# Correntropy: A Localized Similarity Measure

Weifeng Liu, P. P. Pokharel, and J. C. Principe, *Fellow, IEEE*

***Abstract*—The measure of similarity normally utilized in statistical signal processing is based on second order moments. In this paper, we reveal the probabilistic meaning of correntropy as a new localized similarity measure based on information theoretic learning (ITL) and kernel methods. As such it has vastly different properties when compared with Mean Square Error (MSE) that can be very useful in nonlinear, non-Gaussian signal processing. Two examples are presented to illustrate the technique.**

## I. Introduction

The Computational NeuroEngineering Laboratory at the University of Florida has extended the concept of mean square error adaptation to include descriptors of entropy and divergence so useful in Information Theory [1]. Information Theoretic Learning (ITL) preserves the nonparametric nature of MSE, i.e. the cost function is still directly estimated from data with a Parzen kernel estimator [2], but extracts more information from the data structure for the adaptation process, and yields therefore solutions that are more accurate than MSE for non-Gaussian processes [3]-[5].

The fundamental definition of autocorrelation for random processes was also generalized to auto-correntropy function [6], which measures similarity across lags as the autocorrelation, but when averaged across lags, it yields the entropy of the random variable, hence its name. Therefore correntropy contains higher order moments of the PDF but it is much simpler to estimate directly from samples and bypasses the need for conventional moment expansions. However, this definition only applies to a single random variable (at different lags) and so it can not be generally applied outside the realm of scalar random processes. This paper extends auto-correntropy to cross-correntropy, the function needed to handle the general case of two arbitrary random variables, and provides an intuitive viewpoint to help us apply correntropy judiciously to kernel methods and nonlinear, non-Gaussian signal processing.

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.

The organization of the paper is as follows. After a brief review of information theoretical learning and kernel methods, the definition and its probabilistic meaning of correntropy are presented in section III. Then in section IV, two examples are presented to corroborate our understanding and to inspire readers with possibly numerous applications in their research fields. Finally, section V summarizes the main conclusions.

This work was supported in part by NSF grant ECS-0300340 and ECS-0601271.

The authors are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA, phone: 352-392-2682, fax: 352-392-0044 (e-mail: weifeng@cnel.ufl.edu, pokharel@cnel.ufl.edu, principe@cnel.ufl.edu ).

## II. Information Theoretical Learning and Kernel Methods

Information Theoretic Learning is a framework to non-parametrically adapt systems based on entropy and divergence [1]. Renyi's $\alpha$ -order entropy of a random variable *X* is defined by

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\int f_X^\alpha(x)dx. \tag{1}$$

Estimating the PDF with Parzen estimators for the samples $\{x_i, i=1,2,...N\}$ drawn from the PDF, we obtain the estimator

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} k_\sigma(x-x_i) \tag{2}$$

where $k_\sigma(x)$ is the Gaussian kernel

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

*N* is the number of the samples and $\sigma$ the kernel size. This subscript is usually suppressed unless required for clarity. For $\alpha$ = 2 (quadratic entropy), we obtain a nonparametric estimator of quadratic entropy as

$$\hat{H}_2(X) = -\log IP(x) \tag{4}$$

$$IP(X) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} k_\sigma(x_j - x_i). \tag{5}$$

*IP(X)* is called information potential (IP). The PDF

estimated with Parzen kernels can be thought as defining an information potential field over the space of the samples [5]. It is therefore interesting to use the information potential to define similarity measures in this space, which do not possess the limitation of the conventional moments. Towards this goal we recently proposed a new similarity measure for random processes called auto-correntropy [6].

Let $\{X_t, t \in T\}$ be a stochastic process with $T$ being an index set. The nonlinear transformation induced by the kernel mapping $\Phi$ maps the data into the feature space [8], where the auto-correntropy function $V_X(t_1, t_2)$ is defined from $T \times T$ into $R^+$ given by

$$\begin{aligned} V_X(t_1, t_2) &= E[<\Phi(X_{t_1}), \Phi(X_{t_2})>] \\ &= E[k(X_{t_1} - X_{t_2})] \end{aligned} \tag{6}$$

where $k(\cdot)$ is the Gaussian kernel, and without loss of generality, it will be the only one considered in this paper.

The auto-correntropy estimates the second order moments of the data transformed by the eigen-function of the Gaussian kernel. It shares with the autocorrelation function the fact that it quantifies similarities among pairs of lags, so it is capable of quantifying the time structure of the random process. However, it supersedes the conventional autocorrelation function because this similarity is not limited to second order moments. In fact for the Gaussian kernel all even moments of the random variable contribute to the estimation of similarity due to the nonlinearity provided by the kernel [6]. We have further shown that auto-correntropy is a symmetric, positive-definite function and therefore defines a new reproducing kernel Hilbert space (RKHS) [9]. Based on auto-correntropy it is possible to derive an analytical solution of the optimal linear combiner in this space [7].

## III. The Probabilistic Meaning of Correntropy

In this paper a more general form of correntropy is defined between two arbitrary random variables $X$ and $Y$ given by

$$V(X, Y) = E[<\Phi(X), \Phi(Y)>] = E[k(X - Y)]. \tag{7}$$

when $X = X_{t_1}$ and $Y = X_{t_2}$, this definition reduces to (6).

This new function is called cross-correntropy or simply correntropy. As can be seen correntropy is a straight extension of auto-correntropy, but now two random variables are involved, so it is important to understand under what conditions this is a reasonable measure of similarity between *X* and *Y*. The nonlinearity introduced by the kernels has important implications in assessing the higher order moments of the joint PDF, but the Gaussian kernels also restricts the analysis to a local region of the joint space. Therefore, we need to conduct a systematic analysis.

In practice, the joint PDF is unknown and only finite number of data samples available $\{(x_i, y_i), i = 1, 2, ..., N\}$ to estimate it

$$V(X, Y) = \frac{1}{N} \sum_{i=1}^{N} k(x_i - y_i). \tag{8}$$

We show now that correntropy is actually the integral of the joint PDF of the data along the line $x = y$

$$V(X, Y) \approx \int_{-\infty}^{+\infty} f_{X,Y}(x, y)|_{x=y=u} \, du. \tag{9}$$

Strict equality holds when the kernel size $\sigma$ tends to zero.

From the definition (7),

$$\begin{aligned} V(X, Y) &= E[k(X - Y)] \\ &= \iint_{x,y} k(x - y) f_{XY}(x, y) dx dy. \end{aligned} \tag{10}$$

The Gaussian kernel function $k(x - y)$ is exactly a ridge function along the $x = y$ line. Indeed the Gaussian kernel has high values only when $x_i \approx y_i$, with an exponential fall off when $y$ is dissimilar from $x$. When the kernel size $\sigma$ approaches 0, it becomes a delta function $\Delta(x - y)$ and (10) turns out to be

$$\begin{aligned} \lim_{\sigma \to 0} V(X, Y) &= \iint_{x,y} \Delta(x - y) f_{X,Y}(x, y) dx dy \\ &= \int_{x=-\infty}^{+\infty} f_{X,Y}(x, x) dx. \end{aligned} \tag{11}$$

Generally, we use data samples to estimate correntropy instead of using the expected value.

Assume the data samples $\{(x_i, y_i) \quad i = 1, 2, ..., N\}$ are available to estimate the correntropy by (8).

On the other hand, with these data, the Parzen method can be used to estimate the Joint PDF $f_{X,Y}(x, y)$ as

$$f_{X,Y}(x, y) \approx \frac{1}{N} \sum_{i=1}^{N} k(x - x_i) \cdot k(y - y_i). \tag{12}$$

When the kernel size tends to zero and the product $N\sigma$ to infinity according to the conditions of Parzen method, strict equality holds for (12). Integrating (12) along line $x = y$

$$\int_{-\infty}^{+\infty} f_{X,Y}(x,y)|_{x=y=u}\, du$$
$$\approx \int_{-\infty}^{+\infty} \frac{1}{N}\sum_{i=1}^{N} k(x-x_i)\cdot k(y-y_i)|_{x=y=u}\, du$$
$$= \int_{-\infty}^{+\infty} \frac{1}{N}\sum_{i=1}^{N} k(u-x_i)\cdot k(u-y_i) du \qquad (13)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty} k(u-x_i)\cdot k(u-y_i) du$$
$$= \frac{1}{N}\sum_{i=1}^{N} k_{\sqrt{2}\sigma}(x_i - y_i).$$

The final term is exactly the estimation of correntropy with the kernel size $\sqrt{2}\sigma$. Fig.1 shows diagrammatically that correntropy provides the probability density of the event *p(X=Y)*. This understanding is interesting, because we started with an extension of cross correlation, but indeed we are able to quantify the probability of two events being equal. So correntropy is unable to assess similarity well in the entire join space, but for *X=Y* gives us an estimate of probability density!

In practical applications, the joint PDF is unknown and we only have finite number of data to estimate correntropy. Finite number of data also constrains the kernel size from being too small, since small kernel size may lead to meaningless estimation. Assume the kernel size used in correntropy is $\sigma$ which is relatively small compared with the variance of the data distribution. Therefore, a rectangle approximation with bandwidth $\sqrt{\pi/2}\sigma$ can be used in (10) in the place of Gaussian kernel and we will have a more precise approximation of correntropy as

$$V(X,Y) = E[k_\sigma(X-Y)]$$
$$\approx \frac{P(|Y-X|<\sqrt{\pi/2}\sigma)}{\sqrt{2\pi}\sigma}. \qquad (14)$$

Since correntropy evaluates directly from data the probability $P(|Y-X|<1.25\sigma)$ for a given kernel size $\sigma$, it can be used as a localized similarity measure for supervised applications where the mean square error (MSE) criterion as been traditionally utilized. Indeed when we compare the output of an adaptive system with the desired response for a training set, we are ultimately asking: what is the probability that the two measurements are equal? Therefore, we propose correntropy as a new cost function for adaptive system training, with the advantage that it is a local criterion of similarity and it should be very useful for cases when the measurement noise is non-zero mean, non-Gaussian, with large outliers.

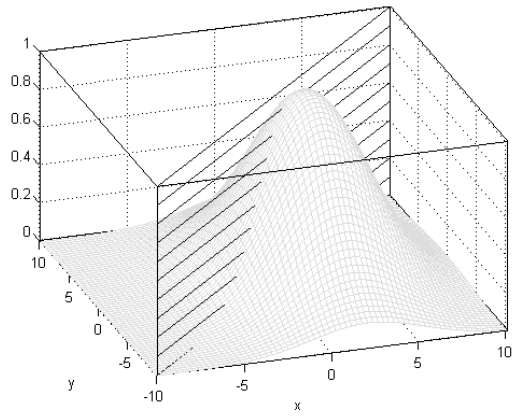



Fig. 1. Correntropy as the integral in the joint space along *x=y* line.

## IV. Applications

### A. Optimal Receiver

In a digital communication system, let *X* and *Y* be respectively the transmitted signal and the received signal corrupted by the additive channel noise *N*.

$$Y = X + N \qquad (15)$$

Suppose $X = s$ which is either -1 or +1 with equal probably. The noise PDF is $f_N(n)$. Therefore, the PDF of $Y$ is

$$f_Y(y) = f_N(y-s). \qquad (16)$$

With the observation available, we want to recover $s$. Under the MSE criterion, we choose $\hat{s}$ such that it minimizes the following cost function.

$$MSE(Y,\hat{s}) = \int_{-\infty}^{\infty} (y-\hat{s})^2 f_Y(y) dy$$
$$= \int_{-\infty}^{\infty} (n+s-\hat{s})^2 f_N(n) dn. \qquad (17)$$

This criterion searches for a location in the noise PDF such that the variance is minimized. In fact, under the MSE criterion, the best estimation of $s$ is

$$\hat{s} = s + \frac{1}{M}\sum_{i=1}^{M} n_i = \frac{1}{M}\sum_{i=1}^{M} y_i \qquad (18)$$

$\{y_i, i=1,2,...,M\}$ and $\{n_i, i=1,2,...,M\}$ are the sample version of the observed signal and noise. This estimation is obviously biased if the noise PDF has non-zero mean.

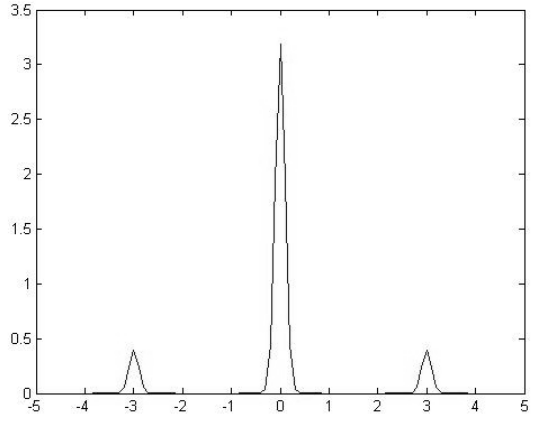


Fig. 2. Noise PDF for example 2. The noise distribution is symmetric but has outliers.

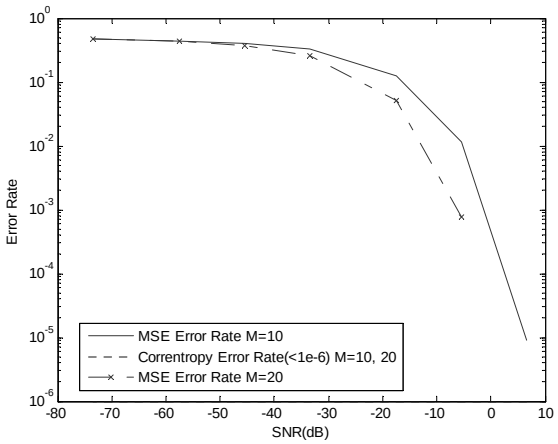



Fig. 3. Error rates by criteria of MSE and correntropy in the case of M = 10 and 20. For MSE, increasing the sample number improves the performance.

Under the correntropy criterion, we choose $\hat{s}$ such that it maximizes the following cost function

$$V(Y,\hat{s}) = \int_{-\infty}^{\infty} k(y-\hat{s})\, f_Y(y)dy = \int_{-\infty}^{\infty} k(n+s-\hat{s})\, f_N(n)dn. \quad (19)$$

And based on our understanding

$$V(Y,\hat{s}) \approx f_N(\hat{s}-s). \quad (20)$$

Therefore, this receiver can always find the best solution if and only if the noise PDF has its global maximum at the origin. Let the noise PDF be (Fig. 2)

$$f_N(n) = 0.8*N(0,0.1) + 0.1*N(3,0.1) + 0.1*N(-3,0.1). \quad (21)$$

We can show that the correntropy method is more robust than MSE, which was also observed in [6]. In the MSE case, the criterion becomes

$$\hat{s} = sign(\frac{1}{M}\sum_{i=1}^{M} y_i). \quad (22)$$

In the correntropy case, the criterion is

$$\hat{s} = sign(\frac{1}{M}\sum_{i=1}^{M} k(y_i,1) - \frac{1}{M}\sum_{i=1}^{M} k(y_i,-1)). \quad (23)$$

We run the simulation for $10^7$ times in the case of $M$ = 10 and 20. The kernel size is 0.1. We simply control the SNR by scaling the noise drawn from (21). In Fig. 3, we can see the ability of correntropy to reject irrelevant noise. In other words, correntropy has the ability of being insensitive to the noise peak in the PDF tail, and effectively handle the effect of the peak at the origin of the PDF, which may be crucial in many kinds of detection applications.

### *B. Function Approximation*

In the second example, we consider the general model of function approximation.

$$Y = f(X) + N \quad (24)$$

$f$ is the unknown function, $N$ is the noise process and $Y$ is the observation. A universal function approximator $g(x;w)$ is used to discover this function and alleviate the effect of noise as much as possible.

Let the noise probability density function be (Fig. 4)

$$f_N(n) = 0.9*N(0,0.1) + 0.1*N(4,0.1) \quad (25)$$

In MSE, the best solution is found by

$$\min J(w) = \frac{1}{N}\sum_{i=1}^{N}(g(x_i)-y_i)^2 \quad (26)$$

whereas with the correntropy criterion, the optimal solution is found by

$$\max J(w) = \frac{1}{N}\sum_{i=1}^{N} k(g(x_i)-y_i) \quad (27)$$

We use polynomial functions for $f(x)$ and $g(x)$ of the same order (n=3). The coefficient of $f(x)$ is arbitrarily chosen as $[0.17425\ 1.6096\ 1.3687\ 1.3559]$.

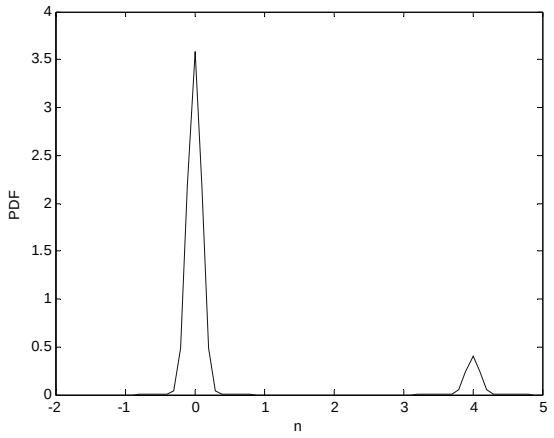



Fig. 4. Noise PDF for example 1. This noise is impulsive and has non-zero mean.

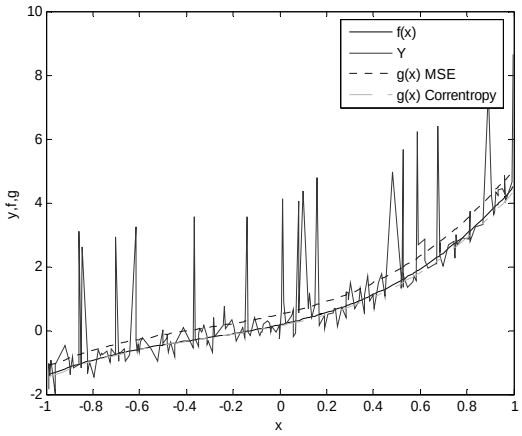
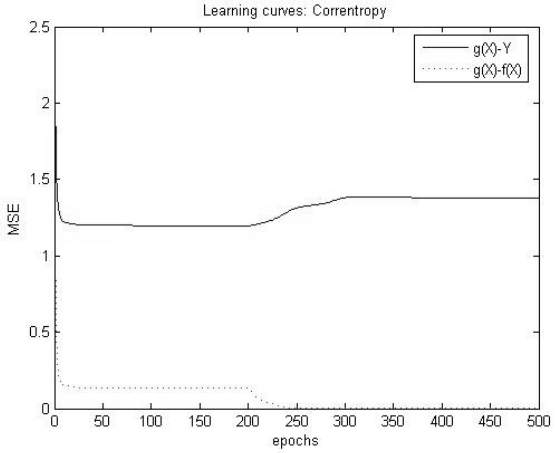



Fig. 6. Regression results with criteria of MSE and correntropy respectively. The observation Y is corrupted with positive impulsive noise; the regression result from MSE (dotted line) is shown shifted above the desired curve; the result from correntropy (dashed line) matches the desired quite well.

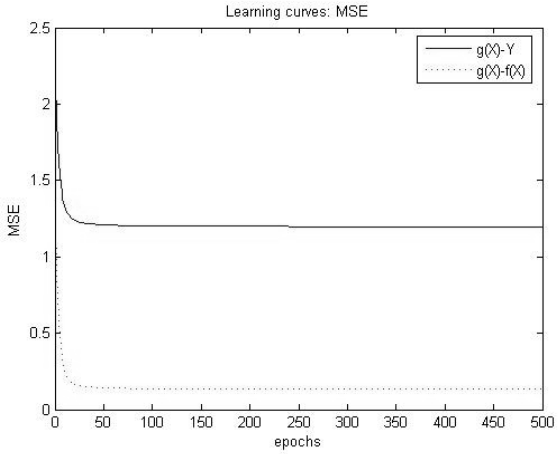
Under the MSE criterion, we set the learning rate 0.001 and train the system for 500 epochs (long enough to make sure it reaches its global solution). In the experiment of maximizing correntropy, we first train the system with MSE criterion during the first 200 epochs, pushing the coefficients close to the global solution (this is equivalent to kernel size annealing) and switch the criterion to maximize the correntropy during the next 300 epochs. The learning rate is set to 0.003 and the kernel size is chosen the same as the noise variance 0.1.

We run 50 Monte Carlo simulations with 50 different initial conditions. The estimated coefficient for MSE is [0.50465 1.6592 1.4503 1.4355] and for correntropy is [0.14304 1.5995 1.2877 1.4337] on average. The ensemble learning curves are shown in Fig.5.

The intrinsic error power between $f(x)$ and $g(x)$ for correntropy method is 0.0053 while for traditional MSE method is 0.1318 on average. When MSE criterion is used,




(a)




(b)

Fig. 5. Ensemble learning curves for criteria of (a) Correntropy and (b) MSE.

$g(x)$ is shifted somewhat by the non-zero-mean noise due to the global property of MSE (Fig.6). Now we understand the importance of correntropy and its local property. In other words, correntropy has the ability of being insensitive to the noise peak in the PDF tail, and effectively handle the effect of the peak at the origin of the PDF in regression. In this sense it implements an ε-norm penalty function in regression.

## V. Conclusion

In this paper, we extend the correlation function for two variables to a new function called correntropy. We proved mathematically that correntropy measures the probability density that two events are equal. This can be done directly from the data with the Gaussian kernel. We further explain the probabilistic meaning of correntropy and its localized property and the effect of kernel size on this localness.

Based on this understanding, the advantage of using correntropy as a cost function to train adaptive systems in non-Gaussian signal processing is also showed experimentally. The correntropy algorithm is applicable in

any noisy conditions provided the global maximum of the noise PDF is at the origin. It outperforms MSE in regression for the case of impulsive noise since correntropy is inherently insensitive to outliers. Further theoretical work is needed to fully understand the properties of correntropy, but the preliminary results in applications are very promising.

We believe that this is the first step to fully understand the correntropy kernel as proposed in [6].

## REFERENCES


[1] Jose C. Principe, Dongxin Xu, Qun Zhao, John Fisher, "Learning from examples with information theoretic criteria," *Journal of VLSI Signal Processing-Systems*, Vol. 26, No. 1-2, pp. 61-77, Aug. 2000.

[2] E. Parzen, "On the estimation of a probability density function and the mode", *Ann. Math. Stat.* 33, 1962, p1065.

[3] D. Erdogmus, J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. on Signal Processing*, vol. 50, pp. 1780-1786, July 2002.

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

[5] Deniz Erdogmus, Jose C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *Trans. on Neural Networks*, Vol. 13, No. 5, pp. 1035-1044, Sept. 2002.

[6] I. Santamaria, P. P. Pokharel, J. C. Principe, "Generalized correlation function: definition, properties and application to blind equalization," *IEEE Trans. Signal Processing*, to be published.

[7] P. P. Pokharel, J. Xu, D. Erdogmus, J. C. Principe, "A closed form solution for a nonlinear Wiener filter", *ICASSP2006*, to be published.

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

[9] E. Parzen, *Statistical Methods on Time Series by Hilbert Space Methods*, Technical Report no. 23, Applied Mathematics and Statistics Laboratory, Stanford University, 1959.