal processing. An example to a simple blind equalization problem using a coded signal is presented. We have shown that the "extra" information provided by the precoder in the time structure of the symbol sequence can be efficiently exploited by the correntropy function, outperforming the CMA algorithm as well as cumulant-based techniques. As an additional advantage

of the method, it has been shown to be robust against impulsive noise. Exploring potential applications of this measure to other problems such as nonlinear controls and nonlinear signal processing are interesting lines for further research.

The mathematical structure of RKHS seems the right tool to pursue the research in correntropy, its properties, and applications. Since the correntropy is a positive function that captures higher order moments of the random process, it defines an RKHS with different properties of the one defined through covariance functions. At the same time, the well-established links between RKHS and statistical estimation, as first outlined by Parzen, remain valid. This gives us hope that practical extensions to statistical estimation theory for non-Gaussian processes are possible and relatively straightforward to implement.

## ACKNOWLEDGMENT

## REFERENCES

[1] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[2] E. Parzen, "Statistical Methods on Time Series by Hilbert Space Methods," Applied Mathematics and Statistics Laboratory, Stanford Univ., Stanford, CA, Tech. Rep. 23, 1959.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[4] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neur. Comput.*, vol. 10, pp. 1299–1319, 1998.

[5] S. Mika, G. Rätsch, J. Wetson, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Int. Workshop Neural Networks Signal Processing IX*, Madison, WI, Aug. 1999, pp. 41–48.

[6] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neur. Comput.*, vol. 16, no. 12, pp. 2664–2699, Dec. 2004.

[8] J. C. Principe, D. Xu, and J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000.

[9] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1780–1786, Jul. 2002.

[10] K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind source separation using Renyi's mutual information," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 174–176, Jun. 2001.

[11] I. Santamaría, D. Erdogmus, and J. C. Principe, "Entropy minimization for supervised digital communications channel equalization," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1184–1192, May 2002.

[12] S. Saitoh, *Theory of Reproducing Kernels and its Applications*. Harlow, U.K.: Longman Scientific and Technical, 1988.

[13] G. Wahba, *Spline Models for Observational Data*, ser. CBMS-NSF Regional Conf. Series in Applied Mathematics. Philadelphia, PA: SIAM, 1990, vol. 59.

[14] R. Jenssen, D. Erdogmus, J. C. Principe, and T. Eltoft, "Toward a unification of information theoretic learning and kernel methods," in *Proc. 2004 IEEE Int. Workshop Machine Learning Signal Processing*, Sao Luis, Brazil, Sep. 2004, pp. 443–451.

[15] ——, "The Laplacian pdf distance: a cost function for clustering in a kernel feature space," presented at the 2004 Neural Information Processing Systems (NIPS), Vancouver, Canada, Dec. 2004.

[16] J.-W Xu, D. Erdogmus, R. Jenssen, and J. C. Principe, "An information-theoretic perspective to kernel independent component analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. V, Philadelphia, PA, Mar. 2005, pp. 249–252.

[17] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.

[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: John Hopkins Univ. Press, 1989.

[19] S. Kay, *Fundamentals of Statistical Signal Processing, Vol. I—Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] D. Erdogmus, K. E. Hild II, and J. C. Principe, "Online entropy manipulation: Stochastic information gradient," *IEEE Signal Process. Lett.*, vol. 10, no. 8, pp. 242–245, Aug. 2003.

[21] E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmosph. Sc.*, vol. 20, pp. 130–141, 1963.

[22] Z. Ding and Y. Li, *Blind Equalization and Identification*. New York: Marcel Dekker, 2001.

[23] A. Benveniste, M. Goursat, and G. Rouget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Trans. Automat. Control*, vol. 25, no. 3, pp. 385–399, Jun. 1980.

[24] J. Sala-Alvarez and G. Vazquez-Grau, "Statistical reference criteria for adaptive signal processing in digital communications," *IEEE Trans. Signal Process.*, vol. 45, no. 1, pp. 14–31, Jan. 1997.

[25] T. Adali, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Trans. Signal Process.*, vol. 45, pp. 1051–1064, Apr. 1997.

[26] M. Lazaro, I. Santamaria, C. Pantaleon, D. Erdogmus, and J. C. Principe, "Matched pdf-based blind equalization," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. IV, Hong Kong, China, Apr. 2003, pp. 297–300.

[27] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase channels," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 312–321, Mar. 1990.

[28] ——, "Super-exponential methods for blind deconvolution," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 504–519, Mar. 1993.

[29] J. K. Tugnait, "Blind estimation and equalization of digital communication FIR channels using cumulant matching," *IEEE Trans. Commun.*, vol. 43, no. 234, pp. 1240–1245, Feb./Mar./Apr. 1995.

[30] D. Hatzinakos and C. L. Nikias, "Blind equalization based on higher-order statistics (HOS)," in *Blind Deconvolution*, S. Haykin, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[31] J. Proakis, *Digital Communications*, NJ: Prentice-Hall, 1988.

[32] A. Lender, "Correlative (partial response) techniques and applications to digital radio systems," in *Digital Communications: Microwave Application*, K. Feher, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[33] J. K. Tugnait and U. Gummadevelli, "Blind equalization and channel estimation with partial response input signals," *IEEE Trans. Commun.*, vol. 45, no. 9, pp. 1025–1031, Sep. 1997.

[34] D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. 28, no. 11, pp. 1867–1875, Nov. 1980.

[35] J. R. Treichler and B. G. Agee, "A new approach to multipath correction of constant modulus signals," *Proc. IEEE Trans. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 31, pp. 349–372, Apr. 1983.

[36] C. R. Johnson Jr., P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown, and R. A. Casas, "Blind equalization using the constant modulus criterion: A review," *Proc. IEEE*, vol. 86, no. 10, pp. 1927–1950, Oct. 1998.
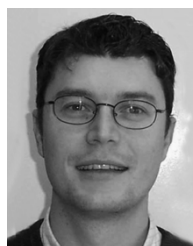
**Ignacio Santamaría** (M'96–SM'05) was born in Spain in 1967. He received the Telecommunication Engineer degree and the Ph.D. degree in electrical engineering from the Polytechnic University of Madrid, Spain, in 1991 and 1995, respectively.

In 1992, he joined the Departamento de Ingeniería de Comunicaciones, Universidad de Cantabria, Spain, where he is currently an Associate Professor. In 2000 and 2004, he spent visiting periods at the Computational NeuroEngineering Laboratory (CNEL), University of Florida. He has more than 90 publications in refereed journals and international conference papers. His current research interests include nonlinear modeling techniques, machine learning theories, and their application to digital communication systems, in particular to inverse problems (deconvolution, equalization, and identification problems).

**Puskal P. Pokharel** (S'04) was born in Lalitpur, Nepal, in 1981. He received the Bachelor of Technology degree in electronics and communication engineering from the Indian Institute of Technology (IIT), Roorkee, India, in 2003. He has been working towards the Ph.D. degree at the Electrical and Computer Engineering Department, University of Florida, Gainesville, under the supervision of J. C. Principe.

His current research interests include digital signal processing, machine learning, information theoretic learning, and their applications.
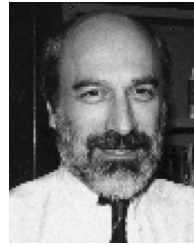
**Jose C. Principe** (M'83–SM'90–F'00) is Professor of Electrical and Computer Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing, machine learning, and artificial neural networks (ANNs) modeling. He is BellSouth Professor and the Founder and Director of the University of Florida Computational Neuro-Engineering Laboratory (CNEL). His primary area of interest is processing of time-varying signals with adaptive neural models. CNEL has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information). He has more than 100 publications in refereed journals, ten book chapters, and 200 conference papers. He directed 35 Ph.D. dissertations and 45 Master's theses. He recently wrote an interactive electronic book titled *Neural and Adaptive Systems: Fundamentals Through Simulation* (New York: Wiley).

Dr. Principe has served as the Chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society, Member of the Board of Governors of the International Neural Network Society, and Editor-in-Chief of the IEEE Transactions on Biomedical Engineering. He is a member of the Advisory Board of the University of Florida Brain Institute.