Signal Processing 91 (2011) 15–27

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A test of independence based on a generalized correlation function

Murali Rao [a], Sohan Seth [b,*], Jianwu Xu [b], Yunmei Chen [a], Hemant Tagare [a], José C. Príncipe [b]

[a] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA
[b] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 11 August 2009
Received in revised form
17 May 2010
Accepted 1 June 2010
Available online 8 June 2010

Keywords:
Correntropy
Test of independence

ABSTRACT

In this paper, we propose a novel test of independence based on the concept of correntropy. We explore correntropy from a statistical perspective and discuss its properties in the context of testing independence. We introduce the novel concept of parametric correntropy and design a test of independence based on it. We further discuss how the proposed test relaxes the assumption of Gaussianity. Finally, we discuss some computational issues related to the proposed method and compare it with state-of-the-art techniques.

1. Introduction

The concept of independence is perhaps the earliest and most fundamental building block of modern statistics [1]. It is undoubtedly an extraordinary concept that has not only shaped the field of statistics but has also opened up applications to a variety of research areas in experimental sciences, engineering and economics. Statistical signal processing and machine learning are two of the areas that have exploited this concept rigorously to come up with innovative tools to solve many practical problems. Independent component analysis (ICA) [2] and feature selection [3] are a few examples that have benefited immensely from this concept.

Although the use of independence in engineering is transparent, estimation of independence from data is still a difficult issue. A measure of independence is a bivariate statistic that takes zero value if and only if the independence condition is satisfied. Over the last century, numerous measures of independence have been proposed [4]. For example, distribution function based measure [5,6], density function based measures [7,8] and its variations such as Hellinger distance based measure [9], characteristic function based measure [10] and mutual information based measure [11], and quadratic form based measures [12–14]. However, these methods often suffer from higher ($\mathcal{O}(N^2)$ where N is the number of samples) computational cost and/or involvement of free parameters, such as choice of kernel.

In this paper, we discuss a novel test of independence from a very different perspective using the concept of correntropy. Correntropy is a generalization of correlation that extracts not only the second order information but also higher order moments of the joint distribution [15,16]. In the last few years, this concept has been successfully applied in several engineering applications such as time series modeling [17,18], non-linearity test [19], matched filtering [20], object recognition [21] and independent component analysis [22]. Although, correntropy is similar to correlation by definition, recent studies have shown that it performs better than correlation while dealing with non-linear systems and non-Gaussian noise environments, without any significant increase in the computational cost. Inspired by these results, we investigate the applicability of this method in designing a test of independence. We show that it is indeed possible to

* Corresponding author.
E-mail addresses: rao@math.ufl.edu (M. Rao), sohan@cnel.ufl.edu (S. Seth), jianwu@cnel.ufl.edu (J. Xu), yun@math.ufl.edu (Y. Chen), hemant.tagare@yale.edu (H. Tagare), principe@cnel.ufl.edu (J.C. Príncipe).

0165-1684/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2010.06.002

construct a test of independence using correntropy and the proposed test is $\mathcal{O}(N\log N)$ in computation and parameter free.

The rest of the paper is organized as follows. In Section 2, we describe formally the concept of correntropy. We explore correntropy from a statistical perspective and, in particular, investigate properties of correntropy in the context of test of independence. We show that correntropy exhibits properties very similar to those of correlation in many aspects, but, it is a more powerful similarity measure. In addition, we describe some properties of correntropy that leads to a better understanding of this concept. In Section 3, we introduce an extension of correntropy and design a test of independence. We introduce the concept of parametric correntropy and show that independence can be inferred if the value of parametric correntropy is zero for all parameter choices. We further extend this idea and show that with some assumptions on the underlying probability distribution it is possible to infer independence by checking only two parametric correntropy values. This result leads to a computationally effective test of independence that generalizes the assumption of Gaussianity and rivals the power of other recently introduced independent tests based on very different principles. In Section 4 we discuss some computational issues regarding the proposed work and show that under specific kernel the computational cost of the test of independence reduces to $\mathcal{O}(N\log N)$. We provide a brief overview of the available measures of independence and discuss their pros and cons compared to the proposed method and also the connections among these methods. Next, we describe some experimental results to validate the proposed method. Finally, in Section 5 we conclude the paper with a brief overview of the proposed work and some future research directions.

2. Correlation and correntropy

As mentioned before, correntropy is very similar to correlation. Therefore, it is trivial to extend this idea and define quantities that are equivalent to covariance and correlation coefficient. We call these quantities centered correntropy and correntropy coefficient, respectively. Note that these quantities have already been defined and applied in several engineering problems. However, as we would see, a few interesting properties of these statistics are yet to be explored. In this section we visit correntropy and related statistics from a statistical perspective and explore these properties.

Before proceeding to the definition of correntropy let us discuss the concept of non-negative definiteness.[1]

Definition 1 (*Non-negative definite functions*). A complex valued function $\kappa(x)$, defined on some set $\mathcal{X}$, is said to be non-negative definite if

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\overline{\alpha_j}\kappa(x_i-x_j)\geq 0$$

whenever $\{x_1, x_2,\ldots,x_n\}$ is a finite subset of $\mathcal{X}$ and $\{\alpha_1,\alpha_2,\ldots,\alpha_n\}$ is a finite set of complex numbers.

In this paper, we consider $\mathcal{X}$ to be the real line. The use of non-negative definite functions in engineering applications has recently gained immense popularity due to the advent of kernel machines. However, the use of non-negative definite kernels in this paper is inspired by the relation between non-negative definite functions and positive measures on the real line [24].

Theorem 1 (*Bochner's theorem*). *Every continuous non-negative definite function $\kappa(\cdot)$ on the real line has the following representation*

$$\kappa(z)=\int e^{-i\alpha z}\mu(\mathrm{d}\alpha)$$

for some finite positive measure μ on $\mathbb{R}$, that is, $\kappa(\cdot)$ is the Fourier transform of a positive measure μ.

We will use this particular property of a non-negative definite kernel to prove the properties of correntropy and related quantities.

2.1. Correntropy, centered correntropy and correntropy coefficient

Let $\kappa(\cdot)$ be a real valued, continuous, symmetric and non-negative definite kernel, then, correntropy is defined in the following way.

Definition 2 (*Correntropy*). Given two random variables X and Y, correntropy is defined as

$$V(X,Y)=\mathbf{E}_{X,Y}[\kappa(X-Y)]=\iint\kappa(x-y)\,\mathrm{d}F_{X,Y}(x,y), \tag{1}$$

where $\mathbf{E}$ is the expectation operator and $F_{X,Y}(x,y)$ is the joint probability distribution function of X and Y.

Given realizations $\{(x_1, y_1), (x_2, y_2),\ldots,(x_N, y_N)\}$ correntropy can be estimated using *strong law of large numbers* and therefore the estimation is consistent. It can be seen that if $\kappa(x,y)=xy$ then correntropy actually becomes correlation[2] (see Fig. 1). However, in this paper we only concentrate on shift invariant kernels of the form $\kappa(x,y)=\kappa(x-y)$ that obeys Bochner's theorem. We leave the extension of correntropy to a more general set of kernels as the future work.

Definition 3 (*Centered correntropy*). Given two random variables X and Y, centered correntropy is defined as

$$\begin{aligned}U(X,Y)&=\mathbf{E}_{X,Y}[\kappa(X-Y)]-\mathbf{E}_X\mathbf{E}_Y[\kappa(X-Y)]\\&=\iint\kappa(x-y)\{\mathrm{d}F_{X,Y}(x,y)-\mathrm{d}F_X(x)\,\mathrm{d}F_Y(y)\},\end{aligned} \tag{2}$$

[1] In the literature non-negative functions are sometimes referred as *positive definite* functions [23].

[2] The correlation kernel $\kappa(x,y)=xy$ is also a non-negative definite kernel. However, it is not shift invariant but a separable kernel [23].

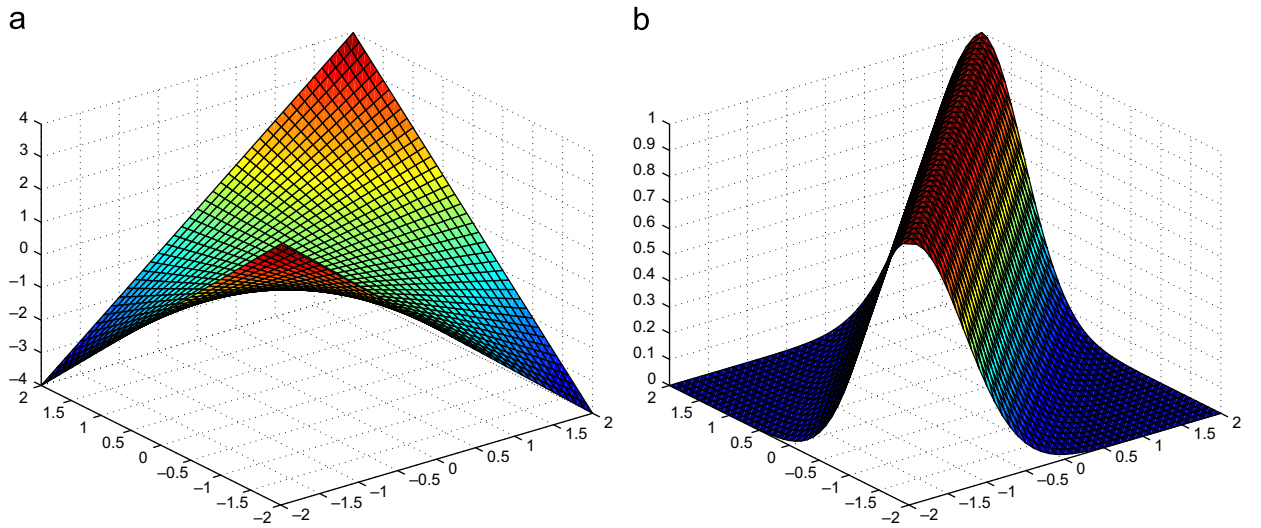

Fig. 1. The figure compares the kernels used for defining correlation and correntropy, respectively. (a) Kernel for correlation and (b) kernel (Gaussian) for correntropy.

where $F_X(x)$ and $F_Y(y)$ are the marginal probability distribution functions of X and Y, respectively.

Given realizations, centered correntropy can again be estimated using strong law of large numbers. Also, note that if $\kappa(x,y)=xy$ then centered correntropy reduces to covariance.

Definition 4 (*Correntropy coefficient*). Given two random variables X and Y, neither of them being a constant with probability 1, the correntropy coefficient is defined as

$$\eta(X,Y)=\frac{U(X,Y)}{\sqrt{U(X,X)U(Y,Y)}}, \tag{3}$$

where $U(X,Y)$ is the centered correntropy of X and Y, and $U(X,X)$ and $U(Y,Y)$ are the centered *autocorrentropy* of X and Y, respectively.

2.2. Properties of correntropy, centered correntropy and correntropy coefficient

Correntropy, centered correntropy and correntropy coefficient exhibit very similar properties as those of correlation, covariance and correlation coefficient. Below we state some of properties to demonstrate this fact. Since these properties are not exactly related to our objective, we provide the proofs in the Appendix.

Property 1. *$V(X,X)>0$ and $U(X,X)\geq 0$ with $U(X,X)=0$ if and only if X is degenerate.*

Property 2. *Both $V(X,Y)$ and $U(X,Y)$ are symmetric non-negative definite functions on the space of random variables.*

Property 3. *$|V(X,Y)|\leq\sqrt{V(X,X)V(Y,Y)}$ and $|U(X,Y)|\leq\sqrt{U(X,X)U(Y,Y)}$.*

Corollary 1. *$-1\leq\eta(X,Y)\leq 1$.*

Property 4. *$\eta(X,Y)=1$ if and only if $Y=X$.*

These properties are very similar to the properties of correlation and related statistics and play a crucial role in many applications where correlation is replaced with correntropy. Being positive definite correntropy induces a metric in the space of random variables named the correntropy induced metric (CIM). This metric has been used in many applications such as [25]. Like correntropy, centered correntropy also induces a metric in the space of random variables called the centered correntropy induced metric (CCIM). CCIM is defined as follows:

$$\begin{aligned}\mathrm{CCIM}(X,Y)&=\sqrt{U(X,X)+U(Y,Y)-2U(X,Y)}\\&=\sqrt{2\kappa(0)-\mathbf{EE}[\kappa(X-X')]-\mathbf{EE}[\kappa(Y-Y')]-2\mathbf{E}[\kappa(X-Y)]+2\mathbf{EE}[\kappa(X-Y)]},\end{aligned}$$

where (X',Y') is independent copy of (X,Y). Like CIM, CCIM can also be treated as a local similarity measure between X and Y [16]. However, unlike CIM, CCIM contains information about the joint distribution as well as the product of the marginal distributions. This inspires us to extend this local measure to a global measure and design a test of independence.

Interestingly enough, the properties exhibited by correntropy and correlation are very similar in this context, too. We show this in the following propositions.[3]

Proposition 1. *If X and Y are jointly normal random variables then centered correntropy of $X-\mathbf{E}[X]$ and $Y-\mathbf{E}[Y]$ is zero if and only if the random variables are independent.*

[3] In the rest of the section, we use the fact that

$$U(X,Y)=\int \mathrm{cov}(e^{-izX},e^{-izY})\mu(\mathrm{d}\alpha).$$

Proof. If X is a normal random variable with mean m and variance σ^2, then

$$\mathbf{E}[e^{i\alpha X}] = \exp\left(im\alpha - \frac{\sigma^2\alpha^2}{2}\right). \tag{4}$$

Let X and Y be jointly normal and $X\sim\mathcal{N}(m_1,\sigma_1^2)$ and $Y\sim\mathcal{N}(m_2,\sigma_2^2)$, then $X-Y$ is also normally distributed and $X-Y\sim\mathcal{N}(m_1-m_2,\sigma_1^2+\sigma_2^2-2\mathrm{cov}(X,Y))$. Therefore $X-m_1\sim\mathcal{N}(0,\sigma_1^2)$, $Y-m_2\sim\mathcal{N}(0,\sigma_2^2)$ and $(X-m_1)-(Y-m_2)\sim\mathcal{N}(0,\sigma_1^2+\sigma_2^2-2\mathrm{cov}(X,Y))$ Using Eq. (4) we find

$$\begin{aligned}
&U(X-m_1,Y-m_2)\\
&\quad= \int \mu(\mathrm{d}\alpha)\exp\left\{\left(-\frac{\sigma_1^2+\sigma_2^2}{2}+\mathrm{cov}(X,Y)\right)\alpha^2\right\}\\
&\qquad-\int \mu(\mathrm{d}\alpha)\exp\left\{\left(-\frac{\sigma_1^2+\sigma_2^2}{2}\right)\alpha^2\right\}\\
&\quad= \int \mu(\mathrm{d}\alpha)\exp\left\{\left(-\frac{\sigma_1^2+\sigma_2^2}{2}\right)\alpha^2\right\}[\exp\{(\mathrm{cov}(X,Y))\alpha^2\}-1].
\end{aligned}$$

Thus $U(X-m_1,Y-m_2)>0$ if $\mathrm{cov}(X,Y)>0$, $U(X-m_1,Y-m_2)<0$ if $\mathrm{cov}(X,Y)<0$ and $U(X-m_1,Y-m_2)=0$ if $\mathrm{cov}(X,Y)=0$. Therefore if (X,Y) is jointly normal then $X-m_1,Y-m_2$ and hence X,Y are independent if and only if $U(X-m_1,Y-m_2)=0$.[4] □

Proposition 2. *Zero centered correntropy does not imply independence.*

Proof. We prove this using a counter example. Let $a_n>0$, $a_n=a_{-n}$ and $\sum_{n=-\infty}^{\infty}a_n=1$ that is $a_0+2\sum_{n=1}^{\infty}a_n=1$. Let μ put mass a_n at n. Then

$$\begin{aligned}
\kappa(x) &= \int e^{-2\pi i\alpha x}\mu(\mathrm{d}\alpha) = \sum_{n=-\infty}^{\infty} e^{-2\pi i x n}a_n\\
&= a_0+\sum_{n=1}^{\infty}(e^{-2\pi i x n}+e^{2\pi i x n})a_n = a_0+2\sum_{n=1}^{\infty}\cos(2\pi x n)a_n.
\end{aligned}$$

Here we change the definition of the kernel slightly by introducing an extra 2π term to simplify the proof. Note that $\kappa(\cdot)$ is non-negative definite but not necessarily positive and $\kappa(m)=a_0+2\sum_{n=1}^{\infty}a_n=1$ for all $m=0,\pm1,\pm2,\ldots$. Thus for any probability measure F concentrated on points (m,n), $m,n=0,\pm1,\pm2,\ldots$, we have

$$\int \kappa(x-y)\,\mathrm{d}F(x,y) = \sum\kappa(m-n)\,\mathrm{d}(m,n) = \sum\mathrm{d}F(m,n) = 1.$$

In particular we see

$$\int \kappa(x-y)\{\mathrm{d}F(x,y)-\mathrm{d}G(x,y)\} = 0$$

for all such measures. This completes the proof. □

[4] It can be shown that

$$\eta(X,Y) = \frac{1\Big/\sqrt{1+\frac{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}{\sigma^2}}-1\Big/\sqrt{1+\frac{\sigma_1^2+\sigma_2^2}{\sigma^2}}}{\sqrt{1-1\Big/\sqrt{1+\frac{2\sigma_1^2}{\sigma^2}}}\sqrt{1-1\Big/\sqrt{1+\frac{2\sigma_2^2}{\sigma^2}}}},$$

where σ is the kernel size of a Gaussian kernel. Therefore, when $\rho=0$, $\eta=0$ and when $\rho=1$, $\eta=1$ if $\sigma_1=\sigma_2$. The proof is omitted because it is straightforward.

From these propositions what we find is that, first, although centered correntropy is zero under independence, zero centered correntropy does not *imply* independence, and second, for Gaussian random variables independence can be inferred from centered correntropy. These properties are similar to those of covariance. However, correntropy is conceptually different from correlation and perhaps better than the latter in the context of detecting independence. For example, consider the following situation. If X and Y are random variables, then $\mathrm{cov}(X-a,Y)=\mathrm{cov}(X,Y)$ for all a. Therefore, $\mathrm{cov}(X,Y)=0$ implies $\mathrm{cov}(X-a,Y)=0$ for all a. Thus, no information is obtained by the parameter a. Suppose on the other hand $U(X-a,Y)=0$ for all a. Then,

$$\int e^{i\alpha a}\mathrm{cov}(e^{-i\alpha X},e^{-i\alpha Y})\mu(\mathrm{d}\alpha) = 0$$

for all a. This implies

$$\mathbf{E}[e^{-i\alpha(X-Y)}] = \mathbf{E}[e^{-i\alpha X}]\mathbf{E}[e^{i\alpha Y}]$$

μ almost everywhere. To appreciate the "degree" of independence this implies, let us assume that all the quantities, in α, above are entire and that support of μ has a limit point. Then, equating the coefficients of α^n on both sides of the equation we get, for all n

$$\begin{aligned}
\mathbf{E}[(X-Y)^n] &= \sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\mathbf{E}[X^k(-Y)^{n-k}]\\
&= \sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\mathbf{E}[X^k]\mathbf{E}[(-Y)^{n-k}],
\end{aligned}$$

which implies that the random variables are "almost independent". This example shows that, correntropy, conveys more information about independence than uncorrelatedness. But, we, still, cannot infer independence exactly, i.e. from a mathematical sense. However, as shown in [22] centered correntropy is an appropriate contrast function for independent component analysis (ICA). In the next section we discuss the necessary steps that would lead to an exact test of independence based on correntropy.

3. Test of independence

In the previous section we have shown that centered correntropy on its own cannot infer independence. In order to design an exact test of independence from centered correntropy, we introduce the concept of parametric centered correntropy.

3.1. Parametric correntropy and test of independence

Definition 5 (*Parametric centered correntropy*). Given two random variables X and Y, the parametric centered correntropy is defined as

$$\begin{aligned}
U_{a,b}(X,Y) &= \mathbf{E}_{X,Y}[\kappa(aX+b-Y)]-\mathbf{E}_X\mathbf{E}_Y[\kappa(aX+b-Y)]\\
&= \iint\kappa(ax+b-y)\{\mathrm{d}F_{X,Y}(x,y)-\mathrm{d}F_X(x)\,\mathrm{d}F_Y(y)\},
\end{aligned}$$

where a and b are scalars in $\mathbb{R}$ and $a\neq0$.

Using the concept of parametric centered correntropy we can design a test of independence due to the following lemma.

Lemma 1 (*Zero parametric centered correntropy and independence*). *Given two random variables X and Y the parametric centered correntropy* $U_{a,b}(X,Y) = 0$ *for all* $a,b \in \mathbb{R}$ *if and only if X and Y are independent.*

Proof.

1. ($\Leftarrow$) The sufficient condition is straightforward.
2. ($\Rightarrow$) Assume $U_{a,b}(X,Y)=0$ for all $a,b \in \mathbb{R}$. From the definition we have,

$$U_{a,b}(X,Y) = \iint \kappa(ax+b-y)\{\mathrm{d}F_{X,Y}(x,y) - \mathrm{d}F_X(x)\,\mathrm{d}F_Y(y)\} = \iint \kappa(ax+b-y)\,\mathrm{d}Q(x,y),$$

where $Q(x,y)=F_{X,Y}(x,y)-F_X(x)F_Y(x)$. Therefore to prove independence we need to show that $\mathrm{d}Q(x,y)=0$. Now, $U_{a,b}(X,Y)=0$ for all $a,b \in \mathbb{R}$ implies

$$\int e^{-i\alpha b} \iint e^{-i\alpha(ax-y)}\mathrm{d}Q(x,y)\mu(\mathrm{d}\alpha) = 0$$

for all $a,b \in \mathbb{R}$, in particular for all $b \in \mathbb{R}$. Hence,

$$\iint e^{-i\alpha(ax-y)}\,\mathrm{d}Q(x,y) = 0$$

for almost all α as μ is always positive. Since the support of μ is $\mathbb{R}$, this holds for all $\alpha,a \in \mathbb{R}$. This is easily written as

$$\int e^{-i(\alpha x+\beta y)}\,\mathrm{d}Q(x,y) = 0$$

for all $\alpha,\beta \in \mathbb{R}$. Thus we conclude that $dQ=0$.[5] □

What this lemma states is that two random variables are independent if and only if the parametric centered correntropy is zero for all possible parameter values. Therefore, using the lemma we can define a test of independence as follows:

Definition 6 (*Correntropy independence measure*). Given two random variables *X* and *Y*, *correntropy independence measure* is defined as follows:

$$\Gamma(X,Y) = \sup_{a,b}|U_{a,b}(X,Y)|, \tag{5}$$

where $a,b \in \mathbb{R}$.

$\Gamma(X,Y)$ is a valid measure of independence since it is zero if and only if *X* and *Y* are independent. However, a test of independence defined in such a way requires searching a two dimensional space which is computationally expensive. But, the search space can be reduced drastically if we put some assumptions on the underlying probability distribution. In the following section we address this issue.

3.2. Generalization of Gaussianity and approximate test of independence

Let *X*,*Y* be normal with mean zero and variance σ_1^2 and σ_2^2, respectively, that is $X \sim \mathcal{N}(0,\sigma_1^2)$ and $Y \sim \mathcal{N}(0,\sigma_2^2)$. Then the general non-degenerate, bi-Gaussian with marginals $\mathcal{N}(0,\sigma_1^2)$ and $\mathcal{N}(0,\sigma_2^2)$ has the following density:

$$f_\rho(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\,\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2\rho xy}{\sigma_1\sigma_2}\right\}\right],$$

where $|\rho| \le 1$. Thus for $0 \le p_i \le 1$, $\sum p_i = 1$, the mixture of Gaussian $\sum_{i=1}^m f_{\rho_i}(x,y)$ has marginals $\mathcal{N}(0,\sigma_1^2)$ and $\mathcal{N}(0,\sigma_2^2)$. Also, if (X,Y) has joint density of the form $\sum_{i=1}^m f_{\rho_i}(x,y)$ then $\mathrm{cov}(X,Y) = \sigma_1\sigma_2\sum_{i=1}^m p_i\rho_i$ and $\mathrm{cov}(X,-Y) = -\mathrm{cov}(X,Y)$ and $X-Y$ and $X+Y$ have densities

$$\sum_{i=1}^m p_i\mathcal{N}(0,\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2) \quad \text{and} \quad \sum_{i=1}^m p_i\mathcal{N}(0,\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2),$$

respectively.

Theorem 2. *Suppose* (X,Y) *has joint density of the form* $\sum_{i=1}^m f_{\rho_i}(x,y)$ *as described above. Then* (X,Y) *is independent if and only if* $U(X,Y)=U(X,-Y)=0$.

Proof. Using Eq. (4) we find

$$\begin{aligned} U(X,Y) &= \int \mu(\mathrm{d}\alpha)\left[\sum_{i=1}^m p_i \exp\left\{\left(-\frac{\sigma_1^2+\sigma_2^2}{2}+\rho_i\sigma_1\sigma_2\right)\alpha^2\right\} - \exp\left(-\frac{\sigma_1^2+\sigma_2^2}{2}\alpha^2\right)\right] \\ &= \int \mu(\mathrm{d}\alpha)\exp\left(-\frac{\sigma_1^2+\sigma_2^2}{2}\alpha^2\right)\left[\sum_{i=1}^m p_i\exp(\rho_i\sigma_1\sigma_2\alpha^2)-1\right]. \end{aligned} \tag{6}$$

Suppose $U(X,Y) = 0$. Then the integrand in Eq. (6) and hence

$$\sum_{i=1}^m p_i(\exp(\rho_i\sigma_1\sigma_2\alpha^2)-1)$$

assumes both positive and negative values. Without loss of generality, we replace α^2 by α. Then for $\alpha > 0$

$$\phi(\alpha) = \sum_{i=1}^m p_i(\exp(\rho_i\sigma_1\sigma_2\alpha)-1)$$

assumes both positive and negative values. Now $\phi(\alpha)$ is convex in $(0,\infty)$ and $\phi(0)=0$. Since ϕ is convex ϕ' is increasing. If ϕ assumes negative values, since $\phi(0)=0$, $\phi'(0)$ must assume negative value that is $\sum p_i\rho_i < 0$ and if $U(X,-Y)=0$ the argument above leads to $\sum p_i\rho_i > 0$. Therefore, if $U(X,Y) = 0$ and $U(X,-Y) = 0$ then $\sum p_i\rho_i = 0$ and hence, $\phi'(0)=0$ which implies $\phi(\alpha) \ge 0$. Then the integrand in (6) is zero μ almost everywhere and if μ has support $\mathbb{R}$, the integrand is identically equal to zero. Then all the $\rho_i = 0$ that is *X*,*Y* are independent. □

This theorem proves that the search space can be restricted drastically with appropriate assumptions on the underlying distribution. Although the assumed distribution is not the most flexible one, it can be easily

[5] The last line of the proof follows from the theory of Fourier transform that if the Fourier transform of a measure is zero everywhere then the measure is zero everywhere [10].

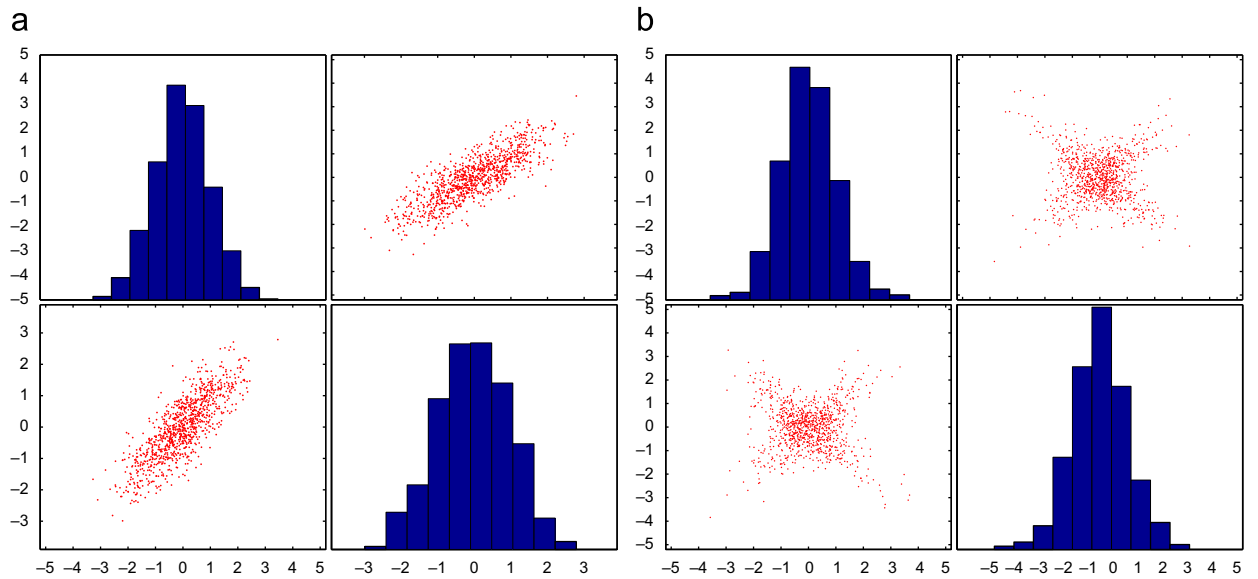

Fig. 2. The figure shows samples generated from the model described in Theorem 2. (a) $p_1 = 1$, $\rho_1 = 0.8$. (b) $p_1 = 0.5$, $p_2 = 0.5$, $\rho_1 = 0.8$, $\rho_2 = -0.8$.

seen that it is actually a generalization of the Gaussianity assumption, i.e., the Gaussianity assumption is a special case when $m = 1$ (see Fig. 2). Moreover, under the stated assumption our search space reduces to only two points which is easily tractable. Therefore, using this assumption, we can modify the proposed independence measure as follows:

Definition 7 (*Approximate correntropy independence measure*). Given two random variables X and Y *approximate correntropy independence measure* is defined as

$$\gamma(X,Y) = \max(|U(X,Y)|,|U(-X,Y)|). \tag{7}$$

Note that $\gamma(X,Y)$ is not an exact measure of independence since it may attain zero value under dependence. However, this measure is optimal under the assumption that the joint density can be expressed as a mixture of bivariate Gaussians with same mean.

4. Simulation

In the previous section, we have introduced two tests of independence; an exact test that does not rely on the underlying distribution and a computationally simpler test of independence that assumes a particular model for the underlying distribution. In this section, we describe some computational issues related to the estimators of these tests and then present some simulation results to corroborate the proposed ideas.

4.1. Computational issues

The proposed tests of independence require computing the parametric centered correntropy for different parameter values. Given realizations $\{(x_i,y_i)\}_{i=1}^N$, the estimator of centered correntropy is given by

$$\hat{U}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x_i-y_i) - \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\kappa(x_j-y_k).$$

Therefore, a straightforward computation of centered correntropy requires $\mathcal{O}(N^2)$ computation which is expensive for many applications.[6] However, the computation complexity can be reduced significantly by choosing Laplacian kernel.

Consider $\kappa(z)$ to be a Laplacian kernel of the form

$$\kappa(z) = e^{-|z|},$$

then,

$$\hat{U}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} e^{-|x_i-y_i|} - \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} e^{-|x_j-y_k|}.$$

In this expression the first term can be computed in $\mathcal{O}(N)$ time. To compute the second term we follow the approach proposed by [27]. Note that this term can be rewritten in the following way:

$$\sum_{j=1}^{N}\left[e^{-x_j}\sum_{\{k:y_k\le x_j\}} e^{y_k} + e^{x_j}\sum_{\{k:y_k > x_j\}} e^{-y_k}\right].$$

Now, let us assume that $\{y_i\}_{i=1}^N$ is sorted in ascending order. If the sequence is not sorted then it can be sorted in $\mathcal{O}(N\log N)$ time using any optimal sorting algorithm. Then,

[6] Since κ is non-negative definite, the estimator can be computed efficiently as described in [26]. However, in this paper we present a different and, perhaps, more efficient method exploiting the fact that the proposed tests are bivariate, i.e., both X and Y are one dimensional random variables.

the expression can just be written as

$$\sum_{j=1}^{n}\left[e^{-x_j}\sum_{k=1}^{K}e^{y_k}+e^{x_j}\sum_{k=K+1}^{n}e^{-y_k}\right],$$

where K is chosen such that $y_K \le x_j$ and $y_{K+1} > x_j$. Now lets assume that we have the cumulative sums

$$\left\{\underline{S}_K=\sum_{k=1}^{K}e^{y_k}\right\}_{K=1}^{N} \quad \text{and} \quad \left\{\overline{S}_K=\sum_{k=K+1}^{n}e^{-y_k}\right\}_{K=1}^{N}$$

of the sorted $\{y_i\}_{i=1}^{N}$. These sums can be computed in $\mathcal{O}(N)$ time. Finally using $\{e^{x_i}\}_{i=1}^{n}$, $\{e^{-x_i}\}_{i=1}^{n}$, $\{\underline{S}_i\}_{i=1}^{n}$ and $\{\overline{S}_i\}_{i=1}^{n}$, the double sum can be computed in $\mathcal{O}(N)$ time. Therefore, the overall complexity of computing centered correntropy using Laplacian kernel becomes $\mathcal{O}(N\log N)$ instead of $\mathcal{O}(N^2)$.

In order to compute $\Gamma(X,Y)$, we need to search a two dimensional space for supremum of a possibly non-linear function. Although, computing a single value of parametric correntropy is $\mathcal{O}(N\log N)$ in time, the search can still be expensive depending on the features of the surface. In order to establish a compromise between computation and accuracy, we suggest to use grid search. The resolution of the grid is, of course, user defined and a finer grid results in better accuracy. Moreover, the grid search only adds a multiplicative constant to the overall complexity, thus, keeping the overall complexity $\mathcal{O}(N\log N)$. We rewrite a as $\tan\theta$ and search over the grid $\theta = 0:\pi/40:\pi$ and $b = -2{:}0.5{:}2$. Note that other sophisticated optimization techniques such as the half quadratic optimization technique can also be applied to efficiently solve this problem [28].

4.2. Review of tests of independence

Test of independence between two random variables has been the focus of research for over half a century. In this section we provide a brief overview of the available methods. The most basic test of independence involves testing equality of the joint and the product of the marginal distribution functions. The two most popular tests of independence based on this approach are the Kolmogorov–Smirnov type test, that uses the maximum absolute deviation between the joint and the product of the marginal distribution functions as a test statistic [6], and the Cramér von–Mises type test (CM), that uses the L_2 distance between the joint and the product of the marginal distributions as a test statistic [5]. However, a distribution function based statistics, although easily computable, usually provide less power compared to density function based statistics. Popular test statistics based on density functions include L_2 distance between the joint and the product of the marginal densities [29] and the mutual information based measure [11]. However, this approach usually suffers from the selection of the appropriate kernel for the density estimation. Several modification of this approach such as using weighted L_2 distance (DC) has been proposed [8]; that does not involve any kernel. Other tests of independence involve comparing the characteristics functions of the joint and the product of the marginal distributions (ECF) [10]. Recently, a new test involving the Hilbert–Schmidt norm of the cross covariance operator (HSIC) has been proposed as a statistic [12] and it has been shown that this test is very similar to the quadratic dependence measure (QDM) proposed in [30]. However, this test has also been studied in a different context by [13]. It is interesting that HSIC, QDM, ECF and the L_2 distance based measures share the same estimator (see Appendix E). Correntropy, on the other hand, has a very different origin when compared with all these methods. It is defined as generalized correlation and corresponds to a novel similarity metric in the space of the random variables [16]. Therefore the correntropy independence test enriches our understanding between test statistics, independence and kernel methods thru QDM as established in Appendix F. Moreover, from Lemma 1, it is evident that the proposed method is strongly related to the difference between the joint and product of the marginal characteristic functions, i.e. $\int \exp(-i\alpha x - i\beta y)\,\mathrm{d}Q(x,y)$. However, this is different from the method proposed in [8], in the sense that it works with weighted L_2 distance between two characteristic functions whereas the proposed method implicitly links to only the difference between characteristic functions and not their distance.

All the statistics discussed above require $\mathcal{O}(N^2)$ space and time complexity. The proposed method, on the other hand, has only $\mathcal{O}(N\log N)$ complexity. In the following sections, we compare the proposed method against HSIC, DC and CM. For HSIC we follow the method described in [12] and set the kernel size parameter to the median distance between samples. We refer to the exact test of independence in (5) as ECI whereas the approximate test of independence in (7) as ACI. In the simulations presented below, we normalize the realizations two have zero mean and unit variance. We use a Laplacian kernel as described above. For all the experiments we use a permutation test to decide the critical threshold of rejection using 1000 surrogates and use another 1000 realizations to compute the empirical power.

4.3. Experiment I

In our first example we generate data from a Gaussian distribution with correlation coefficient $\rho = 0.5$. This is a simple problem since even correlation coefficient can detect the dependence. We test the performance of the proposed methods over different sample sizes, i.e. 25, 50, 100, 150 and 200. Fig. 3 shows the variation of power over different sample sizes for all the methods. In this particular example CM and DC perform better that the rest of the methods. CM should perform better in this example since it is a measure of monotone dependence [31]. It is interesting to see that DC performs equally well. Since the samples are from Gaussian distribution, ACI and ECI perform equally well and they perform slightly better than HSIC.

4.4. Experiment II

Next, we consider two independent random variables X and Y, both having zero mean and unit variance. X is a

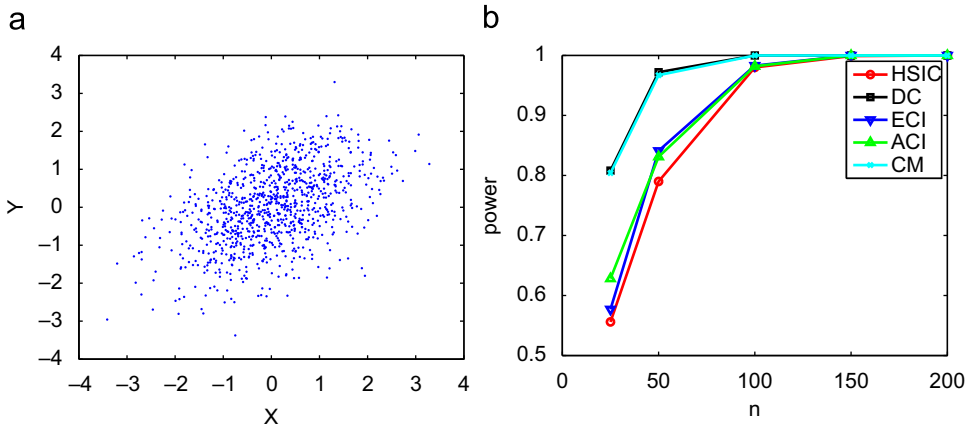

Fig. 3. The figure shows (a) samples from two random variables and (b) the variation of power over different sample sizes for the experiment described in Section 4.3.

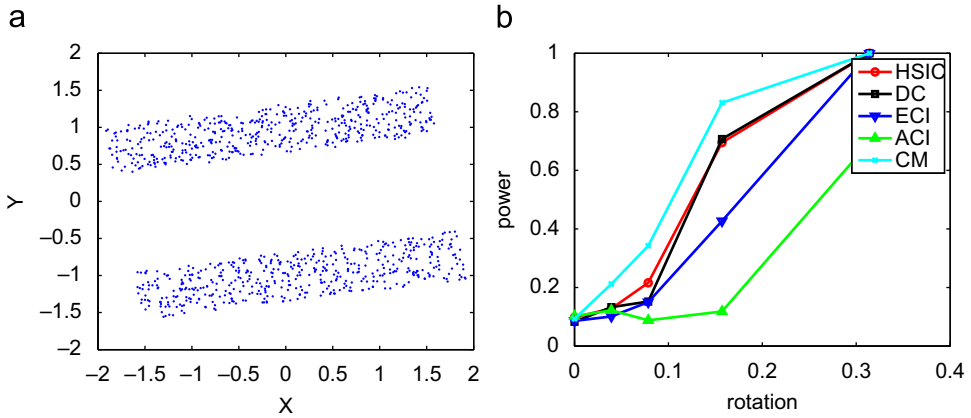

Fig. 4. The figure shows (a) samples from two random variables with rotation $\pi/20$ as described in the experiment in Section 4.4 and (b) variation of power over different rotations for different methods.

uniform random variable whereas Y is a combination of two uniform random variables each having equal probability of occurrence on disjoint support (see Fig. 4). The pair of random variables has an identity covariance matrix. Therefore if we generate a new pair of random variables (X',Y') by rotating this random variable pair (X,Y), the covariance matrix does not change and, thus, the correlation between (X',Y') stays the same, i.e. zero. However, the dependence between the random variables changes. The new variables are independent if and only if the angle of rotation is zero and dependent otherwise. We compute the empirical power of the independence test over different rotations. We use 100 samples. In this example CM again performs the best followed by DC and HSIC that perform equally well and then ECI that performs slightly worse than these methods. ECI fails to perform well on this example since the underlying assumption on the joint distribution is violated.

4.5. Experiment III

Next, we test the performance of the methods on a dataset described in [7]. Here, we consider three random variables X,Y and Z such that X and Y are jointly normal with zero mean and unit covariance matrix whereas Z is uniformly distributed between [0,2]. We construct two new random variables, $X'=XZ$ and $Y'=YZ$. (X',Y') are not independent as they share a common random variable. But this fact is not very clear just from the scatter plot. Once again, we test the performance of the methods over different sample sizes and plot the power over different sample sizes in Fig. 5. In this particular example, both CM and DC perform poorly than the other methods. In this example, ACI performs the best; even better than ECI. A possible reason behind this is the search involved in ECI that adds some inaccuracies in the method. However, ECI still performs similar to HSIC.

4.6. Experiment IV

Next, we test the performance of the methods on a dataset where the assumption made in Section 3.2 is satisfied. We sample data from three Gaussian distributions with correlation coefficients 0, 0.8 and -0.8 with probability 0.6, 0.2 and 0.2, respectively. These two random variables are dependent, but, this fact is not very

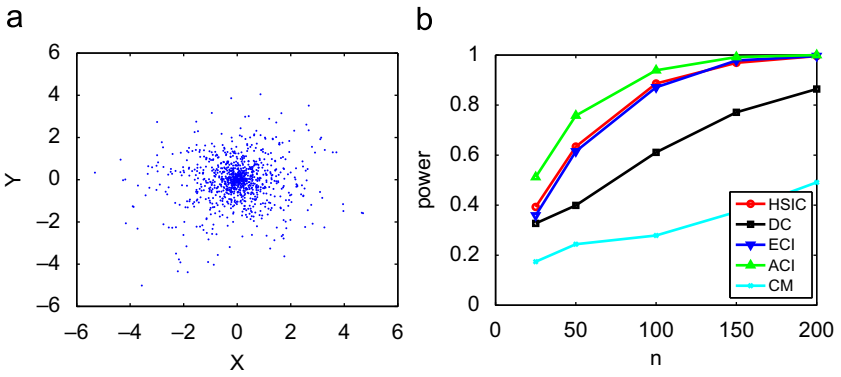

Fig. 5. The figure shows (a) samples from two random variables and (b) the variation of power over different sample sizes for the experiment described in Section 4.5.

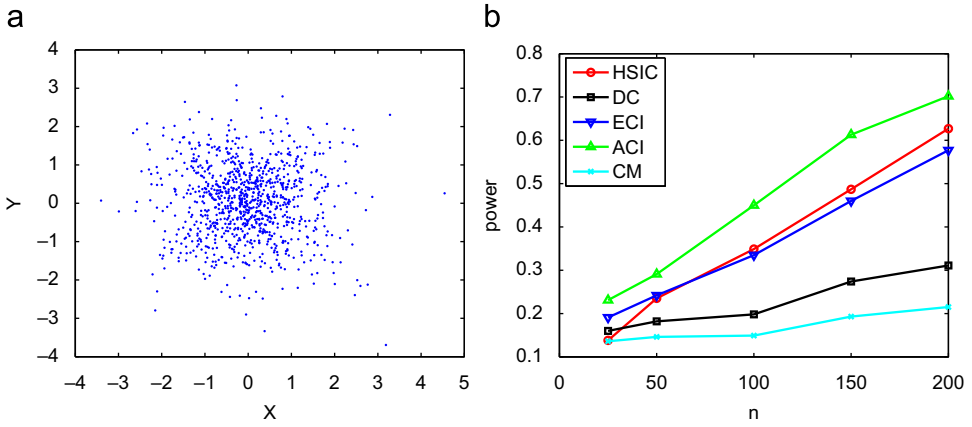

Fig. 6. The figure shows (a) the sample from two random variables and (b) the variation of power over different sample sizes for the experiment described in Section 4.6.

clear just from the scatter plot. Once again, we test the performance of the methods over different sample sizes and plot the power over different sample sizes in Fig. 6. In this particular example ACI again performs the best, followed by HSIC and ECI. These three methods outperform DC and CM. Once again, the worse performance of ECI than ACI is due to the search involved.

4.7. Experiment V

Next, we test the performance of the methods on a linear system corrupted by multiplicative noise. Here we generate data from a Gaussian random variable with correlation coefficient 0.8 such that the random variables share a strong linear connection and then multiply each variable with white Gaussian noise. These two random variables are dependent, but, this fact is not very clear just from the scatter plot. Once again, we test the performance of the methods over different sample sizes and plot the power over different sample sizes in Fig. 7. In this particular example, all the methods perform better than CM. Once again, ACI performs the best followed by HSIC and the rest. ECI performs similar to DC.

4.8. Experiment VI

In our final experiment we compare the computational cost of the proposed methods. We generate samples from the data described in experiment 4.7 for different sample sizes, i.e. 100, 500, 1000 and 2000 and evaluate the different methods 10 times each. We report the time taken to execute these methods for different sample sizes in Fig. 8. All the methods were written in Matlab 7.9. As expected, we observe an almost linear growth in ACI and ECI whereas an almost quadratic growth in the other algorithms. Note that ECI crosses DC at $n=500$. This happens because we search almost 400 grid points for the supremum. ACI, on the other hand, is extremely cost effective since it searches only two grid points.

5. Conclusion

In this paper, we have discussed a novel test of independence based on the concept of correntropy. Correntropy is a generalization of correlation that extracts not only the second order information but also higher

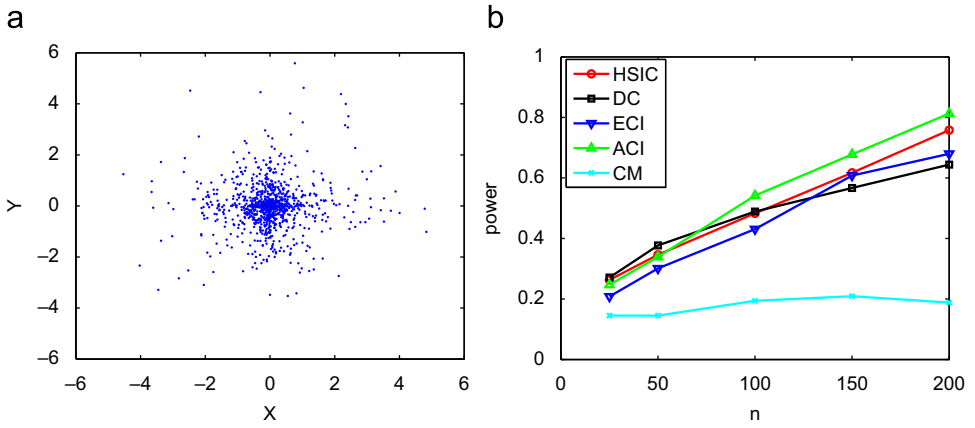

Fig. 7. The figure shows (a) samples from two random variables and (b) the variation of power over different sample sizes for the experiment described in Section 4.7.

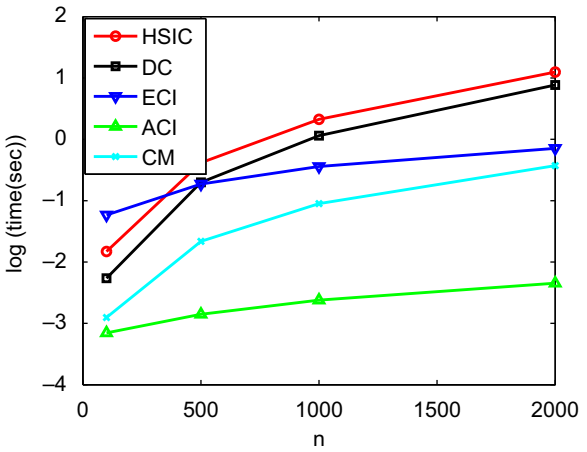

Fig. 8. The figure shows the variation of computation cost over different sample sizes for the experiment described in Section 4.8.

order moments from the joint density function. In recent years, it has been observed that correntropy provides better result than correlation in many diverse engineering applications; especially those involving non-linear and non-Gaussian signals. Inspired by this observation, we have explored the possible use of this concept as a measure of independence; something that cannot be achieved by correlation. In order to design a test of independence, we have introduced a new statistic called parametric centered correntropy and have shown that parametric correntropy is zero for all choices of parameters if and only if the random variables are independent. We have further shown that under some conditions on the underlying distribution, it is possible to infer independence just by checking only two parameter values instead. We have also discussed an efficient way to compute parametric centered correntropy and have compared the proposed methods against many other available ones. The various experimental results have shown the validity of the proposed method.

We have concentrated on the theoretical aspects of centered correntropy and have discussed its use to develop a test of independence. An interesting property of centered correntropy is that it is a positive definite function on the space of random variables, and thus, it induces a reproducing kernel Hilbert space (RKHS) [32]. From this perspective, the proposed method can be interpreted in a different way as follows. Lemma 1 shows that if the inner product (in the RKHS) between Y and all possible rotations (i.e. a) and translations (i.e. b) of X are zero then the two random variables are independent. This approach is similar to maximal correlation (MC) in the sense that in MC the covariance between all possible functions $f(X)$ and $g(Y)$ are considered [33]. We believe that this interpretation can be further exploited in designing effective measures of independence. Notice that this RKHS is significantly different than the RKHS described in HSIC.

Moreover, the proposed method is interesting in the sense that it extends the idea of correntropy and reveals its potential application in a different context, i.e. test of independence. We show that the (approximate) test relaxes the assumption of Gaussianity without any significant increase in the computational cost ($\mathcal{O}(N \log N)$ compared to $\mathcal{O}(N)$) and, thus, is applicable to a more general group of problems. Moreover, our simulation results show the (approximate) test performs well even when the underlying assumption is not satisfied in practice. However, we observe that there are instances where this test fails. But, in these situations the exact test performs comparable to other standard methods such as HSIC and DC.

Finally, the proposed method can be extended to consider independence between multidimensional variables and mutual independence involving three of more variables. However, these extensions require generalizing the definition of correntropy, which, so far, has only been defined on two random variables. Although a possible extension for pairs of multidimensional variables has already been proposed, further work is necessary in this context.

Acknowledgement

Sohan Seth, Jianwu Xu and Jose C. Principe would like to acknowledge support for this project from the National Science Foundation (NSF grant ECS-0601271).

Appendix A. Proof of Property 1

$$V(X,X)=\int \kappa(x-x)\,\mathrm{d}F_X(x)=\kappa(0)>0$$

and

$$\begin{aligned}U(X,X)&=\kappa(0)-\iint \kappa(x-y)\mathrm{d}F_X(x)\,\mathrm{d}F_X(y)\\&=\kappa(0)-\iint \mathrm{d}F_X(x)\,\mathrm{d}F_X(y)\int e^{-i\alpha(x-y)}\mu(\mathrm{d}\alpha)\\&=\kappa(0)-\int \mu(\mathrm{d}\alpha)\int e^{-i\alpha x}\mathrm{d}F_X(x)\int e^{i\alpha y}\mathrm{d}F_X(y)\\&=\kappa(0)-\int \mu(\mathrm{d}\alpha)\left|\int e^{i\alpha x}\mathrm{d}F_X(x)\right|^2\geq 0.\end{aligned}$$

If $U(X,X)=0$ then we have

$$\kappa(0)=\int \mu(\mathrm{d}\alpha)\left|\int e^{i\alpha x}\,\mathrm{d}F_X(x)\right|^2$$

which implies

$$\left|\int e^{i\alpha x}\,\mathrm{d}F_X(x)\right|=1$$

μ almost everywhere. Therefore if 0 is a limit point of the support of μ then we must have $X\equiv c$ where c is a constant, i.e. X is degenerate. This also implies that if X is not a degenerate random variable then we have $U(X,X)>0$.

Appendix B. Proof of Property 2

The symmetry follows from the symmetry of the kernel. Let $\alpha_1,\ldots,\alpha_n\in\mathbb{C}$, and $X_1,\ldots,X_n\in\mathcal{X}$, then

$$\sum_{i=1}^n\sum_{j=1}^n \alpha_i\overline{\alpha_j}V(X_i,X_j)=\mathbf{E}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\overline{\alpha_j}\kappa(X_i,X_j)\geq 0$$

and

$$\begin{aligned}&\sum_{j=1}^n\sum_{k=1}^n \alpha_j\overline{\alpha_k}U(X_j,X_k)\\&=\sum_{j=1}^n\sum_{k=1}^n \alpha_j\overline{\alpha_k}\iint\{\mathrm{d}F_{X_j,X_k}(x,y)-\mathrm{d}F_{X_j}(x)\,\mathrm{d}F_{X_k}(y)\}\int \mu(\mathrm{d}\alpha)e^{-i\alpha(x-y)}\\&=\int \mu(\mathrm{d}\alpha)\sum_{j=1}^n\sum_{k=1}^n \mathbf{E}[\alpha_j(e^{-i\alpha X_j}-\mathbf{E}[e^{-i\alpha X_k}])\overline{\alpha_k}(e^{i\alpha X_k}-\mathbf{E}[e^{i\alpha X_k}])]\\&=\int \mu(\mathrm{d}\alpha)\mathbf{E}\left|\sum_{k=1}^n \alpha_k(e^{-i\alpha X_k}-\mathbf{E}[e^{-i\alpha X_k}])\right|^2\geq 0.\end{aligned}$$

Appendix C. Proof of Property 3

$$\begin{aligned}|V(X,Y)|&=\left|\iint \kappa(x-y)\,\mathrm{d}F_{X,Y}(x,y)\right|\leq\int \mu(\mathrm{d}\alpha)\left|\iint e^{-i\alpha(x-y)}\,\mathrm{d}F_{X,Y}(x,y)\right|\\&\leq\int \mu(\mathrm{d}\alpha)=\kappa(0)=\sqrt{\kappa(0)}\sqrt{\kappa(0)}=\sqrt{V(X,X)}\sqrt{V(Y,Y)}\end{aligned}$$

and

$$\begin{aligned}|U(X,Y)|&=\left|\iint \kappa(x-y)\{\mathrm{d}F_{X,Y}(x,y)-\mathrm{d}F_X(x)\,\mathrm{d}F_Y(y)\}\right|\\&\leq\int \mu\mathrm{d}\alpha\left|\iint e^{-i\alpha(x-y)}\{\mathrm{d}F_{X,Y}(x,y)-\mathrm{d}F_X(x)\,\mathrm{d}F_Y(y)\}\right|\\&=\int \mu(\mathrm{d}\alpha)|\mathrm{cov}(e^{-i\alpha X},e^{-i\alpha Y})|\\&\leq\int \mu(\mathrm{d}\alpha)\sqrt{\mathrm{var}(e^{-i\alpha X})}\sqrt{\mathrm{var}(e^{-i\alpha Y})}\\&\leq\sqrt{\int \mathrm{var}(e^{-iX\alpha})\mu(\mathrm{d}\alpha)}\sqrt{\int \mathrm{var}(e^{-iY\alpha})\mu(\mathrm{d}\alpha)}\\&=\sqrt{U(X,X)}\sqrt{U(Y,Y)}.\end{aligned}$$

Appendix D. Proof of Property 4

1. ($\Rightarrow$) This is straight forward.
2. ($\Leftarrow$)

$$U(X,Y)=\int \mu(\mathrm{d}\alpha)\mathrm{cov}[e^{-i\alpha X},e^{-i\alpha Y}] \tag{8}$$

$$\leq\int \mu(\mathrm{d}\alpha)\sqrt{\mathrm{var}[e^{-i\alpha X}]}\sqrt{\mathrm{var}[e^{-i\alpha Y}]} \tag{9}$$

$$\leq\left(\int \mu(\mathrm{d}\alpha)\mathrm{var}[e^{-i\alpha X}]\right)^{1/2}\left(\int \mu(\mathrm{d}\alpha)\mathrm{var}[e^{-i\alpha Y}]\right)^{1/2} \tag{10}$$

$$=\sqrt{U(X,X)}\sqrt{U(Y,Y)}.$$

When $\eta(X,Y)=1$, the inequalities in the above equations turn into equality. Therefore, equality of Eqs. (9) and (10) implies

$$\mathrm{var}[e^{-i\alpha X}]=\beta^2\mathrm{var}[e^{-i\alpha Y}] \tag{11}$$

μ for almost all α, where β is a constant and since we assume the support of μ in the real line this holds for all α. Equality of Eqs. (8) and (9) implies

$$\mathrm{cov}(e^{-i\alpha X},e^{-i\alpha Y})=\sqrt{\mathrm{var}[e^{-i\alpha X}]}\sqrt{\mathrm{var}[e^{-i\alpha Y}]}$$

and therefore,

$$e^{-i\alpha X}-\mathbf{E}[e^{-i\alpha X}]=\gamma(\alpha)(e^{-i\alpha Y}-\mathbf{E}[e^{-i\alpha Y}]),\quad \gamma(\alpha)>0, \tag{12}$$

where $\gamma(\alpha)$ is constant and from Eqs. (11) and (12) we get, $\gamma(\alpha)=\beta$.

Therefore, we have

$$e^{-i\alpha X}-\mathbf{E}[e^{-i\alpha X}]=\beta[e^{-i\alpha Y}-\mathbf{E}[e^{-i\alpha Y}]] \tag{13}$$

for all α. Multiplying both sides of Eq. (13) by $f\in L^1$ and integrating with respect to α yields

$$\hat{f}(X)-\mathbf{E}[\hat{f}(X)]=\beta[\hat{f}(Y)-\mathbf{E}[\hat{f}(Y)]]$$

almost everywhere where

$$\hat{f}(x)=\int e^{-i\alpha x}f(\alpha)\,\mathrm{d}\alpha.$$

This set of functions is an algebra and we can thus approximate all continuous functions vanishing at infinity, uniformly. Thus we have

$$\varphi(X)-\mathbf{E}[\varphi(X)]=\beta[\varphi(Y)-\mathbf{E}[\varphi(Y)]]$$

for all continuous function functions $\varphi(\cdot)$ and hence, taking limits, all Borel functions.

In particular if A is Borel subset of $\mathbb{R}$

$$\mathbf{1}_A(X)-P(X\in A)=\beta[\mathbf{1}_A(Y)-P(Y\in A)] \tag{14}$$

almost everywhere. Let A be such that $0<P(X\in A)<1$. Now Eq. (14) holds with probability 1. Taking an ω such that $X(\omega)\in A$ and an ω' such that $X(\omega')\notin A$, we get

$$P(X\in A^c)=\beta[\mathbf{1}_A(Y(\omega))-P(Y\in A)], \tag{15}$$

$$-P(X\in A)=\beta[\mathbf{1}_A(Y(\omega'))-P(Y\in A)]. \tag{16}$$

As $0\le P(Y\in A)\le 1$ Eq. (15) forces $\mathbf{1}_A(Y(\omega))=1$ and Eq. (16) forces $\mathbf{1}_A(Y(\omega'))=0$. Thus,

$$P(X\in A^c)=\beta P(Y\in A^c), \tag{17}$$

$$P(X\in A)=\beta P(Y\in A). \tag{18}$$

Adding Eqs. (17) and (18), we get $\beta=1$. Thus we have

$$e^{-i\alpha X}-\mathbf{E}[e^{-i\alpha X}]=e^{-i\alpha Y}-\mathbf{E}[e^{-i\alpha Y}]$$

for all α. Taking derivatives on both sides twice and letting $\alpha=0$ we get

$$X-\mathbf{E}[X]=Y-\mathbf{E}[Y], \tag{19}$$

$$X^2-\mathbf{E}[X^2]=Y^2-\mathbf{E}[Y^2]. \tag{20}$$

Computing X^2 from Eq. (19) and putting it in Eq. (20), we get

$$2Y(\mathbf{E}[X]-\mathbf{E}[Y])=(\mathbf{E}[Y^2]-\mathbf{E}[X^2])-(\mathbf{E}[Y]-\mathbf{E}[X])^2$$

which is a first order equation of Y if $X\neq Y$. Solving this equation we get only one value of Y indicating that Y is a degenerate random variable. But this is a contradiction. Therefore $X=Y$.

Appendix E. Equivalence of several measures of multivariate dependence

In this section we show that many available measures of independence as described in Section 4.2 have similar estimators. Let us start with the characteristic function based measure proposed by [10]. This measure rely on the fact that two density functions are the same if their corresponding characteristic functions are the same. Consider the following function:

$$\gamma(\omega_1,\ldots,\omega_d)=\int \exp(-\omega_1x_1-\cdots-\omega_dx_d)(f_{X_1\ldots X_d}(x_1,\ldots,x_d)-f_{X_1}(x_1)\ldots f_{X_d}(x_d))\,\mathrm{d}x_1\ldots\mathrm{d}x_d.$$

Then γ is zero if and only if the random variables are independent. Using this function a measure of independence can be defined as follows:

$$\varphi=\int \|\gamma(\omega_1,\ldots,\omega_d)\|^2\theta(\omega_1,\ldots,\omega_d)\,\mathrm{d}\omega_1\ldots\mathrm{d}\omega_d,$$

where θ is a positive function centered at zero. If θ is taken to be a spherical multivariate Gaussian then an estimator of Γ is given by

$$\hat{\varphi}=n\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}G(x_k(i)-x_k(j))-\frac{2}{n^{d+1}}\sum_{i=1}^{n}\prod_{k=1}^{d}\sum_{j=1}^{n}G(x_k(i)-x_k(j))+\frac{1}{n^{2d}}\prod_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}G(x_k(i)-x_k(j))\right],$$

where $G(x)$ denotes the Gaussian function.

Next, consider the measure based on reproducing kernel Hilbert space (RKHS) [34]. It has been shown that if a kernel κ is characteristic then there exist a unique and one to one mapping between the space of probability density functions and the mean operator μ in the RKHS. Thus a test of independence can be designed by measuring the distance between two mean operator in the RKHS, i.e.

$$\zeta=\|\mu(f_{X_1\ldots X_d}(x_1,\ldots,x_d))-\mu(f_{X_1}(x_1)\ldots f_{X_d}(x_d))\|_{\mathcal{H}}.$$

The estimator of this quantity is given by

$$\hat{\zeta}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}\kappa(x_k(i)-x_k(j))-\frac{2}{n^{d+1}}\sum_{i=1}^{n}\prod_{k=1}^{d}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j))+\frac{1}{n^{2d}}\prod_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j)),$$

where $\kappa(x)$ is a reproducing kernel.

Next, we consider the test of independence based on L_2 distance between the joint and marginal density functions [29], i.e.

$$\eta=\int (f_{X_1\ldots X_d}(x_1,\ldots,x_d)-f_{X_1}(x_1)\ldots f_{X_d}(x_d))^2\,\mathrm{d}x_1\ldots\mathrm{d}x_d.$$

Using Parzen window estimate the estimation of this statistic is given by

$$\hat{\eta}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}\kappa(x_k(i)-x_k(j))-\frac{2}{n^{d+1}}\sum_{i=1}^{n}\prod_{k=1}^{d}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j))+\frac{1}{n^{2d}}\prod_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j)),$$

where $\kappa(x)$ is a symmetric density function.

Finally, we consider the measure of independence considered in [30]. This measure is based on the fact that a function $f(x)$ is zero everywhere if and only if $\int\kappa(x-a)f(x)\,\mathrm{d}x=0$ for all a where κ is a positive definite function. Using this a measure of independence is given by

$$\psi=\int\left[\int\prod_{i=1}^{d}\kappa(x_i-a_i)(f_{X_1\ldots X_d}(x_1,\ldots,x_d)-f_{X_1}(x_1)\ldots f_{X_d}(x_d))\,\mathrm{d}x_1\ldots\mathrm{d}x_d\right]^2\mathrm{d}a_1\ldots\mathrm{d}a_d.$$

The estimator of this quantity is given by

$$\hat{\psi}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}\kappa(x_k(i)-x_k(j))-\frac{2}{n^{d+1}}\sum_{i=1}^{n}\prod_{k=1}^{d}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j))+\frac{1}{n^{d2}}\prod_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa(x_k(i)-x_k(j)).$$

Thus we see that all these statistics reduces to the same estimator. These methods only differ in the choice of kernels they allow in the estimator. For L_2 the kernel should be a density function itself. However, a Gaussian or

a Cauchy kernel is preferred due to the property that the convolution of two Gaussian/Cauchy kernels is again a Gaussian/Cauchy kernel. For characteristic function based measure again a Gaussian/Cauchy kernel is preferred since they are Fourier transform of Gaussian/Laplacian kernel. For the RKHS based measure the kernels are chosen to be characteristic for example Gaussian/exponential and finally, for the last estimator again the kernels are again chosen to be Gaussian/Cauchy kernel.

Appendix F. Connection between the proposed method and QDM [14]

In QDM, the following fact is used to design a test of independence: given a positive definite kernel κ, the random variables X and Y are independent if and only if

$$\Gamma(a,b) = \mathbf{E}_{X,Y}\kappa(X-a)\kappa(Y-b) - \mathbf{E}_X\kappa(X-a)\mathbf{E}_Y\kappa(Y-b)$$

for all a and b. This fact is used to generate the following measure of independence:

$$\psi = \int \Gamma^2(a,b)\,\mathrm{d}a\,\mathrm{d}b.$$

Note that the expression ψ is very similar to the proposed method except the fact that here a separable kernel is used instead of a shift invariant kernel [23], i.e. roughly speaking, here the kernels are spherical and whereas in the proposed method the kernels are radial. Moreover, since in the final estimator the position of the kernel is integrated, ψ can be thought of as a Cramér–von-Mises type statistic whereas in our case we follow a Kolmogorov–Smirnov type approach by taking the supremum over all kernel positions. Note that in our case it is also possible to integrate over a and b and get a closed form solution.

References

[1] M. Loéve, Probability Theory, D. Van Nostrand Company, Inc.,, Princeton, NJ, 1960.
[2] A. Hyvörinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.
[3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[4] D.D. Mari, S. Kotz, Correlation and Dependence, Imperial College Press, London, 2001.
[5] J.R. Blum, J. Kiefer, M. Rosenblatt, Distribution free tests of independence based on the sample distribution function, Annals of Mathematical Statistics 32 (2) (1961) 485–498.
[6] B.F. Fernández, J.M. Gonázlez-Barrios, Multidimensional dependency measures, Journal of Multivariate Analysis 89 (2) (2004) 351–370.
[7] A. Feuerverger, A consistent test for bivariate dependence, International Statistical Review 61 (3) (1993) 419–433.
[8] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Annals of Mathematical Statistics 35 (6) (2007) 2769–2794.
[9] C.W. Granger, E. Maasoumi, J. Racine, A dependence metric for possibly nonlinear processes, Journal of Time Series Analysis 25 (5) (2004) 649–669.
[10] A. Kankainen, Consistent testing of total independence based on empirical characteristic function, Ph.D. Thesis, University of Jyväskylä, 1995.
[11] C. Granger, J.L. Lin, Using the mutual information coefficient to identify lags in nonlinear models, Journal of Time Series Analysis 15 (4) (1991) 371–384.
[12] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, A.J. Smola, A kernel statistical test of independence, in: Advances in Neural Information Processing Systems, vol. 19, 2007.
[13] C. Diks, V. Panchenko, Nonparametric tests for serial independence based on quadratic forms, CeNDEF Working Papers 05-13, Universiteit van Amsterdam, Center for Nonlinear Dynamics in Economics and Finance, 2005.
[14] S. Achard, D.T. Pham, C. Jutten, Quadratic dependence measure for nonlinear blind sources separation, in: Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, 2003, pp. 263–268.
[15] I. Santamaria, P. Pokharel, J.C. Príncipe, Generalized correlation function: definition, properties, and application to blind equalization, IEEE Transactions on Signal Processing 54 (6) (2006) 2187–2197.
[16] W. Liu, P. Pokharel, J. Príncipe, Correntropy: properties and applications in non-Gaussian signal processing, IEEE Transactions on Signal Processing 55 (11) (2007) 5286–5298.
[17] I. Park, J. Príncipe, Correntropy based Granger causality, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008, pp. 3605–3608.
[18] J. Xu, H. Bakardjian, A. Cichocki, J.C. Príncipe, A new nonlinear similarity measure for multichannel signals, Neural Networks 21 (2008) 222–231.
[19] A. Gunduz, J.C. Príncipe, Correntropy as a novel measure for nonlinearity tests, Signal Processing 89 (1) (2009) 14–23.
[20] P. Pokharel, R. Agrawal, J.C. Príncipe, Correntropy based matched filtering, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2005, pp. 148–155.
[21] K. Jeong, W. Liu, S. Han, E. Hasanbelliu, J.C. Príncipe, The correntropy mace filter, Pattern Recognition 42 (5) (2009) 871–885.
[22] R. Li, W. Liu, J.C. Príncipe, A unifying criterion for blind source separation based on correntropy, Signal Processing 87 (8) (2007) 1872–1881.
[23] M.G. Genton, Class of kernels for machine learning: a statistics perspective, Journal of Machine Learning Research 2 (2001) 299–312.
[24] S. Bochner, Hilbert distances and positive definite functions, Annals of Mathematics 42 (3) (1941) 647–656.
[25] S. Seth, J.C. Príncipe, Compressed signal reconstruction using the correntropy induced metric, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008, pp. 3845–3848.
[26] S. Seth, J.C. Príncipe, On speeding up computation in information theoretic learning, in: Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 2471–2475.
[27] A. Chen, Fast kernel density independent component analysis, Independent Component Analysis and Blind Signal Separation, Springer, Berlin, Heidelberg, 2006, pp. 24–31.
[28] X.T. Yuan, B.G. Hu, Robust feature extraction via information theoretic learning, in: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, , 2009, pp. 1193–1200.
[29] I.A. Ahmad, Q. Li, Testing independence by nonparametric kernel method, Statistics and Probability Letters 34 (2) (1997) 201–210.
[30] S. Achard, Asymptotic properties of a dimension-robust quadratic dependence measure, Comptes Rendus Mathematique 346 (3) (2008) 213–216.
[31] R.B. Nelsen, An Introduction to Copulas, Springer, 1999.
[32] E. Parzen, Statistical inference on time series by Hilbert space methods, Technical Report 23, Statistics Stanford, 1959.
[33] F.R. Bach, M.I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 (2002) 1–48.
[34] A.J. Smola, A. Gretton, L. Song, B. Schölkopf, A Hilbert space embedding for distributions, in: E. Takimoto (Ed.), Algorithmic Learning Theory, Lecture Notes on Computer Science.