A NEW NONPARAMETRIC MEASURE OF CONDITIONAL INDEPENDENCE

Sohan Seth, Il Park and José C. Príncipe

Computational NeuroEngineering Laboratory, University of Florida, Gainesville, USA
sohan@cnel.ufl.edu, memming@cnel.ufl.edu, principe@cnel.ufl.edu

ABSTRACT

In this paper we propose a new measure of conditional independence that is loosely based on measuring the L_2 distance between the conditional joint and the product of the conditional marginal density functions. However, we propose to smooth the arguments prior to measuring the distance and use kernel density estimation to derive the estimator. We show that under suitable conditions the proposed smoothing does not affect the conditional independence but using proper smoothing function helps in choosing the bandwidth parameter robustly. We discuss the computational issues and propose an approximation to evaluate the estimator efficiently. We apply the proposed measure in different experiments to show its validity.

Index Terms— Causality, conditional independence, dimension reduction, Gaussian integral, multivariate density estimation.

1. INTRODUCTION

Let $X = [X_1, X_2, \ldots, X_{d_X}]^\top$, $Y = [Y_1, Y_2, \ldots, Y_{d_Y}]^\top$ and $Z = [Z_1, Z_2, \ldots, Z_{d_Z}]^\top$ be three random vectors of dimensions d_X, d_Y and d_Z respectively. Then X and Y are said to be conditionally independent given Z if knowing Y does not provide any additional information about X when $Z = z$ is known. Mathematically this concept is expressed as

$$f_{X|YZ}(x|y,z) = f_{X|Z}(x|z) \quad \text{or, equivalently}$$
$$f_{XY|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z),\ \forall\, [x;y;z] \in \mathbb{R}^d \qquad (1)$$

where $d = d_X + d_Y + d_Z$, $f_{U,V}(u,v)$ denotes the joint probability density of U and V, $f_{U|V}(u|v)$ denotes the conditional probability density of U given a particular value of $V = v$ and $[\cdot\,;\cdot]$ denotes the operation of concatenating two column vectors. The conditional independence is symbolically expressed as $X \perp Y | Z$.

It can be easily seen that if $X \perp Y | Z$ then

$$f_{XYZ}(xyz) f_Z(z) = f_{XZ}(xz) f_{YZ}(yz),\ \forall\, [x;y;z] \in \mathbb{R}^d. \qquad (2)$$

When $f_Z(z) \neq 0$ (2) can be derived from (1) by replacing the conditional density functions $f_{U|Z}(u|z)$ by the ratio of the joint and the marginal densities as $f_{UZ}(uz)/f_Z(z)$. When $f_Z(z) = 0$ the equation is trivial as all the joint densities involving Z evaluated at $Z = z$ are also zero. Note that this expression does not involve any conditional densities but only joint and marginal densities.

A measure of conditional independence can be derived by measuring the distance between the left and right hand side of (1) or (2). However, note that the notion of distance may vary and we can find different measure of independence, for example see [1]. In this paper, we use (2) and measure the L_2 distance to derive a measure of conditional independence. The advantage of using (2) over (1) is that it does not require estimating any conditional densities to derive an estimator. However, we do not directly measure the L_2-distance but propose to smooth the arguments prior to measuring the distance and then use Parzen density estimate to estimate the distance. We show that under suitable conditions the smoothing does not affect the conditional independence but this simple approach allows us, first, to choose the bandwidth parameters in estimating the measure more robustly and, second, to approximate the final expression of the estimator to reduce computational complexity.

To improve the readability of the paper, we use the following notations throughout the paper. For simplicity we refer to random vectors by a normal capital letter e.g. U and their realizations by normal small letter e.g. u rather than bold letters e.g. $\mathbf{u}$ which are usual vector notation. We denote the i-th element of U i.e. the i-th random variable by U_i, the k-th realization of the random vector U by u_k which is a column vector and the k-th realization of the random variable U_i by u_{ik} i.e. we view the n realizations of U as a matrix of dimension $(d_U \times n)$ where d_U is the dimension of the random vector U. Note that u_i might also denote the i-th element of the vector u, a single realization of U. The proper meaning of u_i will depend on the context. As we are working with three random vectors in the paper, we device the following notation rule to denote the joint random vectors formed by them,

$$\bar{W} = [X;Y;Z],\ \bar{U} = [X;Z],\ \bar{V} = [Y;Z]$$

Unless otherwise stated we denote by $d_{(\cdot)}$ the dimension of a (random) vector. We denote a matrix by bold capital letter e.g. $\mathbf{U}$ and vectors, except realizations of random vectors, by normal bold letters e.g. $\mathbf{u}$. By $\operatorname{diag}\{\sigma_1, \ldots, \sigma_m\}$, we denote a diagonal matrix whose (i,i)-th element is σ_i and, finally, by $\mathbf{K}_{UU}^{\mathbf{\Sigma}_U}$, we denote the $(n \times n)$ Gram matrix formed by U whose (k,l)-th element is given by,

$$\mathbf{K}_{UU}^{\mathbf{\Sigma}_U}(k,l) = \prod_{i=1}^{d_U} \exp\left(-\frac{(u_{ik} - u_{il})^2}{2\sigma_{U_i}^2}\right)$$

where $\mathbf{\Sigma}_U = \operatorname{diag}\left\{\sigma_{U_1}, \ldots, \sigma_{U_{d_U}}\right\}$.

In the following section, we first describe the proposed measure and then derive an estimator. We then discuss the issue of approximating this estimator to reduce computational complexity. In the next section, we describe two experiments to demonstrate the validity of the proposed measure. In the final section we conclude the paper by summarizing the proposed work and briefly describing the future work.

2. MEASURE OF CONDITIONAL INDEPENDENCE

2.1. Description of the measure

If the joint densities are known or if they can be estimated with sufficient accuracy, for example by using kernel density estimate, then

This work is supported by NSF grant ECS 0601271

we can design a test of conditional independence based on testing whether (2) holds. For example, we can define the following measure,

$$\int |f_{XYZ}(x,y,z)f_Z(z) - f_{XZ}(xz)f_{YZ}(yz)|^2 \, dx\, dy\, dz.$$

It is easy to verify that this measure takes zero value under the null hypothesis $X \perp Y|Z$. However, working with the density functions directly poses a problem that we need to estimate appropriate bandwidth parameters for all the random variables. Therefore, in this paper, we consider a different approach; the purpose of which will be apparent soon.

We use the fact that a function $g : \mathbb{R}^d \to \mathbb{R}$ is zero almost everywhere on $\mathbb{R}^d$ if and only if

$$\int_{\mathbb{R}^d} \prod_{i=1}^{d} \theta_i(w_i - \mathbf{w}_i)\, g(w_1, w_2, \ldots, w_d)\, dw_1\, dw_2 \ldots dw_d = 0$$

for almost all $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]^\top \in \mathbb{R}^d$, provided the Fourier transforms of $\theta_i : \mathbb{R} \to \mathbb{R}$ do not vanish $\forall\, i \in \{1, 2, \ldots, d\}$. A simple proof with $\theta_i = \theta_j\ \forall i, j$ can be found in [2] and a general proof can be easily extended. In this paper, we choose θ_i's from a scale family i.e. we work with $\theta_i \equiv \theta_{h_i}$ where h_i is the corresponding scale parameter. In particular, we work with

$$\theta_h(w) = \frac{1}{\sqrt{2\pi}h} \exp\left(\frac{w^2}{2h^2}\right).$$

Using a Gaussian function serves a very particular and important purpose that will be addressed in the following subsections. For sake of clarity, we refer θ_i as the *smoothing function* and h_i as the corresponding *smoothing parameter*.

Let $g(\bar{w}) = f_{XYZ}(xyz)f_Z(z) - f_{XZ}(xz)f_{YZ}(yz)$ and

$$\Theta_{\bar{\mathbf{w}}}(\bar{w}) = \prod_{i=1}^{d_X} \theta_{h_{X_i}}(x_i - \mathbf{x}_i) \prod_{j=1}^{d_Y} \theta_{h_{Y_j}}(y_j - \mathbf{y}_j) \prod_{k=1}^{d_Z} \theta_{h_{Z_k}}(z_k - \mathbf{z}_k)$$

where $\bar{\mathbf{w}} = [\mathbf{x}; \mathbf{y}; \mathbf{z}] \in \mathbb{R}^{d_{\bar{\mathbf{w}}}}$ and $d_{\bar{\mathbf{w}}} = d$. Note that we use the same rule to denote the concatenation of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. Define,

$$H(\bar{\mathbf{w}}) = \int_{\mathbb{R}^{d_{\bar{w}}}} \Theta_{\bar{\mathbf{w}}}(\bar{w}) g(\bar{w})\, d\bar{w}. \tag{3}$$

Then we have a new condition of conditional independence that $X \perp Y|Z$ if and only if $H(\bar{\mathbf{w}}) = 0\, \forall\, \bar{\mathbf{w}} \in \mathbb{R}^{d_{\bar{\mathbf{w}}}}$. The purpose of working with (3) and not (2) is that it allows certain robustness in choosing the bandwidth parameters. We will see this when we derive the estimator of $H(\bar{\mathbf{w}})$. At this point we propose the following measure of conditional independence,

$$\rho = \int_{\mathbb{R}^{d_{\bar{\mathbf{w}}}}} H^2(\bar{\mathbf{w}})\, d\bar{\mathbf{w}}.$$

We call ρ the *quadratic* measure. Thus, $\rho = 0 \Leftrightarrow X \perp Y|Z$.

2.2. Estimation of the measure

In order to estimate ρ, we replace the density functions by their Parzen estimates i.e. in particular we put

$$\hat{f}_{\bar{W}}(\bar{w}) = \frac{1}{n} \prod_{i=1}^{d_X} \kappa_{\sigma_{X_i}}(x_i - x_{is}) \prod_{j=1}^{d_Y} \kappa_{\sigma_{Y_j}}(y_j - y_{js}) \prod_{k=1}^{d_Z} \kappa_{\sigma_{Z_k}}(z_k - z_{ks})$$

where $\kappa_\sigma(u)$ denotes the Parzen kernel with bandwidth σ and n is the number of samples [3]. The marginal densities can be easily obtained by removing appropriate terms from the joint density. In this paper we choose the kernel to be Gaussian.

As we have mentioned before, choosing both the smoothing function θ_h, in the previous subsection, and κ_σ, here, to be Gaussian is very important. It allows us to evaluate the integrals in $H(\bar{\mathbf{w}})$ and ρ in closed form and in terms of Gaussian functions themselves. That in turn allows us to rewrite the expression in terms of Gram matrices of the random vectors and to use the properties of Gram matrices to simplify calculations. Such property of Gaussian function has been extensively used by the Information Theoretic Learning research community [4].

Replacing g by its Parzen density estimate and using the properties of Gaussian functions, we can evaluate $\hat{H}(\bar{\mathbf{w}})$ and $\hat{\rho}$ in closed form. A tedious but straightforward calculation gives $\hat{\rho} =$

$$\frac{1}{D} \sum_{s,t,v,w=1}^{n} \Big[\mathbf{K}_{XX}^{\mathbf{\Sigma}_X}(s,v)\left(\mathbf{K}_{YY}^{\mathbf{\Sigma}_Y}(s,v) + \mathbf{K}_{YY}^{\mathbf{\Sigma}_Y}(t,w) - 2\,\mathbf{K}_{YY}^{\mathbf{\Sigma}_Y}(s,w)\right) \mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}(s,w)\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}(s,w)\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}(t,v)\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}(t,w)\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}}(s,t)\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}}(v,w)\Big]$$

where

$$\mathbf{\Sigma}_X = \operatorname{diag}\left\{\xi_{X_i} = \sqrt{2}\sqrt{\sigma_{X_i}^2 + h_{X_i}^2} \ :\ i \in \{1, \ldots, d_X\}\right\},$$

$$\mathbf{\Sigma}_Y = \operatorname{diag}\left\{\xi_{Y_j} = \sqrt{2}\sqrt{\sigma_{Y_j}^2 + h_{Y_j}^2} \ :\ j \in \{1, \ldots, d_Y\}\right\},$$

$$\mathbf{\Sigma}_{\ddot{Z}} = \operatorname{diag}\left\{\varsigma_{Z_k} = 2\sqrt{\sigma_{Z_k}^2 + 2h_{Z_k}^2} \ :\ k \in \{1, \ldots, d_Z\}\right\}$$

$$\mathbf{\Sigma}_{\dot{Z}} = \operatorname{diag}\left\{2\,\sigma_{Z_k}\sqrt{\frac{\sigma_{Z_k}^2 + 2h_{Z_k}^2}{\sigma_{Z_k}^2 + 4h_{Z_k}^2}} : k \in \{1, \ldots, d_Z\}\right\} \text{ and}$$

$$D = n^2\pi^{\frac{d_X + d_Y + 3d_Z}{2}} 2^{\frac{2d_X + 2d_Y + 3d_Z}{2}} \prod_{i=1}^{d_X} \xi_{X_i} \prod_{j=1}^{d_Y} \xi_{Y_j} \prod_{k=1}^{d_Z} \left(\varsigma_{Z_k}\sigma_{Z_k}^2\right).$$

Note that in the final bandwidth matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$, the individual kernel bandwidths (i.e. σ_{X_i} or σ_{Y_j}) are not important. However, it is the combinations of the kernel bandwidths and the smoothing parameters that play the key role. Since, we can vary the smoothing parameter (i.e. h_{X_i} and h_{Y_j}) for the corresponding kernel bandwidths (i.e. σ_{X_i} and σ_{Y_j} respectively), we can choose a single value (say σ) for all the diagonal elements of $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ provided $\sigma > \sigma_{X_i}, \sigma_{Y_j}\ \forall i, j$.

However, we do not enjoy the same privilege for $\mathbf{\Sigma}_{\dot{Z}}$ and $\mathbf{\Sigma}_{\ddot{Z}}$ since in this case we need to choose two different bandwidth parameters (i.e. one for $\mathbf{\Sigma}_{\dot{Z}}$ and the other for $\mathbf{\Sigma}_{\ddot{Z}}$) for each random variable Z_k but we only have one smoothing parameter (i.e. h_{Z_k}) that can be varied. Therefore, for the conditioning variable we need to pay attention to the selection of a proper kernel bandwidth (i.e. σ_{Z_k}). However, we will later see that the smoothing parameter for the conditioning variable plays a crucial role in approximating the estimator to reduce computational complexity.

2.3. Approximation of the estimator

The estimator $\hat{\rho}$ described in the previous subsection contains four summations. Therefore, to evaluate this expression directly requires $O(n^4)$ computation which is prohibitive in any practical application. To resolve this issue we make use of the smoothing parameter of the conditioning variables. Note that if the diagonal entries ς_{Z_k} of $\mathbf{\Sigma}_{\ddot{Z}}$ is sufficiently large then the elements of the matrix $\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}$ approach 1. Assume that $\operatorname{var} Z_k = 1$ and $\varsigma_{Z_k} > \varsigma\, \forall k \in \{1, \ldots, d_Z\}$. Now,

let Z_i and Z_j be two iid realizations of Z, then

$$1-\mathbb{E}\left[\exp\left\{-\sum_{k=1}^{d_Z}\frac{(Z_{ki}-Z_{kj})^2}{2\varsigma_{Z_k}^2}\right\}\right]<\mathbb{E}\left[\sum_{k=1}^{d_Z}\frac{(Z_{ki}-Z_{kj})^2}{2\varsigma_{Z_k}^2}\right]<\frac{d_Z}{\varsigma^2}.$$

Therefore if, say, $\varsigma=10\sqrt{d_Z}$ then the average value of the elements of $K_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}$ is at least greater than 0.99.

Exploiting this fact we approximate the term $\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}(t,v)$ in the final expression of $\hat{\rho}$ by 1. The resulting expression can then be expressed in the following way,

$$\begin{aligned}\hat{\rho}\approx\frac{1}{D}\Big[&\mathbf{1}^\top\left[(\mathbf{K}_{\bar{W}\bar{W}}^{\mathbf{\Sigma}_{\bar{W}}}\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}})\circ\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}\circ(\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}}\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}})\right]\mathbf{1}\\&+\mathbf{1}^\top\left[(\mathbf{K}_{\bar{U}\bar{U}}^{\mathbf{\Sigma}_{\bar{U}}}\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}})\circ\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}}\circ(\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}}\mathbf{K}_{\bar{V}\bar{V}}^{\mathbf{\Sigma}_{\bar{V}}})\right]\mathbf{1}\\&-2\mathbf{1}^\top\left[(\mathbf{K}_{\bar{U}\bar{U}}^{\mathbf{\Sigma}_{\bar{U}}}\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}})\circ\mathbf{K}_{\bar{V}\bar{V}}^{\mathbf{\Sigma}_{\bar{V}}}\circ(\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\dot{Z}}}\mathbf{K}_{ZZ}^{\mathbf{\Sigma}_{\ddot{Z}}})\right]\mathbf{1}\Big]\end{aligned}$$

where $\circ$ denotes the entrywise matrix multiplication operation and $\mathbf{1}$ is a $(n\times 1)$-dimensional vector of 1's.

A direct evaluation of the approximated $\hat{\rho}$ requires $O(n^3)$ operations. However, to evaluate $\hat{\rho}$ more efficiently we use the incomplete Cholesky decomposition where a $(n\times n)$-dimensional positive definite matrix $\mathbf{K}$ is represented by a $(n\times d)$-dimensional lower triangular matrix $\tilde{\mathbf{G}}$ as $\mathbf{K}\approx\tilde{\mathbf{G}}^\top\tilde{\mathbf{G}}$ where $d\leq n$ [5]. The computation complexity of this method is $O(nd^2)$. Notice that the expression of $\hat{\rho}$ consists of three terms of the form $\mathbf{1}^\top[(\mathbf{K}_1\mathbf{K}_2)\circ\mathbf{K}_3\circ(\mathbf{K}_4\mathbf{K}_5)]\mathbf{1}$ where all $\mathbf{K}$'s are $(n\times n)$-dimensional matrices. First, we compute the incomplete Cholesky decompositions ($\tilde{\mathbf{G}}_i$) of the corresponding Gram matrices ($\mathbf{K}_i$) where $\tilde{\mathbf{G}}_i$ is a $(n\times d_i)$-dimensional matrix. Then we compute $\tilde{\mathbf{G}}_{12}=\tilde{\mathbf{G}}_1^\top\tilde{\mathbf{G}}_2\tilde{\mathbf{G}}_2^\top$ and $\tilde{\mathbf{G}}_{45}=\tilde{\mathbf{G}}_4^\top\tilde{\mathbf{G}}_5\tilde{\mathbf{G}}_5^\top$ such that $\mathbf{K}_1\mathbf{K}_2=\tilde{\mathbf{G}}_1\tilde{\mathbf{G}}_{12}^\top$ and $\mathbf{K}_4\mathbf{K}_4=\tilde{\mathbf{G}}_4\tilde{\mathbf{G}}_{45}^\top$. These computations require $O(2nd_1d_2)$ and $O(2nd_4d_5)$ operations respectively. Finally, we rewrite

$$\mathbf{1}^\top\left[(\tilde{\mathbf{G}}_1\tilde{\mathbf{G}}_{12}^\top)\circ(\tilde{\mathbf{G}}_3\tilde{\mathbf{G}}_3^\top)\circ(\tilde{\mathbf{G}}_4\tilde{\mathbf{G}}_{45}^\top)\right]\mathbf{1}=\sum_{k=1}^{d_1}\sum_{l=1}^{d_3}\sum_{m=1}^{d_4}\left(\sum_{i=1}^{n}\tilde{\mathbf{G}}_1(i,k)\tilde{\mathbf{G}}_3(i,l)\tilde{\mathbf{G}}_4(i,m)\sum_{j=1}^{n}\tilde{\mathbf{G}}_{12}(k,j)\tilde{\mathbf{G}}_3(l,j)\tilde{\mathbf{G}}_{45}(m,j)\right).$$

Computing this term requires $O(4nd_1d_3d_4)$ operations. Therefore as long as we have $d_1,d_2,d_3,d_4,d_5\ll n$, the complexity of evaluating the approximation of $\hat{\rho}$ is almost linear in n.

2.4. Selection of bandwidth parameters

Selecting appropriate bandwidths for multidimensional density estimation is a nontrivial problem. However, assuming that the sample variance of the observed data is 1, it can be shown that the bandwidth parameter that minimizes the mean square error between the actual and the estimated probability density function depends on the number of samples in the following way [6], $\sigma\propto n^{-1/(d+4)}$ where d is the dimension of the multivariate density function. Following this guideline we choose

$$\sigma_{X_i}=\sigma_{Y_j}=\sigma_{Z_i}=cn^{-1/(d+4)}$$

$\forall i\in\{1,\ldots,d_X\},j\in\{1,\ldots,d_Y\},k\in\{1,\ldots,d_Z\}$ where c is a constant. Finally, following the argument in the previous section we select

$$h_{X_i}=h_{Y_j}=h_{Z_i}=5\sqrt{d_Z}$$

$\forall i\in\{1,\ldots,d_X\},j\in\{1,\ldots,d_Y\},k\in\{1,\ldots,d_Z\}$. Note that when $h_{Z_k}\gg\sigma_{Z_k}$ then $\varsigma_{Z_k}\approx 2\sqrt{2}h_{Z_k}$. Therefore the choice of the smoothing parameter is justified.

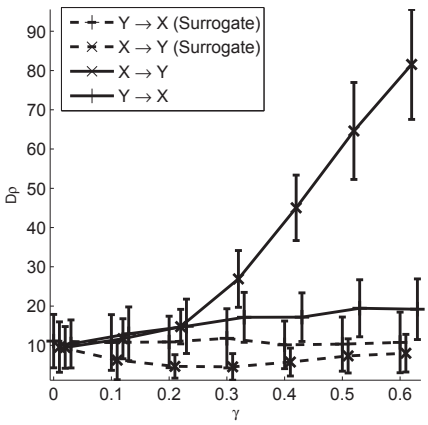

Fig. 1. The figure shows the variation of the quadratic measure $\hat{\rho}$ with the coupling strength γ.

3. SIMULATIONS

In the experiments described below, we always normalize our variables to have zero mean and unity variance and use the bandwidth parameters as described in the previous section. We choose $c=0.5$ for both the experiments.

3.1. Causality

Given two time series $\{U_t\}$ and $\{V_t\}$, $\{U_t\}$ is said to cause $\{V_t\}$ if the present value of $\{V_t\}$ depends on the past value(s) of $\{U_t\}$ i.e. $V_t=h(V_t^-,U_t^-)$ where V_t denotes the current value, V_t^- (or U_t^-) denotes the collection of past values and h is a function; linear or nonlinear. We denote causality by $\{U_t\}\rightarrow\{V_t\}$. Using the concept of conditional independence it is trivial to see that $\{U_t\}\rightarrow\{V_t\}\Leftrightarrow V_t\not\perp U_t^-|V_t^-$.

Let us consider the following time series,

$$\begin{aligned}&X_1(t+1)=1.4-X_1(t)^2+0.3X_2(t),\ X_2(t+1)=X_1(t)\\&Y_1(t+1)=1.4-\{\gamma X_1(t)Y_1(t)+(1-\gamma)Y_2(t)^2\}+0.1Y_2(t)\\&Y_2(t+1)=Y_1(t),\ X_3,X_4,Y_3,Y_4\sim\mathcal{N}(0,0.5^2)\end{aligned}$$

where $0\leq\gamma\leq 0.6$ is the coupling strength. It is obvious that $\{X_t\}\rightarrow\{Y_t\}$ as the current value of $\{Y_t\}$ depends on the past value of $\{X_t\}$ but $\{Y_t\}\not\rightarrow\{X_t\}$ as $\{X_t\}$ evolves by itself. Therefore, $X_t\perp Y_{t-1}\,|\,X_{t-1}$ but $Y_t\not\perp X_{t-1}\,|\,Y_{t-1}$. Also, note that the causal dependence between $\{X_t\}$ and $\{Y_t\}$ increases with the increase in the coupling strength.

We consider this problem to compare our measure with that of [7]. This problem is slightly challenging for the proposed measure as, here, we have $d_X=d_Y=d_Z=4$, i.e. the dimensionality of the problem is 12 and the higher dimensionality of the problem might cause trouble in estimation of multivariate density function. Therefore, to deal with this issue, we work with 500 samples compared to 100 in [7]. Fig. 1 shows the results of the experiment. We clearly see that the figure supports the facts that $\{X_t\}\rightarrow\{Y_t\}$ but $\{X_t\}\not\rightarrow\{Y_t\}$ and also that the causal dependence increases with the increase in the coupling strength.

In order to compare our result to that of [7] we perform a permutation test to decide an empirical threshold of rejection and compute the rejection rate based on that. In Table 1 and 2 we record the result of the hypothesis test. The size of the test is 0.05. The results

Table 1. Performance of I^{NOCCO}, $HSIC$ and Quadratic measure in rejecting the null hypothesis $\{X_t\}$ is *not* causing $\{Y_t\}$.

γ	0	0.1	0.2	0.3	0.4	0.5	0.6
I^{NOCCO}	97	96	93	85	81	68	75
$HSIC$	94	94	92	81	60	73	66
ρ	98	92	89	87	71	79	76

Table 2. Performance of I^{NOCCO}, $HSIC$ and Quadratic measure in rejecting the null hypothesis $\{Y_t\}$ is *not* causing $\{X_t\}$.

γ	0	0.1	0.2	0.3	0.4	0.5	0.6
I^{NOCCO}	96	0	0	0	0	0	0
$HSIC$	93	95	85	56	1	1	1
ρ	99	74	12	0	0	0	0

of Normalized Cross-Covariance Operator (I^{NOCCO}) and Hilbert-Schmidt Independence Criterion ($HSIC$) have been excerpted from [7]. We see that, although, our measure fails to reach the performance level of I^{NOCCO}, it performs better than $HSIC$ in terms of rejecting the null hypothesis, $\{Y_t\}$ is *not* causing $\{X_t\}$.

3.2. Dimensionality reduction

In a regression setting, let U be a set of m features and Y be the target. Then a l-dimensional *effective* subspace of the original m-dimensional feature space ($l < m$) consists of l linear combinations of the m original features that carry most of the information (if not all) about the target. Mathematically, we aim at finding an orthonormal transformation $\mathbf{\Pi} = [\mathbf{\Pi}_S, \mathbf{\Pi}_S^{\perp}]$ such that $Y \perp \mathbf{\Pi}_S^{\perp\top} X \,|\, \mathbf{\Pi}_S^{\top} X$ where $\mathbf{\Pi}_S$ is a $(m \times l)$-dimensional matrix that defines the effective subspace and $\mathbf{\Pi}_S^{\perp}$ denotes the corresponding null space. We use the quadratic measure as the cost function which attains its minimum at $Y \perp \mathbf{\Pi}_S^{\perp\top} X \,|\, \mathbf{\Pi}_S^{\top} X$.

Let us consider the following example, $Y = 2e^{-X_1^2/2} + \xi$ where $X = [X_1, X_2]$ is distributed as $0.5\mathcal{N}(\mathbf{m}, \mathbf{\Sigma^2}) + 0.5\mathcal{N}(-\mathbf{m}, \mathbf{\Sigma^2})$ with $\mathbf{m} = [0.5, 0.5]$ and $\mathbf{\Sigma} = \left[[1; -\sqrt{2}], [-\sqrt{2}; 1]\right]$ and $\xi \sim \mathcal{N}(0, 0.1^2)$ is additive white measurement noise. It can be easily seen that in this particular setting only X_1 is important and therefore the effective subspace is given by $\mathbf{\Pi}_S = [1, 0]^{\top}$.

We choose this example because the nonlinearity of the problem makes it harder to find the effective subspace [8]. The problem involves only 3 random variables. Therefore, we work with 100 samples. Also note that we need to find only a single Givens rotation that generate the orthonormal matrix $\mathbf{\Pi}$. Fig 2 shows the variation of the cost function along the Givens rotation. We clearly see that there is a global minimum near the neighborhood of 0. Exploiting this fact, we use a simple Golden search technique in the neighborhood of 0 to find the minimum. Table 3 records the performance of our method along with other methods such as Canonical correlation Analysis (CCA) and Kernel Dimensionality Reduction (KDR). The results of the other methods have been excerpted from [8]. We see that the proposed measure has been able to extract the effective subspace successfully.

4. SUMMARY

In this paper we propose a new measure of conditional independence that is based on measuring the difference between the conditional joint density and the product of conditional marginal densities. We use kernel density estimation to derive the estimator of the proposed measure. However, we propose to work with smoothed density functions to allow robust selection of the bandwidth parameters. We also propose an approximation for the estimator that can be evaluated efficiently. We apply the measure on two synthetic problems to demonstrate its potential and validity.

The estimator of the proposed measure relies on the kernel density estimate and kernel density estimation often becomes unreliable in multidimension due to unavailability of sufficient samples. Although, we have seen in our experiments that the proposed measure works effectively with finite number of samples, a theoretical study is nonetheless necessary. Also, we work with an approximation of the estimator rather than the actual estimator itself. Although this approximation gives us reasonable result in our experiments, further study of the approximation is still important.

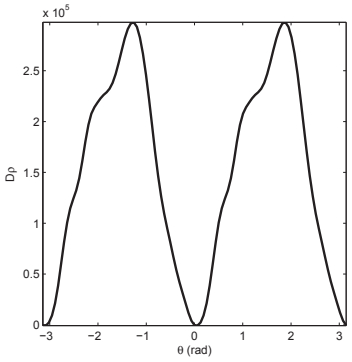

Fig. 2. The figure shows the variation of the $\hat{\rho}$ as a cost function along the Givens rotation θ.

Table 3. Performance of SIR, pHd, CCA, PLS, KDR and the quadratic measure in finding the effective subspace for regression.

	SIR	pHd	CCA	PLS	KDR	ρ
abs. err. (rad)	1.51	0.99	0.18	0.45	0.005	0.047

5. REFERENCES

[1] L. Su and H. White, "A nonparametric Hellinger metric test for conditional independence," *Econometric Theory*, vol. 24, pp. 829–864, 2008.

[2] S. Achard, D. T. Pham, and 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.

[3] E. Parzen, "On the estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[4] J. C. Príncipe, D. Xu, and J. W. Fisher III, *Unsupervised adaptive filtering*. John wiley & sons, inc., 2000, vol. 1, ch. 7, pp. 265–320.

[5] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2001.

[6] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986.

[7] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 489–496.

[8] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *J. Mach. Learn. Res.*, vol. 5, pp. 73–99, 2004.