AN INFORMATION-THEORETIC PERSPECTIVE TO KERNEL INDEPENDENT COMPONENTS ANALYSIS

Jian-Wu Xu[1], Deniz Erdogmus[2], Robert Jenssen[3], Jose C. Principe[1]

[1]CNEL, Dept. of Electrical and Computer Engineering, University of Florida, USA
[2]Dept. of Computer Science and Engineering, Oregon Graduate Institute, OHSU, USA
[3]Dept. of Physics, University of Tromsø, Norway

ABSTRACT

In this paper, we investigate the intriguing relationship between information-theoretic learning (ITL), based on weighted Parzen window density estimator, and kernel-based learning algorithms. We prove the equivalence between kernel independent component analysis (KERNEL ICA) and Cauchy-Schwartz (C-S) independence measure. This link gives a theoretical motivation for the selection of the Mercer kernel, based on density estimation. Demonstrating this equivalence requires introducing a weighted kernel density estimator, a modification of Parzen windowing. We also discuss the role of the weights in the weighted Parzen windowing and KERNEL ICA.

1. INTRODUCTION

Kernel-based learning algorithms have been developed in the machine learning community during the last decades. With the introduction of support vector machine (SVM) theory [1], kernel Fisher discriminant (KFD) [2], and kernel principal component analysis (KPCA) [3], one is able to obtain nonlinear algorithms from linear ones in a simple and elegant way. Kernel-based algorithms are nonlinear versions of linear algorithms where the data has been nonlinearly transformed to a high dimensional feature space where we only need to compute the inner product via the kernel function. The attractiveness of kernel-based algorithms resides in their elegant treatment of nonlinear problems and efficiency for high-dimensional problems. Kernel methods have been successfully applied to time series prediction, DNA and protein analysis, optical pattern and object recognition [4].

Recently, Bach *et al* proposed a class of kernel-based algorithms for independent component analysis (ICA) that utilize contrast functions based on canonical correlations in a reproducing kernel Hilbert space, named KERNEL ICA [5]. The KERNEL ICA is based on novel kernel-based measures of dependence and can be computed efficiently. Minimizing these criteria results in flexible and robust ICA algorithms. One problem with all kernel methods is that it is not theoretically clear how to choose the best kernel function. The most commonly used kernel function is the Gaussian kernel, but it is still an open question how to select the width of Gaussian kernel in general.

In parallel to the developments in kernel-based methods research, independently a research topic called *information-theoretic learning* (ITL) has emerged [6], where kernel-based density estimators form the essence of this learning paradigm. Information-theoretic learning is a signal processing technique that combines information theory and adaptive systems. ITL utilizes information theory as a criterion to update the structure of adaptive system in order to achieve a certain performance. By utilizing Renyi's measure of entropy and approximations to the Kullback-Leibler probability density divergence, ITL is able to extract information beyond second-order statistics directly from data in a non-parametric manner. Information-theoretic learning has achieved excellent results on a number of learning scenarios, e.g. blind source separation [7, 8], time series prediction [9].

In this paper, we examine the KERNEL ICA from an information-theoretic learning perspective. We show that KERNEL ICA is equivalent to minimizing the Cauchy-Schwartz independence measure, when estimated via weighted Parzen windowing, though they have different normalizations. Based on the discussions in this paper, we conjecture that the kernel-based algorithms, including the KERNEL ICA, which are expressed in terms of inner products in the kernel feature space, are in fact learning implicitly by using non-parametric estimates of probability densities in the input space. This new view gives a geometrical interpretation for KERNEL ICA and theoretical criterion for choosing the Mercer kernel used in the kernel-based algorithms such that it would lead to a relatively accurate estimate when used as the Parzen windowing in density estimation. Before we proceed to that, we first show that how the most widely used ITL cost functions, when estimated by Parzen windowing, can

This work was supported by NSF grant ECS-0300340.

be expressed in terms of inner products in a reproducing kernel Hilbert space.

This paper is organized as follows. We review the basic theory of nonlinear kernel feature space and the KERNEL ICA in section 2. Cauchy-Schwartz independence measure is introduced in section 3. Afterwards, in section 4, we show how some of ITL cost functions can be written into quantities defined in the Hilbert feature space via Parzen windowing and prove the equivalence between the KERNEL ICA and Cauchy-Schwartz independence measure. Furthermore, we discuss the role of weights, used in weighted Parzen windowing for probability density estimation and the corresponding kernel space.

2. KERNEL ICA

Kernel-based learning algorithms use the following idea: via a nonlinear mapping

$$\Phi : \Re^t \to \mathcal{F} \qquad \mathbf{x} \to \Phi(\mathbf{x}) \tag{1}$$

the data in the input space $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N} \in \Re^t$ is mapped to a potentially much higher dimensional feature space $\mathcal{F}$. Instead of considering the given learning problem in input space $\Re^t$, one can deal with $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_N)$ in feature space $\mathcal{F}$. When the learning algorithms can be expressed in terms of inner products, this nonlinear mapping becomes particular interesting and useful since one can employ the *kernel trick* to compute the inner products in the feature space via *kernel functions* without knowing the exact nonlinear mapping Φ. This way of addressing the given learning problems allows one to obtain nonlinear algorithms from linear ones in a simple and elegant manner. In essence, by Mercer's theorem [10], the eigen-decomposition of a positive function (the kernel) is utilized to define the following inner product for the transformation space:

$$\kappa_\sigma(x - x') = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x) \varphi_k(x') = \langle \Phi(x), \Phi(x') \rangle \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product, the φ_k's are the eigen-functions of the kernel and λ_k's are the associated eigenvalues. The KERNEL ICA presented by Bach *et al* is a new method to ICA based on a kernel-measure of independence [5]. KERNEL ICA assumes a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ with kernel $\kappa(x, x')$ and feature map $\Phi(x) = \kappa(\cdot, x)$. Then the $\mathcal{F}$-correlation is defined as the maximal correlation between the two random variables $f_1(x_1)$ and $f_2(x_2)$, where f_1 and f_2 range over $\mathcal{F}$:

$$\begin{aligned} \rho &= \max_{f_1, f_2} corr(f_1(x_1), f_2(x_2)) \\ &= \max_{f_1, f_2} \frac{\operatorname{cov}(f_1(x_1), f_2(x_2))}{\sqrt{(\operatorname{var} f_1(x_1))(\operatorname{var} f_2(x_2))}} \end{aligned} \tag{3}$$

Clearly, if the random variables x_1 and x_2 are independent, then the $\mathcal{F}$-correlation is zero. Moreover, the converse is also true provided that the set $\mathcal{F}$ is large enough. This means that $\rho = 0$ implies x_1 and x_2 are independent.

In order to obtain a computationally tractable implementation of $\mathcal{F}$-correlation, the *reproducing property* of RKHS is used to estimate the $\mathcal{F}$-correlation,

$$f(x) = \langle \Phi(x), f \rangle = \langle \kappa(\cdot, x), f \rangle \tag{4}$$

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the linear spaces spanned by the Φ-images of the data samples, then f_1 and f_2 can be decomposed into two parts, i.e.

$$f_1 = \sum_{k=1}^{N} \alpha_1^k \Phi(x_1^k) + f_1^{\perp}, \quad f_2 = \sum_{k=1}^{N} \alpha_2^k \Phi(x_2^k) + f_2^{\perp} \tag{5}$$

where $f_1^{\perp}$ and $f_2^{\perp}$ are orthogonal to $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. Using the empirical data to approximate the population value, the $\mathcal{F}$-correlation can be estimated as

$$\hat{\rho} = \max_{\alpha_1, \alpha_2 \in \Re^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{(\alpha_1^T K_1^2 \alpha_1)(\alpha_2^T K_2^2 \alpha_2)}} \tag{6}$$

where K_1 and K_2 are the Gram matrices associated with the data sets $\{x_1^k\}$ and $\{x_2^k\}$ defined as $(K_i)_{a,b} = \kappa(x_i^a, x_i^b)$.

In the paper [5], Bach *et al* used a regularized version for the expression (6) by penalizing the RKHS norms of f_1 and f_2 in the denominator because (6) is not a consistent estimator in general. The regularized estimator has the same independence characterization property of the $\mathcal{F}$-correlation as (6), since it is the numerator, $\alpha_1^T K_1 K_2 \alpha_2$, in the $\mathcal{F}$-correlation that characterizes the independence property of two random variables. The difference between the direct estimator (6) and the regularized version is only the normalization. This also can be seen in section 4 when we prove the equivalence between the KERNEL ICA and Cauchy-Schwartz (C-S) independence measure.

3. CAUCHY-SCHWARTZ INDEPENDENCE MEASURE

In this section, we introduce the Cauchy-Schwartz (C-S) independence measure, which has been utilized as a cost function in independent component analysis (ICA) [7] and clustering [11].

In information theory, mutual information is a quantity

that characterizes the divergence between two random variables. A well-known divergence measure is the Kullback-Leibler distance

$$K(f,g)=\int f(x)\log\frac{f(x)}{g(x)}dx \qquad (7)$$

where *f(x)* and *g(x)* are two probability density functions (pdf). The Kullback-Leibler measure is difficult to evaluate in practice, without imposing simplifying assumptions about the data, since numerical methods are required to evaluate the integrals. In order to elegantly integrate the non-parametric pdf estimation via Parzen windowing [12], Principe *et al* proposed a new pdf distance measure based on Cauchy-Schwartz inequality between two vectors [6]. Thus we can evaluate pdf distance measure without making any parametric assumptions about the underlying pdfs.

Based on the Cauchy-Schwartz inequality, $\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geq (\mathbf{x}^T\mathbf{y})^2$, we can write $-\log \mathbf{x}^T\mathbf{y}\Big/\sqrt{\|\mathbf{x}\|^2\|\mathbf{y}\|^2} \geq 0$. Replacing the inner product between vectors by inner product between pdfs, we can define the Cauchy-Schwartz independence measure as

$$D_{CS}(f,g)=-\log\frac{\int f(x)g(x)dx}{\sqrt{(\int f^2(x)dx)(\int g^2(x)dx)}} \qquad (8)$$

Notice that $D_{CS}(f,g)\geq 0$ and the equality holds if and only if *f(x)=g(x)*. For two random variables X_1 and X_2, with marginal pdfs $f_1(x_1)$ and $f_2(x_2)$ and joint pdf $f_{1,2}(x_1,x_2)$, the Cauchy-Schwartz independence measure becomes

$$\begin{aligned} D_{CS}(f_1,f_2) &= -\log J \\ &= -\log\frac{\iint f_{1,2}(x_1,x_2)f_1(x_1)f_2(x_2)dx_1dx_2}{\sqrt{(\iint f_{1,2}^2(x_1,x_2)dx_1dx_2)(\iint f_1^2(x_1)f_2^2(x_2)dx_1dx_2)}} \\ &= -\log\frac{E[f_1(x_1)f_2(x_2)]}{\sqrt{E[f_{1,2}(x_1,x_2)]E[f_1(x_1)]E[f_2(x_2)]}} \end{aligned} \qquad (9)$$

As can be seen from above that $D_{CS}(f_1,f_2)\geq 0$. If and only if the two random variables are statistically independent, then $D_{CS}(f_1,f_2)=0$. Hence minimization of Cauchy-Schwartz independence measure leads to minimization of mutual information between two random variables. This is exactly the idea that Cauchy-Schwartz independence measure can be used as a criterion to characterize independence for ICA in [7].

In the next section, we will proceed to prove that the KERNEL ICA is equivalent to Cauchy-Schwartz independence measure, when estimated via weighted Parzen windowing.

4. EQUIVALENCE BETWEEN KERNEL ICA AND C-S INDEPENDENCE MEASURE

In this section, we first show how some widely used cost functions in information-theoretic learning can be estimated directly from data sample through Parzen windowing method. More importantly, these cost functions can be written in terms of inner products in a reproducing kernel Hilbert space, where the Mercer kernel is the windowing function used in Parzen density estimation. Then the proof of equivalence between KERNEL ICA and C-S independence measure will follow naturally.

4.1. ITL Cost Functions in the Kernel Space

One of the most commonly used cost functions in information-theoretic learning is the quadratic Renyi's entropy because it can be easily integrated with the Parzen window estimator [6], thus provides a simple way to estimate the entropy directly from the data samples.

Given the pdf *f(x)* for a random variable *X*, quadratic Renyi's entropy is defined as

$$H(X)=-\log\int f^2(x)dx=-\log E[f(x)] \qquad (10)$$

Since logarithm is a monotonic function, the quantity of interest is its argument $V(X)=\int f^2(x)dx$, which is called *information potential*. For a given pdf *f(x)*, a non-parametric asymptotically unbiased and consistent estimator is given by [12]

$$\hat{f}(x)=\frac{1}{N}\sum_{i=1}^{N}\kappa(x,x_i) \qquad (11)$$

where $\kappa(.)$ is called the Parzen window, or kernel. It is often chosen to be the Gaussian kernel though other kernels are also available, e.g., polynomial kernels. Then approximating the expectation by sample mean, we can estimate the *information potential* direct from data

$$V(X)=\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa(x_i-x_j) \qquad (12)$$

Notice that $\kappa(.)$ is a Gaussian kernel function, Hence we can employ (2) to rewrite (14) as

$$\begin{aligned} V(X) &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\left\langle\Phi(x_i),\Phi(x_j)\right\rangle \\ &= \left\langle\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i),\frac{1}{N}\sum_{j=1}^{N}\Phi(x_j)\right\rangle=\left\|\mathbf{m}^{\Phi}\right\|^2 \end{aligned} \qquad (13)$$

where $\mathbf{m}^{\Phi}$ is the mean vector of the transformed data. Thus, the quadratic information potential turns out to be the inner product of the mean vector of the nonlinearly transformed data in the Hilbert kernel space.

4.2. Equivalence of KERNEL ICA and C-S Independence Measure

To prove the equivalence between KERNEL ICA and Cauchy-Schwartz independence measure, we use weighted Parzen windowing. For a given marginal pdf *f(x)*, the weighted Parzen windowing density estimator is given by

$$\hat{f}(x) = \frac{1}{A}\sum_{i=1}^{N} \alpha_i \kappa(x, x_i) \qquad (14)$$

When Cauchy-Schwartz independence measure (10) is used as a contrast function in ICA, it should be minimized so that the mutual information between random variables is also minimized. As logarithm is a monotonic function, minimizing the C-S quantity is equivalent to maximizing its argument. Approximating the expectation by sample mean in (10) and estimating pdfs with weighed Parzen windowing, we can get

$$\hat{J} = \max_{\alpha_1,\alpha_2 \in \Re^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{(V)(1^T K_1 \alpha_1)(1^T K_2 \alpha_2)}} \qquad (15)$$

where $V = \sum_{j=1}^{N}\sum_{i=1}^{N} \alpha_1^i \kappa(x_1^i, x_1^j)\kappa(x_2^i, x_2^j)\alpha_2^i$,

$1=[1,1,\ldots.]^T$, and $(K_i)_{a,b} = \kappa(x_i^a, x_i^b)$.

Comparing expressions (15) and (6), we notice that they have same numerators and different normalizations. As we already pointed out in section 2, the numerator in KERNEL ICA characterizes the independence measure of two random variables whereas the denominator gives the certain normalization. Hence we conclude that the Cauchy -Schwartz independence measure, estimated via weighed Parzen windowing, is equivalent to the KERNEL ICA.

4.3. Role of the Weights

Recently we showed that the SVM is related to ITL and non-parametric pdf estimation via weighted Parzen windowing. The weights in the Parzen windowing there is associated with the support vectors in SVM [13]. In the Cauchy-Schwartz independence measure with weighted Parzen windowing estimation, we notice that those weights are associated with the coordinates of nonlinear function f_1 and f_2 in the linear spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively.

6. CONCLUSIONS

In this paper, we discuss the connection between information-theoretic learning (ITL), based on the weighted Parzen window density estimator that we have introduced. We demonstrated that the KERNEL ICA algorithm evaluates the independence between the separated outputs through a measure that is equivalent to the C-S mutual information estimated using the weighted Parzen windowing procedure. This discussion reveals an intriguing duality between the Mercer kernels and Parzen windowing (i.e., kernel density estimation). This duality provides a theoretical criterion for selecting the Mercer kernel in kernel-methods for machine learning and signal processing.

7. REFERENCES

[1] V. N. Vapnik, *The Nature of Statistical Learning Theory,* Springer-Verlag, NY, 1995.

[2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller, " Fisher Discriminant Analysis with Kernels," Proc. NNSP'99*,* pp. 41-48, Piscataway, NJ, 1999.

[3] B. Schölkopf, A. J. Smola, and K. R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299-1319, 1998.

[4] B. Schölkopf and A. J. Smola, *Learning with Kernels.* MIT Press, 2001.

[5] F. R. Bach, M. I. Jordan, "Kernel Independent Component Analysis", *Journal of Machine Learning Research,* vol. 3, pp. 1-48, 2002.

[6] J.C. Principe, D. Xu, J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, (Ed. S. Haykin), Wiley, NY, 2000.

[7] D. Xu, J. C. Principe, J. Fisher III, H. –C, Wu, "A novel measure for independent component analysis (ICA)," *Proc. ICASSP'98*, vol. 2, pp. 12-15, 1998.

[8] K. E. Hild, D. Erdogmus, J. C. Principe, "Blind Source Separation using Renyi's Mutual Informaiton," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 174-176, 2001.

[9] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Trans. Neural Networks*, vol. 13, no. 5, pp. 1035-1044, 2002.

[10] J. Mercer, " Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations," *Philos. Trans. Roy. Soc. London,* vol. A. pp. 415-446, 1909.

[11] R. Jenssen, J. C. Principe and T. Eltoft, "Cauchy-Schwartz pdf Divergence Measure for non-Parametric Clustering," *Proc. NORSIG'03,* Bergen, Norway, 2003.

[12] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, CA, 1967.

[13] R. Jenssen, D. Erdogmus, J. Principe, T. Eltoft, "Towards a Unification of Information Theoretic Learning and Kernel Methods," *Proc. MLSP'04*, Sao Luis, Brazil, 2004.