2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Nonlinear Component Analysis Based on Correntropy

Jian-Wu Xu, Puskal P. Pokharel, António R. C. Paiva and José C. Príncipe

***Abstract*— In this paper, we propose a new nonlinear principal component analysis based on a generalized correlation function which we call *correntropy*. The data is nonlinearly transformed to a feature space, and the principal directions are found by eigen-decomposition of the correntropy matrix, which has the same dimension as the standard covariance matrix for the original input data. The correntropy matrix characterizes the nonlinear correlations between the data. With the correntropy function, one can efficiently compute the principal components in the feature space by projecting the transformed data onto those principal directions. We give the derivation of the new method and present simulation results.**

I. Introduction

Principal component analysis (also known as the *Karhunen-Loève* transformation in communication theory) is a powerful tool for feature extraction and data dimensionality reduction in statistical pattern recognition and signal processing. It can be easily performed by eigen-decomposition of the standard covariance matrix or by adaptive algorithms that estimate principal components [2]. Principal component analysis or PCA is really an affine transformation of the coordinate system such that the rate of decrease of data variance is maximized. The projections of the data onto the new coordinate system are called *principal components*. These projections represent the data optimally in a least-square sense. In *feature extraction*, PCA transforms the data in such a way that a small number of principal components can represent the data while retaining most of the intrinsic variance of the data. These are sometimes called *factors* or *latent variables* of the data [5].

While PCA yields a smaller dimensional linear subspace that best represents the full data according to a minimum-square-error criterion, it might be a poor representation if the data structure is non-Gaussian. Hence nonlinear component analysis may be needed. There have been numerous attempts to define nonlinear components analysis in the latest decades. Nonlinear PCA is generally seen as a nonlinear generalization of standard PCA [2], [5]. The principal component is generalized from straight lines to curves. Principal curves were proposed by Hastie [4] to define local directions that pass through the high density parts of the data set. The principal curves are found through an iterative algorithm that minimizes the conditional expectation of projections on the curves. Kramer presented a nonlinear PCA based on auto-associative neural networks. The auto-associative network performs identity mapping from the input data to the output by minimizing the square error [6]. Recently, Schölkopf *et al* applied kernel methodology to obtain a nonlinear form of PCA [13]. This so called Kernel PCA solves the eigen-decomposition of the Gram matrix of the input data in a high-dimensional feature space. The Gram matrix has a dimension given by the number of samples N. The data projections onto the principal directions of the Gram matrix, i.e. the inner product in feature space, are carried out by means of kernel functions in the input space. While the utilization of Mercer kernels provides a tractable way to compute principal components in the high-dimensional feature space, there are still problems of interpretation and computation of the large dimensional Gram matrix. Indeed, the number of eigenfunctions of the Gram matrix is dependent on the number of data samples N, not the size of the data space L. Moreover computing Gram matrices for millions of samples in a small, let us say, two dimensional space becomes wasteful.

In this paper, we propose a new nonlinear PCA technique based on the generalized correlation function, which we call Correntropy PCA. The generalized correlation function quantifies the similarity between the L different components of the L dimensional input data vector (or the time structure in a time series) using the statistical data distribution [12]. The generalized correlation also utilizes a kernel methodology, but in a different form: by applying the kernel to pairs of data vector components, a random vector (or stochastic process) is nonlinearly transformed into a high dimensional function space where the similarity between the components of the transformed random variables (or stochastic process) can be measured by the conventional covariance function. The eigen-decomposition of the covariance of the transformed data yields the principal directions of the nonlinearly transformed data. These linear principal directions in feature space correspond to nonlinear principal directions in the input space. These projections can be efficiently computed by utilizing the generalized correlation function. That means, if one has one million samples in a two dimensional space, it is only necessary to solve a two dimensional eigenvector problem on a matrix whose entries are computed from one million samples. In many applications this is a tremendous computational saving.

The paper is organized as follows. First, we introduce the generalized correlation function in section 2. We present the definitions auto-correntropy and cross-correntropy functions and some properties of correntropy functions. The derivation of the Correntropy PCA is given in section 3. After that, we have a discussion section on some specific issues concerning Correntropy PCA. We present two experimental results comparing the Correntropy PCA with standard linear PCA and kernel PCA by Schölkopf [13] in section 5.

Jian-Wu Xu, Puskal P. Pokharel, António R. C. Paiva, José C. Príncipe are with the Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainsville, FL 32611, USA (phone: 352-392-2682; fax: 352-392-0044; email: {jianwu, pokharel, arpaiva, principe}@cnel.ufl.edu).

0-7803-9490-9/06/$20.00/©2006 IEEE

II. Correntropy Function

In this section, we give the definition and properties of the generalized correlation function. This generalized correlation function extends the correlation function to nonlinear spaces where data has been nonlinearly mapped according to some transformation associated with the data distribution. The correntropy function of two random variables x and y is defined as

$$V(x,y) = E[\kappa(x,y)], \tag{1}$$

where $E[\cdot]$ denotes mathematical expectation and κ is a positive definite kernel function that obeys the Mercer's conditions [7]. One widely used kernel function is the Gaussian kernel given by

$$\kappa(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-y)^2}{2\sigma^2}\}.$$

By using Parzen window to implicitly estimate the probability density function (PDF) [8], it turns out that the value of generalized correlation function coincides with the value of information potential of second order Renyi's entropy for a given data set [9], [12]. Thus the generalized correlation function is called *Correntropy* [12].

Correntropy has very nice properties that make it useful for nonlinear signal processing and machine learning. First and foremost, it is a positive definite function.

Property : Given any symmetric positive definite kernel function $\kappa(x,y)$, the correntropy function defined in (1) is also symmetric and positive definite.

Proof: Given a positive definite kernel function $\kappa(x,y)$, then for any sets of n random variables for x and y $\{x_1, x_2, \cdots, x_n\}$ and $\{y_1, y_2, \cdots, y_n\}$, any set of not all zero real numbers $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$, by definition we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\kappa(x_i,y_j) > 0.$$

Certainly, the expectation of any positive definite function is always positive definite. Thus we have

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\kappa(x_i,y_j)\right] > 0.$$

This is equivalent to

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j E[\kappa(x_i,y_j)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j V(x_i,y_j) > 0.$$

$V(x,y)$ is obviously symmetric since $\kappa(x,y)$ is symmetric. This concludes our proof.

The well known Moore-Aronszajn theorem states that there exists a unique reproducing kernel Hilbert space (RKHS) for any given positive definite function and vice versa [1]. So for the correntropy function $V(x,y)$ defined in (1), there is a unique RKHS associated with correntropy function. The nonlinear principal component analysis using correntropy function proposed in this paper is based on this property. As we know that the conventional correlation function for any two random variables is not necessary positive definite (the autocorrelation function for any given random process is positive definite), thus there is no such RKHS associate with convectional correlation function. By imposing a positive definite kernel function on the argument inside the expectation, the correntropy function becomes positive definite which makes it a reproducing kernel. The data then will be nonlinearly transformed into this RKHS, called feature space, and we only need to compute the inner product via the correntropy function to perform nonlinear algorithms. We provide two ways of nonlinear mapping to get insights into the feature space associated with the correntropy function. One is through the reproducing kernel mapping and the other is by the Mercer kernel mapping.

A. The reproducing kernel mapping: As we have already seen that correntropy function corresponds a unique reproducing kernel Hilbert space, this suggests that we can map data into the feature space by means of RKHS. Let $\mathcal{H}$ be the RKHS, then for any functional $f \in \mathcal{H}$, by Riesz represent theorem [11], we have

$$f(x) = < f, V(x,\cdot) >$$

The feature space can be constructed by containing all finite linear combinations of the form $\sum \alpha_i V(x_i,\cdot)$ and the inner product is given by

$$V(x,y) = < V(x,\cdot), V(y,\cdot) > . \tag{2}$$

B. The Mercer kernel mapping: Mercer theorem is one of the fundamental theorems in reproducing kernel Hilbert space research [7]. Suppose V is a continuous symmetric positive function on a closed finite interval T $\times$ T. Let $\{\varphi_k(x),\ k = 1, 2,...\}$ be a sequence of normalized eigenfunction of V, and $\{\lambda_k,\ k = 1, 2, ...\}$ be the sequence of corresponding non-negative eigenvalues. In other word, for all integers k and j,

$$\int_T V(x,y)\varphi_k(x)dx = \lambda_k\varphi_k(y), \quad x,y \in T$$

$$\int_T \varphi_k(x)\varphi_j(x)dx = \delta_{k,j}$$

where $\delta_{k,j}$ is the Kronecker delta function, i.e., equal to 1 or 0 according as $k = j$ or $k \neq j$. Then

$$V(x,y) = \sum_{k=0}^{\infty}\lambda_k\varphi_k(x)\varphi_k(y)$$

where the series above converge absolutely and uniformly on T $\times$ T [11].

It follows that $V(x,y)$ can be rewritten as an inner product between two vectors in the feature space, i.e.,

$$V(x,y) = < \Pi(x), \Pi(y) >$$
$$\Pi : x \mapsto \sqrt{\lambda_k}\varphi_k(x), \qquad k = 1, 2, ... \tag{3}$$

Comparing equations (2) and (3), we notice that there is an equivalence between $V(x,\cdot)$ and $\Pi(x)$. Both functional mappings construct the reproducing kernel Hilbert space associated with the correntropy function $V(x,y)$. In the next

section, we will apply Mercer kernel mapping associated with correntropy function to principal component analysis.

III. CORRENTROPY PCA

Given a set of zero mean vector observations $\mathbf{x}_j$, $j = 1, \ldots, N$, $\mathbf{x}_j \in \mathbb{R}^L$, $\sum_{j=1}^N \mathbf{x_j} = \mathbf{0}$, CORRENTROPY PCA seeks a direction in the feature space such that the variance of the data projected onto this direction is maximized. Unlike the kernel method which transforms data into a feature space sample by sample, CORRENTROPY PCA maps data component-wise into a feature space, i.e., the RKHS associated with the correntropy function. By the equation (3) in the above section, we have

$$\begin{aligned} \Pi : \mathbf{R}^L &\mapsto F \\ \mathbf{x} \mapsto [\Pi(x_1), \Pi(x_2), \cdots, \Pi(x_L)]&, \end{aligned}$$

where x_i denotes the i^{th} component of the original input data sample $\mathbf{x}$. This nonlinear mapping transforms the component-wise data into a high dimensional RKHS which is associated with the correntropy function. By the definition of correntropy function, we have

$$\begin{aligned} &< \Pi(x_i), \Pi(x_j) >= V(x_i, x_j) \qquad (4) \\ &= E[\kappa(x_i, x_j)] = \frac{1}{N}\sum_{k=1}^{N} \kappa(x_{ik}, x_{jk}), \quad \forall\ i, j = 1, 2, ..., L, \end{aligned}$$

where x_{ik} is the i^{th} component of the k^{th} input data sample. The expectation runs over all the data samples.

Let us first assume the transformed data in the feature space is zero mean, and we will come back later to this issue in the next section. Then the covariance matrix of the transformed data in the feature space is given by

$$C = \frac{1}{L}\sum_{i=1}^{L} \Pi(x_i)\Pi^T(x_i)$$

We now have to find the eigenvalues $\lambda \geq 0$ and non-zero eigenvectors satisfying

$$C\mathbf{q} = \lambda\mathbf{q}.$$

All the solutions $\mathbf{q}$ must lie in the span of $\Pi(x_1), \ldots, \Pi(x_L)$, i.e., we can write $\mathbf{q}$ as the form of linear combination of all the $\Pi(x_1), \ldots, \Pi(x_L)$,

$$\mathbf{q} = \sum_{i=1}^{L} \beta_i \Pi(x_i) \qquad (5)$$

And we may instead consider the set of equations,

$$< \Pi(x_k), C\mathbf{q} >=< \Pi(x_k), \lambda\mathbf{q} >, \forall\ k = 1, \ldots, L. \qquad (6)$$

Combining equations (5) and (6), we get

$$\begin{aligned} &< \Pi(x_k), \frac{1}{L}\sum_{j=1}^{L} \Pi(j_i)\Pi^T(j_i) \cdot \sum_{i=1}^{L} \beta_i \Pi(x_i) > \\ &= \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{L} \beta_i < \Pi(x_k), \Pi(x_j) >< \Pi(x_j), \Pi(x_i) > \\ &= \lambda \sum_{i=1}^{L} \beta_i < \Pi(x_k), \Pi(x_i) >, \forall\ k = 1, \ldots, L. \qquad (7) \end{aligned}$$

By equation (4), we can define an $L \times L$ correntropy matrix V by

$$\begin{aligned} V_{ij} := E[\kappa(x_i - x_j)] = \frac{1}{N}\sum_{k=1}^{N} \kappa(x_{ik}, x_{jk}), \qquad (8) \\ \forall\ i, j = 1, 2, ..., L. \end{aligned}$$

Let k in (7) runs from 1 to L, and write the result in matrix form, we can get

$$V^2\overline{\beta} = L\lambda V\overline{\beta}, \qquad (9)$$

where $\overline{\beta}$ denotes the column vector with entries $\beta_1, \ldots, \beta_L$. It can be shown that the solutions of equation (9) are equivalent to the solutions to the following eigenvalue problem,

$$V\overline{\beta} = L\lambda\overline{\beta}, \qquad (10)$$

for nonzero eigenvalues.

For the purpose of principal component extraction, we need to compute the projections onto the eigenvectors $\mathbf{q}$ in the feature space. Let $\mathbf{x}$ be a test point, the projection of $\mathbf{x}$ onto the principal direction mapped back to input space is given by

$$P(\mathbf{x}) = \sum_{i=1}^{L} \beta_i \frac{1}{N}\sum_{j=1}^{N} \kappa(x_{ij}, x_i), \qquad (11)$$

this is so called a nonlinear principal component.

In summary, we need to take the following steps to compute the nonlinear principal components: (1) compute the correntropy matrix V by equation (8), where the expected value is substituted by the average, (2) compute its eigenvectors and eigenvalues through SVD, and (3) compute the projections of a test point onto the eigenvectors by (11).

IV. DISCUSSIONS

In this section, we will discuss some specific issues concerning the CORRENTROPY PCA. First issue is the data centering in feature space. In the derivation of CORRENTROPY PCA above, we assume data is centered in the RKHS. But this might not be true even the original data has been preprocessed to be zero mean since data centering in feature space and original input space is very different. Now we define the centered data in feature space as the following,

$$\begin{aligned} \overline{\Pi(x_i)} &= \Pi(x_i) - E[\Pi(x_i)] \\ &= \Pi(x_i) - \frac{1}{N}\sum_{k=1}^{N} \Pi(x_{ik}). \end{aligned}$$

Then the inner product between any two centered vectors in the feature space is given by

$$< \overline{\Pi(x_i)}, \overline{\Pi(x_j)} >=< \Pi(x_i), \Pi(x_j) > - < \Pi(x_i), \frac{1}{N}\sum_{m=1}^{N}\Pi(x_{jm}) > - < \frac{1}{N}\sum_{k=1}^{N}\Pi(x_{ik}), \Pi(x_j) > + < \frac{1}{N}\sum_{k=1}^{N}\Pi(x_{ik}), \frac{1}{N}\sum_{m=1}^{N}\Pi(x_{jm}) > = E[\kappa(x_i - x_j)] - \frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\kappa(x_{ik} - x_{jm}).$$

The term $\frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\kappa(x_{ik} - x_{jm})$ is called *cross information potential* between i^{th} and j^{th} dimension of the original data [9]. Notice that the correntropy function for the centered data in the feature space is the original correntropy function minus the cross information potential.

V. Simulations

In this section, we will present two experimental results to show the effectiveness of CORRENTROPY PCA in finding nonlinear principal directions. The first experiment compares the standard linear PCA and CORRENTROPY PCA to extract features from a two dimensional mixture of Gaussian distributed data. Specifically, the probability density function is a mixture of Gaussian modes with the following form

$$f(\mathbf{x}) = 1/2(\mathcal{N}(\mathbf{m_1}, \mathbf{\Sigma_1}) + \mathcal{N}(\mathbf{m_2}, \mathbf{\Sigma_2})),$$

where $\mathcal{N}(\mathbf{m_1}, \mathbf{\Sigma_1})$ and $\mathcal{N}(\mathbf{m_2}, \mathbf{\Sigma_2})$ are two Gaussian distribution with the mean vectors and variance matrices given by

$$\mathbf{m_1} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mathbf{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix},$$
$$\mathbf{m_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{\Sigma_2} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In figure 1, we plot the contours of the data and of the constant projections produced by linear PCA and CORRENTROPY PCA respectively. The principal directions would be perpendicular to these constant projections. 200 samples are used and kernel size is chosen to be 2. The result confirms that linear PCA only provides the linear directions that maximizes the variance. But since the underlying data is a mixture of two Gaussian modes, linear PCA fails to consider the directions of the individual modes but only averages these directions. On the contrary, CORRENTROPY PCA is more tuned to the underlying structure of the data in the input space. CORRENTROPY PCA generates a nonlinear principal direction that follows locally the directions of the individual modes so that the variance of principal component projected onto this nonlinear curve is maximized. The experiment shows that CORRENTROPY PCA is superior in describing the underlying structure of the data when compared to the linear PCA method.

Our second experiment compared the kernel PCA, proposed by Schölkopf *et al* in [13], with CORRENTROPY PCA.

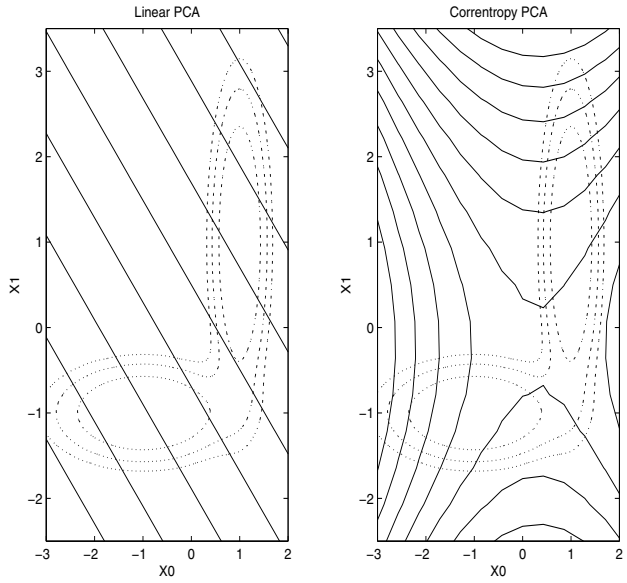

Fig. 1. Linear PCA versus CORRENTROPY PCA for a two-dimensional mixture of Gaussian distributed data

We use the same experiment setup as in [13] in order to illustrate the performance of CORRENTROPY PCA. The data is two-dimensional with three clusters(Gaussian distribution with standard deviation 0.1). The number of data samples and the kernel size are chosen to be 90 and 0.1 respectively. Since the number of principal components for kernel PCA depends on the number of data samples, there are many eigen-directions in feature space that are difficult to identify in the input space, so we plot the two principal components with the largest eigenvalues from kernel PCA. However the number of principal components for CORRENTROPY PCA is equal to the dimension of input space, so there is no ambiguity. Figure 2 shows that both kernel PCA and CORRENTROPY PCA can extract the nonlinear principal components form the data. While kernel PCA tends to find the local structure for a given data set as the contours circle around different data clusters suggest, CORRENTROPY PCA seeks the underlying global structure of the data set. The contour in the left bottom plot shows that CORRENTROPY PCA can be tuned to the data structure by changing the kernel size in the Gaussian kernel, and locate the principal direction.

In experiments comparing the performance of CORRENTROPY PCA with standard linear PCA and kernel PCA for nonlinear feature extraction, we found two advantages of our method. First, CORRENTROPY PCA can be more tuned to the underlying data structure than linear PCA so that it can extract the nonlinear principal components from the data, very much like principal curves. There is no ambiguity since the number of nonlinear principal components is the same as the dimensionality of the input space. In kernel PCA it is very difficult to choose the eigen-directions if we can not visualize the data, since the eigenvectors project locally to the input space. Therefore, it is not easy to separate major

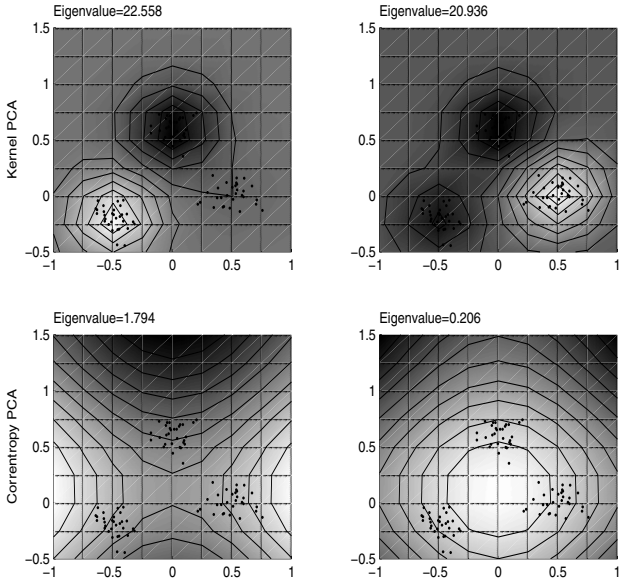

Fig. 2. Kernel PCA versus CORRENTROPY PCA for a two-dimensional mixture of Gaussian distributed data

and minor components. Second, CORRENTROPY PCA has a tremendous computational complexity advantage over kernel PCA. For example, in the second simulation, we only need to compute an eigen-decomposition for a 2×2 matrix using CORRENTROPY PCA while we have to do eigen-decomposition for a 90×90 matrix using kernel PCA. As the training set increases, the computational complexity of kernel PCA will increase dramatically but the size of the correntropy matrix remains the same. Of course the computation of each entry of the matrix, the correntropy between components increases with the square of the number of samples. New fast techniques $O(N)$ to compute each entry have been developed [3].

VI. CONCLUSION

In this paper we have presented a novel approach for principal component analysis using a new function called correntropy that generalizes the concept of auto-correlation to nonlinear spaces. Our approach is based on finding the eigenvectors of the correntropy matrix (same as the dimension of the data) unlike the Gram matrix used by other kernel methods, where the dimension of the Gram matrix is dependent on the number of the data. Yet, the final principle curves we get using this method adequately covers the data in the direction of maximum spread (variance in the feature space). Since we are dealing with a finite dimensional matrix, we get a number of principle curves equal to the dimension of the data space, and at the same time the computational complexity is drastically reduced compared to the kernel methods. In general this approach offers a new method of analyzing data. The study also suggests that the concept of correntropy can be used for de-correlating the data in the feature space (whitening), which can be applied in the context of independent component analysis. The future research will apply CORRENTROPY PCA to feature extraction in real data problems and also compare it with other nonlinear principal component analysis methods.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under grant ECS-0300340. The work of J.-W. Xu was supported by Alumni Graduate Fellowship from University of Florida. The work of A. R. C. Paiva was supported by Fundação para a Ciência e a Tecnologia under grant SFRH/BD/18217/2004.

REFERENCES

[1] N. Aronszajn, "Theory of Reproducing Kernels," *Trans. of American Mathematical Society*, Vol. 68, No. 3, pp. 337-404, May, 1950.
[2] K.I. Diamantaras, S.Y. Kung, *Principal Component Nerual Networks*, Wiley, New York, 1996.
[3] S. Han, S. Rao, J.C. Príncipe, "Estimating the Information Potential with the Fast Gauss Transform," in *Proc. of ICA*, 2006.
[4] T. Hastie, W, Stuetzle, "Principal Curves," *Journal of the American Statistical Association*, vol. 84, No. 406, pp. 502-516, 1989.
[5] I.T. Jolliffe, *Principal Component Analysis*, 2nd edn. Springer-Verlag, New York, 2000.
[6] M. Kramer, "Nonlinear Principal Component Analysis using Autoassociative Neural Networks," *AIchE Journal*, vol. 37 pp. 233-243, 1991.
[7] J. Mercer, "Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations,"*Philosophical Trans. of Royal Society of London, Series A,* Vol. 209, pp. 415-446, 1909.
[8] E. Parzen, "On the Estimation of a Probability Density Function and the Mode," *The Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.
[9] J.C. Príncipe, D. Xu, J.W. Fisher, "Information Theoretic Learning," in Unsupervised Adaptive Filtering, Edited by S.Haykin, Wiley, New York, pp. 265-319, 2000.
[10] A. Rényi, "On Measures of Entropy and Information," in Selected Paper of Alfréd Rényi. vol 2, Akademiai Kiado, Hungary, pp. 565-580, 1976.
[11] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, New York, Ungar, 1955.
[12] I. Santamaría, P.P.Pokharel, J.C. Príncipe, "Generalized Correlation Function: Definition, Properties and Application to Blind Equalization," to appear in *IEEE Trans. on Signal Processing*.
[13] B. Schölkopf, A. Smola, K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299-1219, 1998.