Correntropy: Properties and Applications in Non-Gaussian Signal Processing

Weifeng Liu, *Student Member, IEEE*, Puskal P. Pokharel, *Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

***Abstract*—The optimality of second-order statistics depends heavily on the assumption of Gaussianity. In this paper, we elucidate further the probabilistic and geometric meaning of the recently defined correntropy function as a localized similarity measure. A close relationship between correntropy and M-estimation is established. Connections and differences between correntropy and kernel methods are presented. As such correntropy has vastly different properties compared with second-order statistics that can be very useful in non-Gaussian signal processing, especially in the impulsive noise environment. Examples are presented to illustrate the technique.**

***Index Terms*—Generalized correlation function, information theoretic learning, kernel methods, metric, temporal principal component analysis (TPCA).**

I. Introduction

SECOND-ORDER statistics in the form of correlation and in particular the mean square error (MSE) are probably the most widely utilized methodologies for quantifying how similar two random variables are. Successful engineering solutions from these methodologies rely heavily on the Gaussianity and linearity assumptions. Recently, our group has extended the concept of mean square error adaptation to include information theoretic criteria [1], which has been named information theoretic learning (ITL). ITL preserves the nonparametric nature of correlation learning and MSE adaptation, i.e., the cost function is still directly estimated from data via a Parzen kernel estimator [2], but it extracts more information from the data for adaptation, and yields, therefore, solutions that are more accurate than MSE in non-Gaussian and nonlinear signal processing [3]–[8].

Inspired by ITL, we recently extended the fundamental definition of correlation function for random processes with a generalized correlation function called correntropy [9], which contains higher-order moments of the probability density function (pdf), but it is much simpler to estimate directly from samples than conventional moment expansions. The original definition only applies to a single random process, and it is more precisely called autocorrentropy. In this paper, we extend the definition to the general case of two arbitrary random variables and provide for the first time its probabilistic and geometric meaning. This theoretical framework will help understand and apply correntropy judiciously to nonlinear, non-Gaussian signal processing.

Manuscript received April 17, 2006; revised January 26, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. David J. Miller. This work was supported in part by the NSF by Grants ECS-0300340 and ECS-0601271.

The authors are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: weifeng@cnel.ufl.edu; pokharel@cnel.ufl.edu; principe@cnel.ufl.edu).

Digital Object Identifier 10.1109/TSP.2007.896065

We show that correntropy is directly related to the probability of how similar two random variables are in a neighborhood of the joint space controlled by the kernel bandwidth, i.e., the kernel bandwidth acts as a zoom lens, controlling the "observation window" in which similarity is assessed. This adjustable window provides an effective mechanism to eliminate the detrimental effect of outliers, and it is intrinsically different from the use of a threshold in conventional techniques.

Statistics estimated from data samples usually have a geometric meaning. For instance, MSE gives the 2-norm distance in sample space. We show that correntropy induces a new metric which is equivalent to the 2-norm distance if points are close, behaves similarly to the 1-norm distance as points get further apart and eventually approaches the zero-norm as they are far apart. This geometric interpretation elucidates the robustness of correntropy for outlier rejection.

The organization of the paper is as follows. After a very brief review of ITL and kernel methods to introduce the terminology, the definition and properties of correntropy are presented in Section III. In Section IV, the difference between MSE and correntropy is presented and the advantage of correntropy is theoretically explained based on its connection to robust M-estimation. Then in Section V, some examples are presented to corroborate our understanding and to inspire readers on how to apply correntropy to their research fields. Finally, Section VI summarizes the main conclusions and future lines of research.

II. Brief Background on ITL and Kernel Methods

ITL is a framework to nonparametrically adapt systems based on entropy and divergence. Renyi's quadratic entropy of a random variable X with pdf $f_X(x)$ is defined by

$$H_2(X) = -\log \int f_X^2(x)\,dx. \tag{1}$$

The Parzen estimate of the pdf, given a set of i.i.d. data $\{x_i\}_{i=1}^N$ drawn from the distribution, is

$$\hat{f}_{X;\sigma}(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(x - x_i) \tag{2}$$

where $\kappa_\sigma(x - x_i)$ is the Gaussian kernel

$$\kappa_\sigma(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x_i)^2}{2\sigma^2}\right) \tag{3}$$

where N is the number of the data points and σ the kernel size. For simplicity, we will drop the subscript and denote it as $\kappa(\cdot)$ when the meaning is clear from the context. The Gaussian kernel will be the only one considered in this paper (for other Mercer

kernels, most of the discussions still hold with minor modifications). The kernel size or bandwidth is a free parameter that must be chosen by the user using concepts of density estimation, such as Silverman's rule [23] or maximum likelihood. We have experimentally verified that the kernel size affects much less the performance of ITL algorithms than density estimation [24], but a thorough treatment of this issue is beyond the scope of this paper.

From the viewpoint of kernel methods, the kernel function (3) satisfies Mercer's Theorem [11], so that it induces a nonlinear mapping Φ which transforms data from the input space to an infinite dimensional reproducing kernel Hilbert space (RKHS) $\mathbf{F}$ where the following holds:

$$\kappa_\sigma(x - x_i) = \langle \Phi(x), \Phi(x_i) \rangle_{\mathbf{F}} \tag{4}$$

where $\langle \cdot, \cdot \rangle_{\mathbf{F}}$ denotes inner product in $\mathbf{F}$.

A nonparametric estimate of quadratic entropy directly from samples is obtained as [3]

$$\hat{H}_2(X) = -\log \mathrm{IP}(X) \tag{5}$$

$$\mathrm{IP}(X) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} \kappa_{\sqrt{2}\sigma}(x_j - x_i). \tag{6}$$

IP(X) stands for information potential (IP), and we have recently shown that it represents the mean square of the projected data, i.e., the first statistical moment of the RKHS data. It is therefore interesting to further define similarity measures in this space that are not hindered by the conventional moment expansions. Towards this goal we recently proposed a similarity measure called correntropy [9] defined for random processes.

Let $\{X(t), t \in T\}$ be a stochastic process with T being an index set. The nonlinear mapping Φ induced by the Gaussian kernel maps the data into the feature space $\mathbf{F}$, where the autocorrentropy function $V_X(t, s)$ is defined from $T \times T$ into $\mathbf{R}^+$ given by

$$\begin{aligned} V_X(t, s) &= \mathbf{E}[\langle \Phi(X(t)), \Phi(X(s)) \rangle_{\mathbf{F}}] \\ &= \mathbf{E}[\kappa_\sigma(X(t) - X(s))]. \end{aligned} \tag{7}$$

We call (7) the autocorrentropy function due to the analogy with the autocorrelation of random processes and the property that its average over the lags is the IP, i.e., the argument of Renyi's entropy [9]. We have shown that autocorrentropy is a symmetric, positive-definite function and therefore defines a new RKHS [10]. Based on autocorrentropy it is possible to derive the analytical solution of the optimal linear combiner in this space [12].

III. Definition and Properties of Cross Correntropy

A. Definition

A more general form of correntropy between two arbitrary scalar random variables is defined as follows.

Definition: Cross correntropy is a generalized similarity measure between two arbitrary scalar random variables X and Y defined by

$$V_\sigma(X, Y) = \mathbf{E}[\kappa_\sigma(X - Y)]. \tag{8}$$

In this paper it will be simply called correntropy. The extension of (8) to arbitrary dimensions will be addressed in a future work.

In practice, the joint pdf is unknown and only a finite number of data $\{(x_i, y_i)\}_{i=1}^N$ are available, leading to the sample estimator of correntropy

$$\hat{V}_{N,\sigma}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x_i - y_i). \tag{9}$$

B. Properties

Some important properties of correntropy are presented next. The first three are extensions of the properties presented in [9] and will, therefore, not be proved here.

Property 1: Correntropy is symmetric: $V(X, Y) = V(Y, X)$.

Property 2: Correntropy is positive and bounded: $0 < V(X, Y) \leq 1/\sqrt{2\pi}\sigma$. It reaches its maximum if and only if $X = Y$.

Property 3: Correntropy involves all the even moments of the random variable $E = Y - X$: $V_\sigma(X, Y) = (1)/(\sqrt{2\pi}\sigma) \sum_{n=0}^{\infty}((-1)^n)/(2^n n!)\mathbf{E}[((X - Y)^{2n})/(\sigma^{2n})]$. As σ increases, the high-order moments decay faster, so the second-order moment tends to dominate and correntropy approaches correlation. This has been verified in practice for kernel sizes 20 times larger than the value given by Silverman's rule for the data. Due to the expected value operator, the issue of kernel size selection in correntropy is different from density estimation. As will be practically demonstrated, the performance sensitivity of correntropy to the kernel size is much less than what could be expected from density estimation.

Property 4: Assume i.i.d. data $\{(x_i, y_i)\}_{i=1}^N$ are drawn from the joint pdf $f_{X,Y}(x, y)$, and $\hat{f}_{X,Y;\sigma}(x, y)$ its Parzen estimate with kernel size σ. The correntropy estimate with kernel size $\sigma' = \sqrt{2}\sigma$ is the integral of $\hat{f}_{X,Y;\sigma}(x, y)$ along the line $x = y$ [18]

$$\hat{V}_{\sqrt{2}\sigma}(X, Y) = \int_{-\infty}^{+\infty} \hat{f}_{X,Y;\sigma}(x, y)|_{x=y=u} du. \tag{10}$$

Proof: Using the two dimensional radially symmetric Gaussian kernel to estimate the joint pdf, we have

$$\hat{f}_{X,Y;\sigma}(x, y) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x_i \\ y_i \end{pmatrix} \right) \tag{11}$$

where

$$K_\sigma \left(\begin{pmatrix} x \\ y \end{pmatrix} \right) = \frac{1}{2\pi |\sum|^{1/2}} \exp \left(-\frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right) \tag{12}$$

with $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$.

It is easy to see that

$$K_\sigma \left(\begin{pmatrix} x \\ y \end{pmatrix} \right) = \kappa_\sigma(x) \cdot \kappa_\sigma(y). \tag{13}$$

So

$$\hat{f}_{X,Y;\sigma}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x-x_i)\cdot\kappa_\sigma(y-y_i). \tag{14}$$

Integrating (14) along the line $x=y$, we obtain

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\hat{f}_{X,Y;\sigma}(x,y)|_{x=y=u}du \\
&\quad=\int_{-\infty}^{+\infty}\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x-x_i)\cdot\kappa_\sigma(y-y_i)|_{x=y=u}du \\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\kappa_\sigma(u-x_i)\cdot\kappa_\sigma(u-y_i)du \\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\kappa_{\sqrt{2}\sigma}(x_i-y_i)=\hat{V}_{\sqrt{2}\sigma}(X,Y).
\end{aligned} \tag{15}$$

This completes the proof.

According to the conditions of the Parzen method [2], when σ goes to zero and the product $N\sigma$ to infinity, $\hat{f}_{X,Y;\sigma}(x,y)$ approaches the true pdf $f_{X,Y}(x,y)$, therefore we also have

$$\begin{aligned}
&\lim_{\sigma\to 0}V(X,Y) \\
&=\lim_{\sigma\to 0}\iint_{x,y}\kappa_\sigma(x-y)f_{XY}(x,y)\,dxdy \\
&=\iint_{x,y}\delta(x-y)f_{X,Y}(x,y)\,dxdy \\
&=\int_{x=-\infty}^{+\infty}f_{X,Y}(x,x)\,dx
\end{aligned} \tag{16}$$

In practical applications, the estimation of correntropy is done only with a finite number of samples, which sets a lower bound on the kernel size, since too small a kernel size will lead to meaningless estimation [2]. When the kernel size used in correntropy is σ, its rectangle approximation has a bandwidth $\sqrt{\pi/2}\sigma$ and let us assume that the joint pdf is smooth in this bandwidth. Then an alternative interpretation of correntropy is

$$V_\sigma(X,Y)\approx P(|Y-X|<\sqrt{\pi/2}\sigma)/\sqrt{2\pi}\sigma. \tag{17}$$

Property 5: Assume the samples $\{(x_i,y_i)\}_{i=1}^N$ are drawn from the joint pdf $f_{X,Y}(x,y)$. Define the error random variable $E=Y-X$, and $\hat{f}_{E;\sigma}(e)$ as the Parzen estimate of the error pdf from data $\{(e_i=x_i-y_i)\}_{i=1}^N$. Then $\hat{V}_\sigma(X,Y)$ is the value of $\hat{f}_{E;\sigma}(e)$ evaluated at the point $e=0$, i.e.

$$\hat{V}_\sigma(X,Y)=\hat{f}_{E;\sigma}(0). \tag{18}$$

Proof: By (9)

$$\begin{aligned}
\hat{V}_\sigma(X,Y)&=\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_i-y_i) \\
&=\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(e_i)=\hat{f}_{E;\sigma}(0)
\end{aligned} \tag{19}$$

which completes the proof.

It is important to study the statistical properties of the correntropy estimator. First note that

$$\begin{aligned}
V_\sigma(X,Y)&=\mathbf{E}[\kappa_\sigma(X-Y)]=\mathbf{E}[\kappa_\sigma(E)] \\
&=\int_e\kappa_\sigma(e)f_E(e)\,de.
\end{aligned} \tag{20}$$

It is also obvious that

$$f_E(0)=p(X=Y)=\int_{x=-\infty}^{+\infty}f_{X,Y}(x,x)\,dx. \tag{21}$$

Now, the study of the mean and variance of $\hat{V}_{N;\sigma}(X,Y)$ is quite straightforward in the context of Parzen estimation [2]. Indeed

$$\mathbf{E}[\hat{V}_{N,\sigma}(X,Y)]=V_\sigma(X,Y) \tag{22}$$

$$\lim_{N\to\infty,\sigma\to 0}\mathbf{E}[\hat{V}_{N,\sigma}(X,Y)]=f_E(0) \tag{23}$$

$$\mathbf{var}[\hat{V}_{N,\sigma}(X,Y)]=N^{-1}\mathbf{var}[\kappa_\sigma(E)] \tag{24}$$

$$\begin{aligned}
&\lim_{N\to\infty,\sigma\to 0}N\sigma\mathbf{var}[\hat{V}_{N,\sigma}(X,Y)] \\
&\quad=f_E(0)\int_{-\infty}^{\infty}(\kappa_1(z))^2dz
\end{aligned} \tag{25}$$

where $\kappa_1(z)$ is the Gaussian kernel with $\sigma=1$.

Among these properties, (25) is the most important and its proof can be essentially found in [2]. Therefore, under the condition of $N\to\infty$, $\hat{V}_{N,\sigma}(X,Y)$ is an unbiased estimator of $V_\sigma(X,Y)$ and consistent in mean square. Further, under the conditions $N\sigma\to\infty$ and $\sigma\to 0$, $\hat{V}_{N,\sigma}(X,Y)$ is an asymptotically unbiased estimator of $f_E(0)$ and consistent in mean square.

Let us assume the error pdf is Gaussian, i.e.

$$f_E(e)=\frac{1}{\sqrt{2\pi}\sigma_E}\exp\left(-\frac{e^2}{2\sigma_E^2}\right). \tag{26}$$

where σ_E is the standard deviation of the error.

$$V_\sigma=1\Big/\sqrt{2\pi\left(\sigma^2+\sigma_E^2\right)} \tag{27}$$

$$V_0=:\lim_{\sigma\to 0}V_\sigma=1/\sqrt{2\pi}\sigma_E. \tag{28}$$

A simple calculation shows $|(V_\sigma-V_0)/V_0|<0.05$ if $\sigma<0.32\sigma_E$. In the case of $\sigma^2\ll\sigma_E^2$, we obtain

$$(V_0-V_\sigma)/V_0\approx\sigma^2/\left(2\sigma_E^2\right). \tag{29}$$

This weak sensitivity of the correntropy estimator with respect to σ contrasts with the large dependence of the density estimation w.r.t. σ [23], and can be understood by the fact that correntropy is a "central moment" in joint space.

Further, if the assumption (26) holds in regression problems, maximizing V_σ in (27) is essentially minimizing the error variance σ_E. σ becomes immaterial provided the variance of the estimator of (25) is reasonably upper-bounded. For instance, given N we can choose σ so that the following condition holds:

$$|\sqrt{\mathbf{var}(\hat{V}_{N,\sigma})}/\mathbf{E}(\hat{V}_{N,\sigma})|<0.05. \tag{30}$$

Asymptotically (N is sufficiently large and σ sufficiently small according to (25)) we have

$$|\sqrt{\mathbf{var}(\hat{V}_{N,\sigma})}/\mathbf{E}(\hat{V}_{N,\sigma})| \approx \sqrt{(N\sigma f_E(0))^{-1}\int_{-\infty}^{\infty}(\kappa_1(z))^2 dz}. \quad (31)$$

Property 6: Correntropy is a second-order statistic of the mapped feature space data.

Proof: Assume the dimension of the feature space is M (eventually infinite as in the case of Gaussian kernel) and the kernel mapping is $\Phi(X) = [\varphi_1(X)\ \varphi_2(X)\ldots\varphi_M(X)]^T$.

The second-order statistics between $\Phi(X)$ and $\Phi(Y)$ is expressed by the following correlation matrix:

$$\mathbf{R}_{XY} = \mathbf{E}[\Phi(X)\Phi(Y)^T] = \begin{bmatrix} \mathbf{E}[\varphi_1(X)\varphi_1(Y)] & \cdots & \mathbf{E}[\varphi_1(X)\varphi_M(Y)] \\ \vdots & \ddots & \vdots \\ \mathbf{E}[\varphi_M(X)\varphi_1(Y)] & \cdots & \mathbf{E}[\varphi_M(X)\varphi_M(Y)] \end{bmatrix}_{M\times M}. \quad (32)$$

Meanwhile

$$V(X,Y) = \mathbf{E}[\Phi(X)^T\Phi(Y)] = \mathrm{trace}(\mathbf{R}_{XY}). \quad (33)$$

The trace of $\mathbf{R}_{XY}$ is equal to the sum of the eigenvalues, which clearly shows that correntropy is a second-order statistic in the feature space induced by the Gaussian kernel. This property should be contrasted with the cross variance operator defined in kernel methods [22] and shows that correntropy is its trace (assuming centered data in RKHS).

Property 7: If X and Y are statistically independent

$$V(X,Y) = \langle \mathbf{E}[\Phi(X)], \mathbf{E}[\Phi(Y)]\rangle_{\mathbf{F}}. \quad (34)$$

Proof: Using the notation in property 6

$$\begin{aligned} V(X,Y) &= \mathbf{E}\left[\sum_{i=1}^{M}\varphi_i(X)\varphi_i(Y)\right] \\ &= \sum_{i=1}^{M}\mathbf{E}[\varphi_i(X)]\mathbf{E}[\varphi_i(Y)] \\ &= \langle \mathbf{E}[\Phi(X)], \mathbf{E}[\Phi(Y)]\rangle_{\mathbf{F}} \end{aligned} \quad (35)$$

by the independence assumption. This completes the proof.

This property can be called uncorrelatedness in feature space and is a new, easily computable measure of independence between X and Y. Additionally, this property can be interpreted in terms of pdf. If X and Y are independent

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \quad (36)$$

Using Parzen window to estimate these pdfs

$$\hat{f}_{X,Y;\sigma}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x-x_i)\cdot\kappa_\sigma(y-y_i) \quad (37)$$

$$\hat{f}_{X;\sigma}(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x-x_i) \quad (38)$$

$$\hat{f}_{Y;\sigma}(y) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(y-y_i). \quad (39)$$

Integrating (36) along the line $x = y$ and using (37)–(39) yields

$$\frac{1}{N}\sum_{i=1}^{N}\kappa_{\sqrt{2}\sigma}(x_i-y_i) \approx \frac{1}{N^2}\sum_{j}^{N}\sum_{i}^{N}\kappa_{\sqrt{2}\sigma}(x_j-y_i) \quad (40)$$

which is a sample estimate approximation of (34). The approximation in (40) is due to the Parzen estimates. When σ tends to zero and the product $N\sigma$ to infinity, strict equality holds. Using the analogy of potential fields, the term on the right-hand side (RHS) of (40) is called the cross information potential (CIP). When $X = Y$, it reduces to IP. From the viewpoint of kernel methods, $f_X(\cdot) = \mathbf{E}[\Phi(X)]$, $f_Y(\cdot) = \mathbf{E}[\Phi(Y)]$ are two points in the RKHS, and CIP is exactly the inner product between the vectors created by these two pdfs.

Equation (34) bears resemblance to the constrained covariance proposed by Gretton *et al.* in [22], which is a strong measure of independence according to the work of Jacod and Protter on independence characterization through covariance operator in function spaces [22]. These authors constrained the covariance operator in a closed ball of a reproducing kernel Hilbert space and converted the measure into a matrix norm of Gram matrices. However, our measure starts directly from Parzen estimates of pdfs and is a much simpler, while possibly weaker, measure of independence. Further analysis and applications of this property in independent component analysis (ICA) will be pursued in future work.

Property 8: Correntropy, as a sample estimator, induces a metric in the sample space. Given two vectors $X = (x_1, x_2, \ldots, x_N)^T$ and $Y = (y_1, y_2, \ldots, y_N)^T$ in the sample space, the function $\mathrm{CIM}(X,Y) = (\kappa(0) - V(X,Y))^{1/2}$ defines a metric in the sample space and is named as the correntropy induced metric (CIM).

Proof: This property highlights the geometric meaning of correntropy in the sample space. To be a metric CIM must obey the following properties:

1) Nonnegativity. $\mathrm{CIM}(X,Y) \geq 0$ by Property 2.
2) Identity of indiscernibles. $\mathrm{CIM}(X,Y) = 0$ if and only if $X = Y$ by Property 2.
3) Symmetric by Property 1.
4) Triangle inequality: $\mathrm{CIM}(X,Z) \leq \mathrm{CIM}(X,Y) + CIM(Y,Z)$. The proof is based on the kernel mapping and a vector construction in a feature space which is a well defined Hilbert space. For X and Y, we construct two new vectors $\tilde{X} = [\Phi(x_1); \Phi(x_2); \ldots; \Phi(x_N)]$ and $\tilde{Y} = [\Phi(y_1); \Phi(y_2); \ldots; \Phi(y_N)]$ in the Hilbert space F^N. The Euclidean distance $\mathrm{ED}(\tilde{X}, \tilde{Y})$ is

$$\begin{aligned} \mathrm{ED}(\tilde{X},\tilde{Y}) &= (\langle(\tilde{X}-\tilde{Y}),(\tilde{X}-\tilde{Y})\rangle)^{1/2} \\ &= (\langle\tilde{X},\tilde{X}\rangle - 2\langle\tilde{X},\tilde{Y}\rangle + \langle\tilde{Y},\tilde{Y}\rangle)^{1/2} \\ &= \left(\sum_{i=1}^{N}\kappa(x_i-x_i) - 2\sum_{i=1}^{N}\kappa(x_i-y_i) + \sum_{i=1}^{N}\kappa(y_i-y_i)\right)^{1/2} \\ &= [2N\cdot(\kappa(0) - V(X,Y))]^{1/2} \\ &= \sqrt{2N}\cdot\mathrm{CIM}(X,Y). \end{aligned} \quad (41)$$

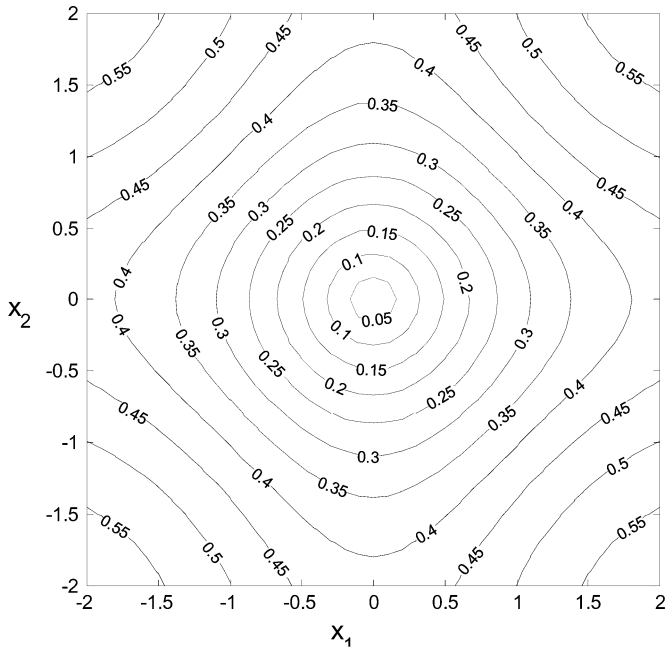


Fig. 1. Contours of $\mathrm{CIM}(X,0)$ in 2-D sample space (kernel size is set to 1).

Therefore

$$\begin{aligned}\mathrm{CIM}(X,Z) &= \mathrm{ED}(\tilde{X},\tilde{Z})/\sqrt{2N} \\ &\le \mathrm{ED}(\tilde{X},\tilde{Y})/\sqrt{2N} + \mathrm{ED}(\tilde{Z},\tilde{Y})/\sqrt{2N} \\ &= \mathrm{CIM}(X,Y) + \mathrm{CIM}(Y,Z). \end{aligned} \tag{42}$$

This completes the proof.

It can also be shown that this metric is translation invariant for translation invariant kernels like the Gaussian kernel, so we can denote $V(X,Y)$ as $V(Y-X)$. However, CIM is not homogenous so it does not induce a norm on the sample space. Fig. 1 shows the contours of the distance from X to the origin in a 2-D space. The interesting observation is as follows: when two points are close, CIM behaves like an L2 norm (which is clear from the Taylor expansion of property 3) and we call this area the Euclidean zone; outside of the Euclidean zone CIM behaves like an L1 norm which is named the Transition zone; eventually in the Rectification zone as two points are further apart, the metric saturates and becomes insensitive to distance (approaching L0 norm). This property inspired us to investigate the inherent robustness of CIM. Another important observation is that the kernel bandwidth σ controls the scale of the CIM norm. A small kernel size leads to a tight linear (Euclidean) region and to a large L0 region, while a larger kernel size will enlarge the linear region. Note also that for points far away from the origin, the metric becomes radially anisotropic, i.e., the distance becomes dependent on the direction. It is remarkable that a *single* parameter has such a tremendous impact on the evaluation of radial distances from any given point unlike what happens in the more traditional L_p norms. While bringing flexibility, there is still a need to choose appropriately the kernel size in practical applications.

Property 9: Let $\{x_i\}_{i=1}^N$ be a data set. The correntropy kernel induces a scalar nonlinear mapping η which maps the signal as $\{\eta_x(i)\}_{i=1}^N$ while preserving the similarity measure in the sense

$$\begin{aligned} E[\eta_x(i)\cdot\eta_x(i+t)] &= V(i,i+t) \\ &= \mathbf{E}[\kappa(x(i)-x(i+t))], \quad 0 \le t \le N-1. \end{aligned} \tag{43}$$

The square of the mean of the transformed data is an asymptotic estimate of the information potential of the original data as $N \to \infty$.

Proof: The existence of this nonlinear mapping η is proved in [12] using results from [10]. Here we prove the second part of the property. Denote m_η as the mean of the transformed data

$$m_\eta = \frac{1}{N}\sum_{i=1}^{N}\eta_x(i). \tag{44}$$

Therefore

$$m_\eta^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\eta_x(i)\eta_x(j) \tag{45}$$

Rewriting (43) in the sample estimate form, asymptotically we have

$$\sum_{i=1}^{N-t}\eta_x(i)\cdot\eta_x(i+t) = \sum_{i=1}^{N-t}\kappa(x(i)-x(i+t)) \tag{46}$$

with fixed t and as $N \to \infty$.

We arrange the double summation (45) as an array and sum along the diagonal direction which yield exactly the autocorrelation function of the transformed data at different lags, thus the correntropy function of the input data at different lags, i.e.

$$\begin{aligned} &\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\eta_x(i)\eta_x(j) \\ &\quad= \frac{1}{N^2}\left(\sum_{t=0}^{N-1}\sum_{i=1}^{N-t}\eta_x(i)\eta_x(i+t) + \sum_{t=1}^{N-1}\sum_{i=1+t}^{N}\eta_x(i)\eta_x(i-t)\right) \\ &\quad\approx \frac{1}{N^2}\left(\sum_{t=0}^{N-1}\sum_{i=1}^{N-t}\kappa(x(i)-x(i+t)) + \sum_{t=1}^{N-1}\sum_{i=1+t}^{N}\kappa(x(i)-x(i-t))\right) \\ &\quad= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x(i)-x(j)). \end{aligned} \tag{47}$$

As observed in (47), when the summation indices are far from the main diagonal, smaller and smaller data sizes are involved which leads to poorer approximations. Notice that this is exactly the same problem when the autocorrelation function is estimated from windowed data. As N approaches infinity, the estimation error goes to zero asymptotically. In other words, the mean of the transformed data induced by the correntropy kernel

asymptotically estimates the square root of the information potential and thus the entropy of the original data. This property corroborates further the name of correntropy given to this similarity measure.

Property 10: The autocorrentropy function defined by (7) is a reproducing kernel.

This property is proved in [9] and will be only interpreted here. Correntropy can be interpreted in two vastly different feature spaces. One is the RKHS induced by the Gaussian kernel, which is widely used in kernel machine learning. The elements on this RKHS are infinite dimensional vectors expressed by the eigenfunctions of the Gaussian kernel [11], and they lie on the first quadrant of a sphere since $\|\Phi(x)\|^2 = \kappa(0) = 1/\sqrt{2\pi}\sigma$ [9]. Correntropy performs statistical inference on the projected samples in this sphere. A full treatment of this view requires a differential geometry approach to take advantage of the manifold properties. The second feature space is the RKHS induced by the correntropy kernel itself where elements are random variables and the inner product is defined by the correlation [10]. With this interpretation, correntropy can be readily used for statistical inference, and provides a straightforward way to apply conventional optimal algorithms based on inner products in a RKHS that is nonlinearly related to the input space. Solutions of a nonlinear (with respect to the input space) Wiener filter [12] and of a nonlinear minimum average correlation energy (MACE) filter [25] have already been achieved. The problem is that products of transformed samples must be approximated by the kernel evaluated at the samples to implement the computation, which is only valid in the mean (see property 9). This limits performance but the results are very promising.

In this section, for the first time a detailed explanation of the probabilistic meanings of correntropy and an analysis of the mean-variance of the correntropy estimator are presented in properties 4 and 5. Properties 6 and 7 show interpretations of correntropy in connection to kernel methods. Property 8 is one of the main results in this paper, defining a new metric in sample space which is the mathematical foundation for regression applications in Section V. The proof in property 9 is also novel and together with property 10 it will be used to derive correntropy temporal principal component analysis (TPCA) in the following section.

IV. Comparison Between MSE and Correntropy

Let X and Y be two random variables and $E = Y - X$. $\mathrm{MSE}(X,Y)$ is defined as

$$\begin{aligned}\mathrm{MSE}(X,Y) &= \mathbf{E}[(X-Y)^2] \\ &= \iint_{x,y} (x-y)^2 f_{XY}(x,y)dxdy = \int_e e^2 f_E(e)de\end{aligned} \tag{48}$$

whereas

$$\begin{aligned}V(X,Y) &= \mathbf{E}[\kappa(X-Y)] \\ &= \iint_{x,y} \kappa(x-y) f_{XY}(x,y)dxdy = \int_e \kappa(e) f_E(e)de.\end{aligned} \tag{49}$$

Notice that the MSE is a quadratic function in the joint space with a valley along the $x = y$ line. Since similarity quantifies how different X is from Y in probability, this intuitively explains why MSE is a similarity measure in the joint space. However, the quadratic increase for values away from the $x = y$ line has the net effect of amplifying the contribution of samples that are far away from the mean value of the error distribution and it is why Gaussian distributed residuals provide optimality for the MSE procedure. But it is also the reason why other data distributions will make the MSE nonoptimal, in particular if the error distribution has outliers, is non-symmetric, or has nonzero mean.

Comparing MSE with correntropy, we conclude that these two similarity measures are assessing similarity in rather different ways: Correntropy is *local* whereas MSE is *global*. By global, we mean that all the samples in the joint space will contribute appreciably to the value of the similarity measure while the locality of correntropy means that the value is primarily dictated by the kernel function along the $x = y$ line. Therefore, correntropy of the error (51) can be used as a new cost function for adaptive systems training, which will be called the maximum correntropy criterion (MCC). MCC has the advantage that it is a local criterion of similarity and it should be very useful for cases when the measurement noise is nonzero mean, non-Gaussian, with large outliers. It is also easier to estimate than the MEE criterion proposed in [1].

Furthermore, we can put MCC in a more general framework by showing that it bears a close relationship with M-estimation [20]. M-estimation is a generalized maximum likelihood method proposed by Huber to estimate parameters θ under the cost function $\min_\theta \sum_{i=1}^N \rho(e_i \mid \theta)$, where ρ is a differentiable function satisfying:

1) $\rho(e) \geq 0$;
2) $\rho(0) = 0$;
3) $\rho(e) = \rho(-e)$;
4) $\rho(e_i) \geq \rho(e_j)$ for $|e_i| > |e_j|$.

In the case of adaptive systems, θ is a set of adjustable parameters and e_i are errors produced by the system during supervised learning. This general estimation is also equivalent to a weighted least square problem as

$$\min_\theta \sum_{i=1}^N w(e_i) e_i^2. \tag{50}$$

The weight function $w(e)$ is defined by $w(e) = \rho'(e)/e$ where ρ' is the derivative of ρ.

Defining $\rho(e) = (1 - \exp(-e^2/2\sigma^2))/\sqrt{2\pi}\sigma$ it is easy to see that ρ satisfies all the conditions listed above. Moreover, it corresponds to the kernel of the error as can be easily shown as

$$\begin{aligned}\min_\theta \sum_{i=1}^N \rho(e_i) &= \min_\theta \sum_{i=1}^N \left(1 - \exp\left(-e_i^2/2\sigma^2\right)\right)/\sqrt{2\pi}\sigma \\ &\Leftrightarrow \max_\theta \sum_{i=1}^N \exp\left(-e_i^2/2\sigma^2\right)/\sqrt{2\pi}\sigma \\ &= \max_\theta \sum_{i=1}^N \kappa_\sigma(e_i).\end{aligned} \tag{51}$$

The weighting function in this case is

$$w(e) = \exp(-e^2/2\sigma^2)/\sqrt{2\pi}\sigma^3. \tag{52}$$

For comparison, the weighting function of Bi-square is

$$w_{\mathrm{Bi}}(e) = \begin{cases} [1-(e/h)^2]^2 & |e| \le h \\ 0 & |e| > h \end{cases} \tag{53}$$

where h is a tuning constant. It turns out that the square of the Taylor expansion of (52) to the first-order is the weight function of Bi-square and the kernel size σ serves as the tuning constant in Bi-square. Notice that the weighting function is solely determined by the choice of ρ in the cost function and does not depend on the adaptive system. For example, the MSE cost function uses a constant weighting function. In that sense, the Gaussian like weighting function attenuates the large error terms so that outliers would have a less impact on the adaptation.

It is the first time a close relationship between M-estimation and methods of ITL is established, although their superior performances in impulsive environments were repeatedly reported [7]–[9]. It is also interesting that there is no threshold in correntropy. The kernel size controls all the properties of the estimator. Moreover, this connection may provide one practical way to choose an appropriate kernel size for correntropy.

V. Applications

A. Robust Regression

In the first example, we consider the general model of regression $Y = f(X) + Z$ where f is an unknown function, Z is a noise process and Y is the observation. A parametric approximator $g(x;w)$ (specified below) is used to discover this function and alleviate the effect of noise as much as possible. Let the noise probability density function be an impulsive Gaussian mixture $p_Z(z) = 0.9 \times \mathbf{N}(0, 0.1) + 0.1 \times \mathbf{N}(4, 0.1)$.

In MSE, the optimal solution is found by

$$\min_w J(w) = \frac{1}{M}\sum_{i=1}^{M}(g(x_i;w) - y_i)^2. \tag{54}$$

Here $\{(x_i, y_i)\}_{i=1}^{M}$ are the training data. Under the maximum correntropy criterion (MCC), the optimal solution is found by

$$\max_w J(w) = \frac{1}{M}\sum_{i=1}^{M}\kappa_\sigma(g(x_i;w) - y_i). \tag{55}$$

The first example uses a first degree polynomial system for simplicity, i.e., $g(x;w) = w_1 x + w_2$. $f(x) = ax + b$ with $a = 1$ and $b = 0$. Since the ultimate performance of the MCC criterion is under investigation, the kernel size is chosen by systematically searching for the best result using either *a priori* knowledge of the noise distribution, or simply scanning when the problem justifies this solution. Performance sensitivity with respect to kernel size will be quantified experimentally in Tables and will be compared with the kernel size estimated by Silverman's rule, one of the most widely used kernel density estimation heuristics. The data length is set small on purpose

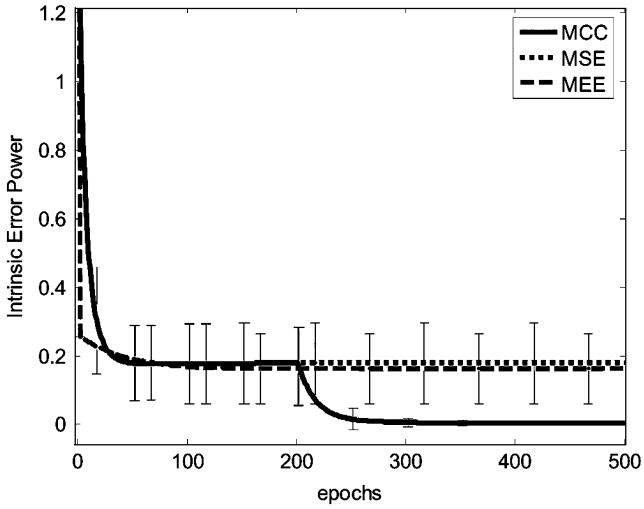


Fig. 2. Average learning curves with error bars of MCC, MSE, and MEE.

$M = 100$. Steepest descent is used for both criteria. Under the MSE criterion, the learning rate is set to 0.001 and the system is trained for 500 epochs (long enough to guarantee it reaches its global solution). In the MCC experiment, we first train the system with MSE criterion during the first 200 epochs (which is equivalent to kernel size annealing [1] as shown in Properties 3 and 8), and switch the criterion to MCC during the next 300 epochs. The learning rate is set to 0.001 and the kernel size is 0.5 which performs best on test data. We run 50 Monte Carlo simulations for the same data with 50 different starting points. The average estimated coefficients for MSE are [0.484 0.679] and [0.020 0.983] for MCC. The average learning curves are shown in Fig. 2, along with its standard deviation. For comparison, we also include the result of minimum error entropy (MEE) [1] with the bias set at the mean of the desired response. MEE is independent of the mean of the distribution, but with the bias set as explained it is also sensitive to the nonzero mean noise.

When MSE criterion is used, $g(x)$ is shifted by the nonzero-mean noise and slanted by the outliers due to the global property of MSE (Fig. 3). Now we see the importance of correntropy with its local property. In other words, correntropy has the ability of being insensitive to the peak in the noise pdf tail, and effectively handle the bulk of residuals around the origin (Property 5).

Although the main purpose of this example is to highlight the robustness of correntropy, we feel obligated to compare with the existing robust fitting methods such as least absolute residuals (LAR) and bi-square weights (BW) [17]. The parameters of these algorithms are the recommended settings in MATLAB. Another set of 50 Monte Carlo simulations are run with different noise realizations and different starting points. All the results are summarized in Table I in terms of intrinsic error power (IEP) which is estimated as $E[(g(X;w) - f(X))^2]$ on the test set. All the results in the tables are in the form of "average $\pm$ standard deviation." Notice that the intrinsic error power compares the difference between the model output and the true system (without the noise added). The performance of MCC is much better than LAR, and when regarded as a L1 norm alternative,

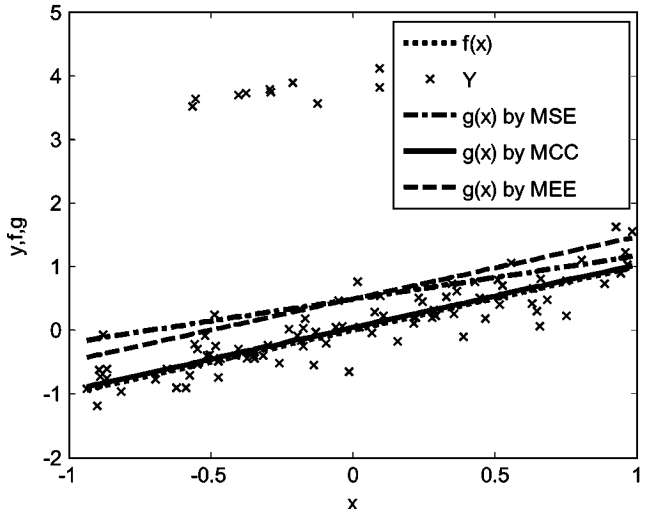


Fig. 3. Regression results with criteria of MSE, MCC, and MEE, respectively. The observation Y is corrupted with positive impulsive noise; the fit from MSE (dash-dot line) is shown shifted and skewed; the fit from MCC (solid line) matches the desired (dotted) quite well; the fit from MEE is shifted but not skewed.

TABLE I
REGRESSION RESULTS SUMMARY

	a	b	IEP
MSE	1.0048±0.1941	0.3969±0.1221	0.1874±0.1121
MCC	0.9998±0.0550	0.0012±0.0355	0.0025±0.0026
MEE	0.9964±0.0546	0.3966±0.1215	0.1738±0.1049
LAR	1.0032±0.0861	0.0472±0.0503	0.0072±0.0066
BW	1.0007±0.0569	0.0010±0.0359	0.0025±0.0025

TABLE II
EFFECTS OF KERNEL SIZE ON MCC

σ	IEP
0.1	0.0511±0.0556
0.2	0.0059±0.0050
0.5	0.0024±0.0024
1.0	0.0022±0.0022
2.0	0.0113±0.0109
3.0	0.0502±0.0326
4.0	0.0950±0.0558
Silverman's	0.0053±0.0052

correntropy is differentiable everywhere and to every order. Furthermore, notice that there is no threshold for MCC, just the selection of the kernel size. Moreover, the algorithm complexity of MEE is $O(M^2)$ whereas MSE, MCC, BS, and LAR are all $O(M)$.

Next, the effect of the kernel size on MCC is demonstrated. We choose seven kernel sizes: 0.1, 0.2, 0.5, 1, 2, 3, and 4. For each kernel size, 50 Monte Carlo simulations with different noise realizations are run to estimate the intrinsic error power and its standard deviation. The results are presented in Table II. MCC performs very well when the kernel size is in the range of $[0.2, 2]$. Intuitively, if the data size is large, a small kernel size shall be used so that MCC searches with high precision (small estimator bias) for the maximum position of the error pdf. However, if the data size is small, the kernel size has to be chosen as a compromise between estimation efficiency (small estimator variance) and outlier rejection. Nevertheless, MCC using large kernel sizes will perform no worse than MSE due to correntropy unique metric structure as shown in property 8. If one sets the kernel size by applying Silverman's rule to the error signal of each iteration [23]

$$\sigma_{Sm} = 1.06 * \min\{\sigma_E, R/1.34\} * M^{-1/5} \tag{56}$$

where σ_E is the standard deviation of the error and R is the error interquartile range, σ_{Sm} varies between $[1.8, 3]$ during the adaptation, which is in the neighborhood of the best values as shown in Table II.

TABLE III
EFFECTS OF THE MEAN OF THE OUTLIERS

Outliers mean	IEP by MCC	IEP by BW	IEP by LAR
0.2	0.0026±0.0021	0.0023±0.0019	0.0030±0.0026
0.5	0.0045±0.0047	0.0048±0.0043	0.0054±0.0050
1.0	0.0028±0.0033	0.0053±0.0059	0.0048±0.0048
2.0	0.0024±0.0023	0.0026±0.0027	0.0060±0.0053

TABLE IV
EFFECTS OF THE VARIANCE OF THE OUTLIERS

Outliers variance	IEP by MCC	IEP by BW	IEP by LAR
0.2	0.0031±0.0032	0.0026±0.0024	0.0072±0.0062
0.5	0.0021±0.0018	0.0022±0.0019	0.0059±0.0048
1.0	0.0025±0.0025	0.0025±0.0024	0.0055±0.0051
2.0	0.0034±0.0038	0.0032±0.0034	0.0061±0.0056
4.0	0.0022±0.0022	0.0023±0.0024	0.0069±0.0068

In the fourth set of simulations, we investigate the effect of the mean and variance of the outliers. First, we keep the variance at 0.1 and set the mean of the outliers to be 0.2, 0.5, 1, and 2, respectively. The performance of MCC, BW and LAR are summarized in Table III. In short, large-mean outliers are easy to reject while small-value outliers naturally have small effect on the estimation. Next, we set the mean of the outliers to be 4 and vary the variance. The results are shown in Table IV. As we see, MCC performs very well under a variety of circumstances. The kernel size is set at 0.5 throughout this set of simulations.

A second, more complex and nonlinear, regression experiment is conducted to demonstrate the efficiency of MCC. Let the noise pdf be $p_Z(z) = (1-\varepsilon)\times\mathbf{N}(0, 0.1)+\varepsilon\times\mathbf{N}(4, 0.1)$ and $f(X) = \mathrm{sinc}(X), X \in [-2, 2]$. A multilayer perceptron (MLP) is used as the function approximator $g(x; w)$ with 1 input unit, 7 hidden units with tanh nonlinearity, and 1 linear output. The data length is $M = 200$. Under the MSE criterion, the MLP is trained for 500 epochs with learning rate 0.01 and momentum rate 0.5. Under the LAR criterion, 600 epochs are used with learning rate 0.002 and momentum rate 0.5. In the MCC case, MSE criterion is used for the first 200 epochs and switched to MCC for the next 400 epochs with learning rate 0.05 and momentum rate 0.5. Different values of ε are tried to test the efficiency of MCC against LAR. 50 Monte Carlo simulations are run for each value. The results are in Table V. The kernel size in MCC is chosen as $\sigma = 1$ for best results. An outstanding result is that MCC can attain the same efficiency as MSE when

TABLE V
NONLINEAR REGRESSION RESULTS SUMMARY

ε	IEP by MCC	IEP by MSE	IEP by LAR
0.1	0.0059±0.0026	0.2283±0.0832	0.0115±0.0068
0.05	0.0046±0.0021	0.0641±0.0325	0.0083±0.0041
0.01	0.0039±0.0017	0.0128±0.0124	0.0058±0.0025
0	0.0040±0.0019	0.0042±0.0020	0.0061±0.0028

TABLE VI
EFFECTS OF KERNEL SIZE ON MCC IN NONLINEAR REGRESSION

σ	IEP
0.2	0.01476±0.01294
0.5	0.00594±0.00243
1.5	0.00556±0.00252
2.0	0.00854±0.00737
4.0	0.01553±0.00840
10	0.06020±0.02489
Silverman's	0.01220±0.00612

the noise is purely Gaussian due to its unique property of "mix norm" whereas LAR cannot.

Next we fix $\varepsilon = 0.05$ and choose different kernel sizes to show that a wide range of values can be used in this problem. The results are in Table VI. A noteworthy observation and corollary of property 8 is that when large kernel sizes are employed, MCC reduces to MSE. In other words, MCC using large kernel sizes will never perform worse than MSE. The kernel size estimated by Silverman's rule on the error signal is around 0.2 during adaptation. For this problem this heuristic is not the best possible value, but it is still far better that the MSE solution.

B. TPCA With Correntropy

In this example, we present a correntropy extension to the Karhunen-Loeve transform that will be called the TPCA, which is widely utilized in subspace projections [15]. Suppose the signal is $\{x(i), i = 1, 2, \ldots, N + L - 1\}$ and we can map this signal as a trajectory of N points in the reconstruction space of dimension L. With the data matrix

$$\mathbf{X} = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \\ \vdots & \vdots & \ddots & \vdots \\ x(L) & x(L+1) & \cdots & x(N+L-1) \end{bmatrix}_{L\times N}. \tag{57}$$

Principal component analysis (PCA) estimates the eigenfilters and principal components (PC) [15]. We review this technique in a different way here to extend the method easily to correntropy TPCA. The autocorrelation matrix and Gram matrix are denoted as $\mathbf{R}$ and $\mathbf{K}$, respectively, and written as

$$\mathbf{R} = \mathbf{X}\mathbf{X}^T \approx N \times \begin{bmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & r(0) & r(1) \\ r(L-1) & \cdots & r(1) & r(0) \end{bmatrix}_{L\times L} \tag{58}$$

$$\mathbf{K} = \mathbf{X}^T\mathbf{X} \approx L \times \begin{bmatrix} r(0) & r(1) & \cdots & r(N-1) \\ r(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & r(0) & r(1) \\ r(N-1) & \cdots & r(1) & r(0) \end{bmatrix}_{N\times N} \tag{59}$$

where $r(k) = \mathbf{E}[x(i)x(i+k)]$ is the autocorrelation function of x. When N and L are large, (58) and (59) are good approximations. In the following derivation, we will see that L is actually not involved in the new algorithm, so we can always assume L is set appropriately to the application. Assuming $L < N$, by singular value decomposition (SVD) we have

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{60}$$

where $\mathbf{U}, \mathbf{V}$ are two orthonormal matrices and $\mathbf{D}$ is a pseudodiagonal $L\times N$ matrix with singular values $\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots \sqrt{\lambda_L}\}$ as its entries. Therefore

$$\mathbf{R} = \mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{D}^T\mathbf{U}^T \tag{61}$$

$$\mathbf{K} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T. \tag{62}$$

From (61) and (62), the columns of $\mathbf{U}$ and $\mathbf{V}$ are eigenvectors of $\mathbf{R}$ and $\mathbf{K}$, respectively. Rewriting (60) as

$$\mathbf{U}^T\mathbf{X} = \mathbf{D}\mathbf{V}^T \tag{63}$$

or equivalently

$$U_i^T\mathbf{X} = \sqrt{\lambda_i}V_i^T, \quad i = 1, 2, \ldots, L. \tag{64}$$

Here U_i and V_i are the ith columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. This equation simply shows that the projected data onto the ith eigenvector of $\mathbf{R}$ is exactly the scaled ith eigenvector of $\mathbf{K}$. This derivation provides another viewpoint to understand why Kernel PCA obtains the principal components from the Gram matrix [16]. As we see, for the conventional PCA, we can either obtain the principal components by eigendecomposing the autocorrelation matrix and then projecting the data or by eigendecomposing the Gram matrix directly.

Moreover, by property 9, there exists a scalar nonlinear mapping $\eta(\cdot)$ (not Φ) which maps the signal as $\{\eta_x(i), i = 1, 2, \ldots, N + L - 1\}$ while preserving the similarity measure

$$\mathbf{E}[\eta_x(i) \cdot \eta_x(j)] = \mathbf{E}[\kappa(x(i) - x(j))]. \tag{65}$$

In other words, the autocorrelation function of $\eta_x(i)$ is given by the correntropy function of x. With these results, the correntropy extension to temporal PCA is straightforward. We simply replace the autocorrelation entries with the correntropy entries in (59) and obtain the principal components by eigendecomposition of the new Gram matrix $\mathbf{K}$. Therefore, correntropy TPCA is similar to kernel PCA in the sense that none has access to the projected data, so the only way to obtain the principal components is by eigendecomposing the Gram matrix. However, we are talking about entirely different feature spaces as explained in property 10.

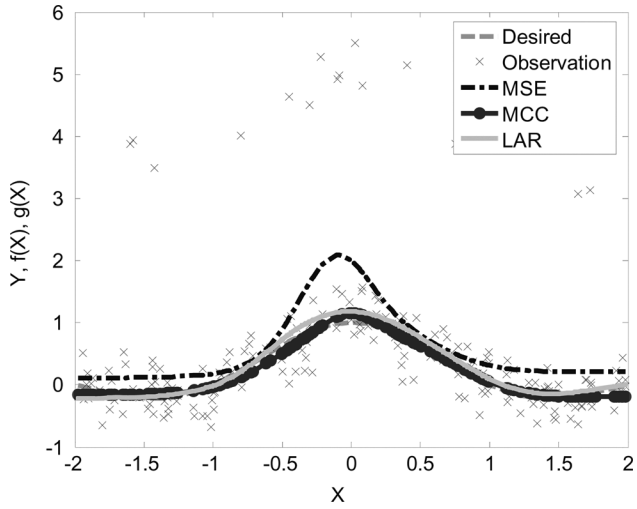


Fig. 4. Nonlinear regression results by MSE, MCC, and LAR, respectively.

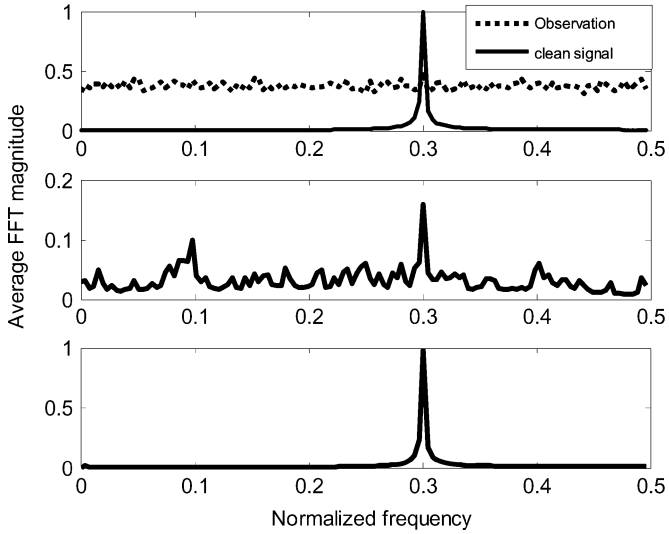


Fig. 5. PCA of sinusoidal signal in impulsive noise. (top) FFT of clean sinusoidal signal and noisy observation ($A = 5$). (middle) Average of FFT of first principal component by EVD N-by-N autocorrelation matrix. (bottom) Average of FFT of second principal component by EVD N-by-N correntropy matrix.

As a practical example, we apply correntropy TPCA (CTPCA) to a sinusoidal signal corrupted with impulsive noise

$$x(m) = \sin(2\pi f m) + A \cdot z(m) \tag{66}$$

for $m = 1, 2, \ldots, N + L - 1$. $z(m)$ is a white noise process drawn from the pdf

$$p_Z(z) = 0.8 \times \mathbf{N}(0, 0.1) + 0.1 \times \mathbf{N}(4, 0.1) + 0.1 \times \mathbf{N}(-4, 0.1). \tag{67}$$

We set $N = 256$, $f = 0.3$ and generate 3 N data to estimate N point autocorrelation and correntropy functions. For TPCA, the larger the subspace (larger L), the higher the signal-noise-ratio (SNR) is. For a fair comparison, we choose to eigendecompose the N-by-N Gram matrix for both methods. Results by eigendecomposing the L-by-L autocorrelation matrix and then projecting the data are also presented for comparison. For each case in Table VII, 1000 Monte Carlo trials with different noise realizations are run to evaluate the improvement of CTPCA upon TPCA. For $A = 5$, the probability of detecting the sinusoidal signal successfully as the largest peak in the spectrum is 100% for CTPCA, compared with 15% for N-by-N autocorrelation (Fig. 5). The kernel size is set to $\sigma = 1$ in CTPCA throughout this simulation. In this particular application of finding sinusoids in noise, the kernel size can be scanned until the best line spectrum is obtained.

In CTPCA, the second principal component is used instead of the first one, because the transformed data is not centered in the feature space and the mean introduces a large dc component that is picked as the first component. This can be easily shown as follows. We denote $\mathbf{C}$ as the correntropy matrix, $\mathbf{C}_c$ as the centered correntropy matrix, m_η as the mean of the transformed data, 1_N as an N -by-1 column vector with all entries equal to 1 and $1_{N\times N}$ as an N-by-N matrix with all entries equal to 1. Thus

$$\mathbf{C}_c = \mathbf{C} - m_\eta^2 \cdot 1_N \cdot 1_N^T. \tag{68}$$

If we normalize the eigenvector 1_N to be unit norm, the corresponding eigenvalue is Nm_η^2 which can be very large if N is large as in this example. Therefore, the first principal component of the correntropy matrix in this example is always a dc component.

The method widely used in kernel methods to calculate the centered Gram matrix [19] also applies here

$$\mathbf{C}_c = \mathbf{C} - 1_{N\times N} \cdot \mathbf{C}/N - \mathbf{C} \cdot 1_{N\times N}/N + 1_{N\times N} \cdot \mathbf{C} \cdot 1_{N\times N}/N^2. \tag{69}$$

Simulation results in Table VII shows the centering CTPCA works very well.

Another way to center the correntropy matrix is to estimate the square of the mean in (68) directly by the Information Potential as shown in Property 9. In this experiment, the approximation involved in estimating the square of the mean by IP will introduce about 2% error and after this error is amplified by N, the total error could exceed the eigenvalues corresponding to the signal space. Roughly there is about 10% degeneration in performance by using the IP method.

The second set of simulations is run to show how the kernel size affects the performance of CTPCA and to throw some light on how to choose it appropriately. Let $A = 5$. One thousand Monte Carlo simulations are run for each kernel size and the results are listed in Table VIII.

A graphical description of this table is achieved by plotting the correntropy spectrum density (i.e., the Fourier transform of the correntropy function) as a function of the kernel size (Fig. 6). The y axis shows the average normalized amplitude across the 1000 runs, which can be interpreted as the percentage of time that the particular frequency was the highest. We see that for kernel sizes between 1 and 2, the frequency of the sinusoid was the highest. Kernel size 10 is very similar to the power spectrum of the data, and it shows that in only 10% of the runs the highest peak in the spectrum corresponded to the sinusoid. In a sense, Fig. 6 exemplifies a new type of frequency analysis

TABLE VII
RESULTS OF TIME SERIES PCA

A	CTPCA (2^{nd} PC)	Centering CTPCA (1^{st} PC)	PCA by N-by-N Gram R (N=256)	PCA by L-by-L autocorrelation matrix (L=4)	PCA by L-by-L autocorrelation matrix (L=30)	PCA by L-by-L autocorrelation matrix (L=100)
5	100%	100%	15%	3%	4%	8%
4	100%	100%	27%	6%	9%	17%
2	100%	100%	99%	47%	73%	90%

TABLE VIII
THE EFFECTS OF THE KERNEL SIZE ON TIME SERIES PCA

Kernel size	Centering CTPCA
0.1	48%
0.5	93%
1.0	100%
1.5	99%
2.0	98%
3.0	95%
3.5	90%
4.0	83%
8.0	10%

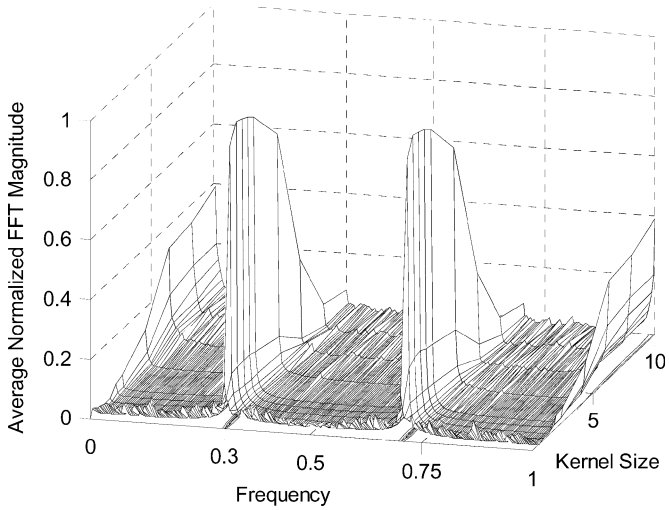


Fig. 6. Average of normalized FFT magnitude of first principal component of centering correntropy matrix over the change of kernel size.

based on correntropy, where a single parameter (the kernel size) in the correntropy function is able to scan between the conventional linear frequency analysis (large kernel sizes) and a nonlinear (higher-order moments) frequency analysis (smaller kernel sizes).

An intuitive explanation of the behavior of CTPCA is also provided by Property 8. The signal in this example has a relatively small dynamic range and mainly cluster in the Euclidean zone of the CIM space (we can always adjust the kernel size to implement this condition as long as we have sufficient data). The impulsive noise produces outliers in the rectification zone. Correntropy TPCA employs two steps: forward nonlinear transformation and backward linear approximation. By forward nonlinear transformation, we can move the "bad" points along the contours in CIM space to approach the bulk of data in the L2 norm sense; then by implicit backward linear approximation, we simply treat correntropy entries as second-order statistics. This perspective justifies our correntropy-extension to second-order methods and explains why this method works well in impulsive noise environments.

This insight gives also a clue on how to choose an appropriate kernel size for this application, that is, choosing a kernel size such that the normal signal dynamics lie in the Euclidean zone while the impulsive outliers are kept in the Rectification zone. Indeed, the normal signal dynamics in this example is about 2 and the impulsive outliers are above 8, so an appropriate kernel size is in the range of $[0.5, 3.5]$ according to the 3σ condition for outlier rejection and the variance condition of (25).

We also tried Kernel PCA [16] on this problem and no reasonable results were achieved. When the Gaussian kernel with large kernel size is employed, Kernel PCA is almost equivalent to linear PCA but not better. When a polynomial kernel is used, the higher its degree the larger the amplification of the effect of outliers, thus worsens the situation.

VI. CONCLUSION

This paper explains the probabilistic and geometric meanings of correntropy, and shines light on its connections with M-estimation, ITL, and kernel methods, hoping the insights gained here will be helpful in other research contexts.

As a measure of similarity, correntropy directly indicates the probability density of how close two random variables are in a specific "window" controlled by the kernel size. As a second-order statistics in the feature space, correntropy possesses a lot of properties quantifying the data pdf directly. Based on this understanding, the advantage of using correntropy in non-Gaussian signal processing is also showed theoretically and experimentally. As a new cost function, MCC adaptation is applicable in any noise environments when its distribution has the maximum at the origin. It outperforms MSE in the case of impulsive noise since correntropy is inherently insensitive to outliers. MCC is theoretically equivalent to M-estimation and has a weighting function similar to Bi-square. It is infinitely differentiable compared with L1 norm and can attain the same efficiency as MSE in the case of Gaussian noise due to its unique 'mix norm' property. The computational complexity of MCC is $\mathrm{O}(N)$, much simpler compared with other methods of higher-order statistics like cumulant methods and even MEE.

As a new metric in the sample space, CIM exhibits a property of 'mix norm'. This space is locally linear but highly nonlinear globally. Accordingly, we divide the space and label them Euclidean zone, Transition zone, and Rectification zone with respect to the data of interest. Correntropy extends second-order methods by 'filtering' outliers in an explicit forward nonlinear transformation and implicit backward linear approximation.

Two case studies are also presented. The first investigates the performance of MCC on the adaptation of two function approximators (polynomial and MLP) with non-zero mean measurement noise. Optimal solutions obtained by MSE, MEE, MCC and existing robust fitting methods were compared. The second case addresses temporal principal component analysis of a sinusoidal signal corrupted by impulsive noise. The improvement of correntropy-based PCA was shown compared with the conventional autocorrelation based method. These results indicate the obvious advantage of correntropy versus second-order statistics and therefore this study offers a feasible alternative to second-order statistics in the detection of sinewaves in non-Gaussian high noise environments. The perspectives gained here apply to many other estimation algorithms employing nonlinearities. For example, in the FastICA algorithm [13], a Gaussian nonlinearity is used to estimate the kurtosis instead of using the fourth power function directly.

Like in the Parzen estimation, any symmetric kernel can be used and most of the results presented in the paper still hold. The reasons we prefer the Gaussian kernel are: first, Gaussian kernel is smooth everywhere; second, and most importantly, the integral of the product of two Gaussians is still a Gaussian as shown in the calculation in (6), (15) and (40), i.e., the Gaussian kernel gives much simpler and more elegant expressions here.

Correntropy is ultimately dependent upon the kernel size, which should be selected according to the application. Although the kernel size is introduced in correntropy through the Parzen estimation step, its net effect differs greatly due to the expected value operator in the definition of correntropy. It is easy to show that correntropy defaults to correlation for larger than recommended kernel sizes (for density estimation). The use of higher-order moment information in correntropy is controlled smoothly by the kernel size, which is very appealing and unique. We have shown that, when selected appropriately for the application, correntropy provides results comparable to robust statistical methods, and the performance sensitivity to kernel size should be much smaller than the selection of thresholds due to the smooth dependence of correntropy on the kernel size. Nevertheless, methods to estimate the kernel size in practical applications are necessary, and are presently being investigated. Further work includes the extension of correntropy to arbitrary dimensional random variables, and further analysis of property 7 as an ICA criterion. One of the most interesting and perhaps most important lines of research is to understand the mathematical structure and properties of the reproducing kernel Hilbert space induced by the correntropy kernel, based on which we believe that practical extensions to statistical estimation theory for non-Gaussian processes are possible.

REFERENCES

[1] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, pp. 1780–1786, Jul. 2002.

[2] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–, 1962.

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

[4] D. Erdogmus and J. C. Principe, "Generalized information potential criterion for adaptive system training," *Trans. Neural Netw.*, vol. 13, no. 5, pp. 1035–1044, Sept. 2002.

[5] D. Erdogmus, J. C. Principe, and K. E. Hild II, "Beyond second-order statistics for learning: A pair-wise interaction model for entropy estimation," *Natural Comput.*, vol. 1, no. 1, pp. 85–108, May 2002.

[6] K. E. Hild II, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1385–1392, Sep. 2006.

[7] D. Erdogmus, R. Agrawal, and J. C. Principe, "A mutual information extension to the matched filter," *Signal Process.*, vol. 85, no. 5, pp. 927–935, May 2005.

[8] K.-H. Jeong, P. P. Pokharel, J.-W. Xu, S. Han, and J. C. Principe, "Kernel based synthetic discriminant function for object recognition," in *Int. Conf. Acoust., Speech, Signal Process.*, France, May 2006, pp. 765–768.

[9] I. Santamaria, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: Definition, properties and application to blind equalization," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2187–2197, Jun. 2006.

[10] E. Parzen, "Statistical methods on time series by Hilbert space methods," Appl. Math. Statistics Lab., Stanford Univ., Tech. Rep. 23, 1959.

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

[12] P. P. Pokharel, J. Xu, D. Erdogmus, and J. C. Principe, "A closed form solution for a nonlinear wiener Filter," presented at the Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2006.

[13] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, 1999.

[14] R. Jenssen, D. Erdogmus, J. C. Principe, and T. Eltoft, "Towards a unification of information theoretic learning and kernel methods," in *Proc. 2004 IEEE Int. Work. Mach. Learn. Signal Process.*, Sao Luis, Brazil, Sep. 2004, pp. 443–451.

[15] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York: Wiley, 2000, pp. 371–375.

[16] B. Scholkopf, A. J. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[17] W. DuMouchel and F. O'Brien, K. Berk and L. Malone, Eds., "Integrating a robust option into a multiple regression computing environment," in *Computing Science and Statistics: Proc. 21st Symp. Interface*, Alexandria, VA, 1989, pp. 297–301, Amer. Statist. Assoc..

[18] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: A localized similarity measure," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2006, pp. 4919–4924.

[19] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[20] P. Huber, *Robust Statistics*. New York: Wiley, 1981.

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

[22] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *J. Mach. Learning Res. 6*, pp. 2075–2129, 2005.

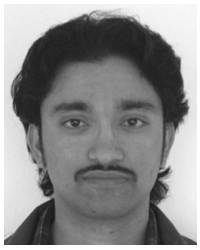
[23] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman and Hall, 1986.

[24] D. Erdogmus and J. Principe, "From linear adaptive filtering to nonlinear information processing," *IEEE Signal Process. Mag.*, Nov. 2006.

[25] K. H. Jeong and J. C. Principe, "The correntropy MACE filter for image recognition," in *Proc. IEEE Int. Work. Mach. Learning for Signal Process.*, Ireland, 2006, pp. 9–14.

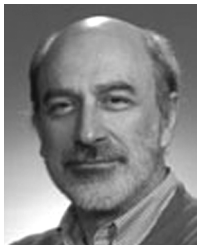
Weifeng Liu (S'06) was raised in Shanghai, China. He received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University in 2003 and 2005, respectively.

In 2005, he joined the Computational NeuroEngineering Laboratory (CNEL), University of Florida, Gainesville, as a Ph.D. degree student. His research interests are signal processing and data inference using machine learning and pattern recognition techniques.



Puskal P. Pokharel (S'04–M'07) was born in Lalitpur, Nepal, in 1981. He received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology (IIT), Roorkee, in 2003, and the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2005.

Presently, he is pursuing the Ph.D. degree at the University of Florida Compuational NeuroEngineering Laboratory (CNEL). 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) received the B.S. degree from the University of Porto, Portugal, in 1972 and the M.Sc. and Ph.D. degrees from the University of Florida in 1974 and 1979, respectively.

He is a Distinguished Professor of Electrical and Biomedical Engineering with the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is a BellSouth Professor and Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). He is involved in biomedical signal processing, in particular, the electroencephalogram (EEG) and the modeling and applications of adaptive systems. He has more than 129 publications in refereed journals, 15 book chapters, and over 300 conference papers. He has directed more than 50 Ph.D. dissertations and 61 master's degree theses.

Dr. Principe is Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, President of the International Neural Network Society, and formal Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is an AIMBE Fellow and a recipient of the IEEE Engineering in Medicine and Biology Society Career Service Award. He is also a member of the Scientific Board of the Food and Drug Administration, and a member of the Advisory Board of the McKnight Brain Institute at the University of Florida.