Copula-based Kernel Dependency Measures

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Carnegie Mellon University, School of Computer Science, 5000 Forbes ave, Pittsburgh, PA 15213 USA

Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
University of Cambridge, Department of Engineering, Trumpington Street, Cambridge CB2 1PZ, UK

Jeff Schneider SCHNEIDE@CS.CMU.EDU
Carnegie Mellon University, School of Computer Science, 5000 Forbes ave, Pittsburgh, PA 15213 USA

Abstract

The paper presents a new copula based method for measuring dependence between random variables. Our approach extends the Maximum Mean Discrepancy to the copula of the joint distribution. We prove that this approach has several advantageous properties. Similarly to Shannon mutual information, the proposed dependence measure is invariant to any strictly increasing transformation of the marginal variables. This is important in many applications, for example in feature selection. The estimator is consistent, robust to outliers, and uses rank statistics only. We derive upper bounds on the convergence rate and propose independence tests too. We illustrate the theoretical contributions through a series of experiments in feature selection and low-dimensional embedding of distributions.

1. Introduction

Measuring dependence between random variables is an important problem in statistics, information theory, and machine learning with a wide range of applications in science and engineering. The most well-known dependence measure is the Shannon mutual information, which has found numerous applications recently. Although this is the most popular dependence measure, it is only one of the many other existing ones. In particular, it is a special case of the Rényi-α (Rényi, 1961) and Tsallis-α mutual information (Tsallis, 1988). Other interesting dependence measures include the maximal correlation coefficient (Rényi, 1959), kernel mutual information (Gretton et al., 2003), the generalized variance and kernel canonical correlation analysis (Bach, 2002), the Hilbert-Schmidt independence criterion (Gretton et al., 2005), the Schweizer-Wolff measure (Schweizer & Wolff, 1981), and the distance based correlation (Székely et al., 2007).

There is a tremendous list of dependence applications. They have been used, for example, in causality detection, feature selection, active learning, structure learning, boosting, image registration, independent component and subspace analysis. For more applications and references, please see the supplementary material.

One reason why so many dependence measures have been defined in the literature is that the problem is challenging and researchers and practitioners are not satisfied with the available measures and estimators (Fernandes & Gloor, 2010). As Schweizer & Wolff (1981) formalized in their dependence axioms, a good dependence measure I has to have several properties. The most important ones are as follows. (i) Dependence $I(\mathbf{X})$ is defined for $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ d-dimensional random variables. (ii) $I(X_1, \ldots, X_d)$ is invariant to permutation. (iii) $0 \leq I(\mathbf{X})$, and $I(\mathbf{X}) = 0$ iff $(X_1, \ldots, X_d)$ are independent variables. (iv) $I(X_1, \ldots, X_d)$ is invariant to strictly increasing transformation of X_i variables. For more discussion on these axioms, see the Appendix. Among the above mentioned dependence measures, only the Rényi, Tsallis information, and the Schweizer-Wolff measure is invariant to strictly increasing transformations.

In addition to these constraints on the dependence measure, we also want an efficient estimator that is consistent, robust to outliers, has fast convergence rate, and can be used in high-dimensions too. De-

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

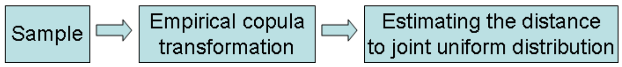


Figure 1. Illustration of the proposed dependence measure. Using empirical copula transformation, first we transform the data to have uniform marginals, then measure the distance to the joint uniform distribution with reproducing kernel based divergence estimators.

pendence estimation is very challenging, especially in nonparametric situations when we cannot assume that the observations have an underlying density function belonging to some parametric family. Many of the above mentioned dependence measures can be defined as some functionals of the density, thus an obvious way for their estimation would be to estimate the densities first. The density function, however, is a nuisance parameter in our case, and its estimation—especially in higher dimensions—is known to be very difficult.

Due to these difficulties, all the existing dependence estimators have their own shortcomings. For example, the bound on the convergence rate of the Rényi and Tsallis information estimator (Pál et al., 2010) suffers from the curse of dimensionality. The available reproducing kernel based dependence measures are not invariant to strictly increasing transformation of the X_i marginal random variables. The estimator of Székely et al. (2007) is not robust; one single large enough outlier can arbitrarily ruin the estimator.

The main contributions of the paper are as follows. (i) We introduce a new dependency measure I that satisfies the above listed axioms. (ii) We prove that I can be efficiently estimated, and the calculation of the estimator is simple. The estimator is consistent, robust to outliers, and uses rank statistics only. (iii) We also provide an upper bound on the rate of convergence and derive a test of independence. This bound shows that the estimator can be efficiently used in large dimensions too.

Our main idea is to combine empirical copula transformations with reproducing kernel based divergence estimators. We will show that the empirical copula transformation only slightly affects the convergence rate, but the resulting dependence estimator possesses all the above mentioned required properties. The proposed method is illustrated in Figure 1.

One might wonder why it is important for a dependence measure to be invariant to strictly increasing transformations of the marginal variables. One reason for this is that in many scenarios we need to compare the estimated dependencies. This is the case for example in feature selection and low-dimensional embedding of random variables. In these problems we can think of dependence as a "distance" between random variables in the sense that when the dependence is large, then the random variables are "close" to each other, and when the dependence is small, then the variables are far. However, if certain variables are measured on different scales, then this distance can be much different from the distance using other scales. As a result, it might happen that different features would be selected by the feature selection algorithm if we measured a quantity e.g. in grams, kilograms, pounds, or if we used log-scale. This is an odd situation that can be avoided with dependence measures that are invariant to strictly increasing transformations of the marginal variables. As an application, we will show how the proposed dependence measure can be used for feature selection and low-dimensional embedding of distributions.

The proofs can be found in the supplementary material. There we also discuss the robustness properties of the estimators and show how to use them in independence tests.

Notation: In the rest of the paper $X \sim P$ will denote that the random variable X has distribution P. $\mathbb{E}(X)$ and $\sigma(X)$ stand for the expectation and standard deviation of X, respectively. For a random variable $X \in \mathbb{R}$, $\Xi[X]$ denotes the standardized variable, that is, $\Xi[X] \doteq (X - \mathbb{E}[X])/\sigma(X)$, which has zero mean and unit variance. $U[a, b]$ stands for the uniform distribution in the interval $[a, b]$. $X_{1:m}$ is shorthand notation for the set of random variables $\{X_1, \ldots, X_m\}$. The cardinality of a set S is denoted by $|S|$.

2. Maximum Mean Discrepancy

In this section we review some important properties of the Maximum Mean Discrepancy (MMD), which is a quantity used to measure the distance between distributions (Borgwardt et al., 2006; Fortet & Mourier, 1953). An appealing property of this quantity is that it can be efficiently estimated from independent and identically distributed (i.i.d.) samples.

Definition 1. *Let $\mathcal{F}$ be a class of functions, P, Q be probability distributions. The MMD between P and Q on the function class $\mathcal{F}$ is defined as follows,*

$$\mathcal{M}[\mathcal{F}, P, Q] \doteq \sup_{f \in \mathcal{F}} (\mathbb{E}_{\mathbf{X} \sim P}[f(\mathbf{X})] - \mathbb{E}_{\mathbf{Y} \sim Q}[f(\mathbf{Y})]).$$

Let $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$ be a reproducing kernel Hilbert Space (RKHS) with feature map $\phi(x) \in \mathcal{H}$ ($x \in \mathcal{X}$), and kernel $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. It is well known that $\phi(x) = k(\cdot, x)$, and $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$, which is called the reproducing property of the RKHS. Later we will also need the definition of universal kernels.

Definition 2 (Universal kernel)**.** *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is universal whenever the associated RKHS $\mathcal{H}$ is dense in $C(\mathcal{X})$, the space of bounded continuous functions over $\mathcal{X}$, with respect to the L_∞ norm.*

Steinwart (2001) has shown that the Gaussian and Laplace kernels are universal. Let $\mu[P] \doteq \mathbb{E}_{\mathbf{X}\sim P}[\phi(\mathbf{X})] = \mathbb{E}_{\mathbf{X}\sim P}[k(\cdot, \mathbf{X})]$. A sufficient condition for this quantity to exist is $\mathbb{E}_{\mathbf{X}\sim P, \mathbf{X}'\sim P} k(\mathbf{X}, \mathbf{X}') < \infty$, where $\mathbf{X}$ and $\mathbf{X}'$ are independent variables having distribution P.

For general $\mathcal{F}$ function sets, $\mathcal{M}[\mathcal{F}, P, Q]$ can be difficult to calculate and is not even symmetric in P and Q. Nonetheless, when $\mathcal{F}$ is a unit ball of RKHS $\mathcal{H}$, then for all $f \in \mathcal{F}$ we also have that $-f \in \mathcal{F}$, which implies that $\mathcal{M}[\mathcal{F}, P, Q] = \mathcal{M}[\mathcal{F}, Q, P]$. Furthermore, in this case $\mathcal{M}^2[\mathcal{F}, P, Q]$ has a simple form that makes efficient estimations possible. This is stated formally in the following lemma (Borgwardt et al., 2006).

Lemma 3. *When $\mathcal{F}$ is a unit ball of RKHS $\mathcal{H}$ and $\mu[P] < \infty$, $\mu[Q] < \infty$, then*

$$\mathcal{M}^2[\mathcal{F}, P, Q] = \|\mu[P] - \mu[Q]\|^2_{\mathcal{H}} = \mathbb{E}_{\mathbf{X}, \mathbf{X}'\sim P}[k(\mathbf{X}, \mathbf{X}')] - 2\mathbb{E}_{\mathbf{X}\sim P, \mathbf{Y}\sim Q}[k(\mathbf{X}, \mathbf{Y})] + \mathbb{E}_{\mathbf{Y}, \mathbf{Y}'\sim Q}[k(\mathbf{Y}, \mathbf{Y}')],$$

where $\mathbf{X}$ and $\mathbf{X}'$ have distribution P, $\mathbf{Y}$ and $\mathbf{Y}'$ have distribution Q, and these random variables are all independent from each other.

In the remainder of the paper we will always assume that $\mathcal{F}$ is a unit ball of RKHS $\mathcal{H}$. Let $\mathbf{X}_{1:m} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ be an independent and identically distributed (i.i.d.) sample drawn from distribution P, and similarly let $\mathbf{Y}_{1:n} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ be an i.i.d. sample with distribution Q.

A biased (but asymptotically unbiased) estimator for $\mathcal{M}[\mathcal{F}, P, Q]$ can be easily given using the law of large numbers:

$$\mathcal{M}_b[\mathcal{F}, \mathbf{X}_{1:m}, \mathbf{Y}_{1:n}] \doteq \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(\mathbf{X}_i, \mathbf{X}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{Y}_i, \mathbf{Y}_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(\mathbf{X}_i, \mathbf{Y}_j)\right]^{1/2}. \quad (1)$$

An unbiased estimator for $\mathcal{M}^2[\mathcal{F}, P, Q]$ (when $m = n$) has also been derived in Borgwardt et al. (2006):

$$\mathcal{M}^2_u[\mathcal{F}, \mathbf{X}_{1:m}, \mathbf{Y}_{1:m}] = \frac{1}{m(m-1)} \sum_{i,j} h(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j), \quad (2)$$

which is a one sample U-statistic with $h(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j) \doteq k(\mathbf{X}_i, \mathbf{X}_j) + k(\mathbf{Y}_i, \mathbf{Y}_j) - k(\mathbf{X}_i, \mathbf{Y}_j) - k(\mathbf{X}_j, \mathbf{Y}_i)$, where $\mathbf{\Lambda}_i \doteq (\mathbf{X}_i, \mathbf{Y}_i)$, and $\mathbf{\Lambda}_{1:m} = (\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_m)$ are i.i.d. random variables. From the r.h.s. of Lemma 3, one can see that $\mathbb{E}[h(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)] = \mathcal{M}^2[\mathcal{F}, P, Q]$, which proves the unbiasedness of the estimator $\mathcal{M}^2_u[\mathcal{F}, \mathbf{X}_{1:m}, \mathbf{Y}_{1:m}]$.

3. The Copula of Distributions

Below we review a few important properties of the copula of multivariate distributions that we will use in our work (Nelsen, 1998).

The copula plays an important role when we study the dependence among random variables. The marginal variables $X^1, \ldots, X^d$ are independent from each other, if and only if the copula distribution is the multivariate uniform distribution. In turn, we can measure the dependence of the $X^1, \ldots, X^d$ random variables by measuring how far the copula distribution is from the uniform distribution. The copula contains all the information that we need to measure dependence, and it is invariant to any nonlinear strictly increasing transformations of the marginal variables.

The copula can be defined by the Sklar's theorem (Sklar, 1959) as follows. Let $\mathbf{X} = (X^1, \ldots, X^d) \in \mathbb{R}^d$ be a random variable. Denote the marginal cdf's of X^j by $F_j : \mathbb{R} \to [0, 1]$. Sklar's theorem states that a multivariate cumulative distribution function $H(x_1, \ldots, x_d) = \Pr(X^1 \leq x_1, \ldots, X^d \leq x_d)$ can be written as $H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$, where C is a unique distribution function on the range of the F_i cdf functions. This distribution function is called the copula of the joint distribution H. The distribution of the copula C is the same as the joint distribution of $\mathbf{Z} = (Z^1, \ldots, Z^d) \doteq \mathbf{F}(\mathbf{X}) = (F_1(X^1), \ldots F_d(X^d)) \in \mathbb{R}^d$ random variables. When the F_i cumulative distribution functions are invertible, then $\mathbf{F}(\mathbf{X})$ have uniformly distributed marginal distributions on $[0, 1]$, and the copula distribution can be calculated as $C(y_1, \ldots, y_d) = H\left(F_1^{-1}(y_1), \ldots, F_d^{-1}(y_d)\right)$, where $0 \leq y_i \leq 1$. The relation of the joint distribution H, marginal distributions F_i, and copula C is illustrated in Figure 2.

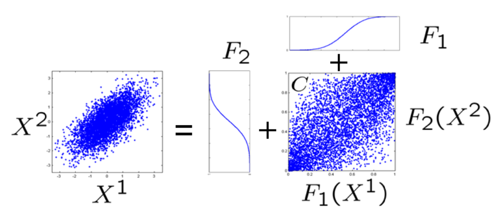


Figure 2. Illustration of the copula. On the left: random samples from a 2-dimensional distribution H. On the right: the copula transformed sample points. They are distributed according to the copula C. Every distribution function H can be rewritten with its copula distribution C and marginal distributions F_1, F_2 as $H(X^1 \leq x_1, X^2 \leq x_2) = C(F_1(X^1 \leq x_1), F_2(X^2 \leq x_2))$. The copula C captures all the dependence between X^1 and X^2. The marginal variables, X^1 and X^2, are independent iff the copula distribution C is the uniform distrbution.

4. Dependence Estimation

Let $\mathbf{U} = (U^1, \ldots, U^d) \in [0,1]^d$ be a random variable with uniform distribution on the d-dimensional unit cube, $\mathbf{U} \sim U[0,1]^d$. We define the dependence among continuous random variables $X^1, \ldots, X^d$ as the *MMD distance between the joint copula and the d-dimensional uniform distribution*:

$$I(X^1, \ldots, X^d) \doteq \mathcal{M}(\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}}).$$

Definition 4. *Let $x_1, x_2 \in \mathbb{R}$. A function $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing, if $g(x_1) < g(x_2)$ for all $x_1 < x_2$.*

It is easy to see that $I(X^1, \ldots, X^d) \geq 0$, and $I(X^1, \ldots, X^d) = I(g_1(X^1), \ldots, g_d(X^d))$ for any g_i strictly increasing functions. In other words, $I(X^1, \ldots, X^d)$ is invariant to strictly increasing transformations of the marginal variables.

The following lemma states that $I(X^1, \ldots, X^d)$ is indeed a well-defined dependence measure when kernel k is universal.

Lemma 5. *Let the kernel k be universal on $[0,1]^d \times [0,1]^d$. Then $I(X^1, \ldots, X^d) = 0$, if and only if $X^1, \ldots, X^d$ are independent of each other.*

In what follows we will provide a consistent estimator for $I(\mathbf{X}) = I(X^1, \ldots, X^d)$. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel function of RKHS $\mathcal{H}$, and let $\mathbf{Z} \doteq \mathbf{F}(\mathbf{X})$ be a random variable drawn from the copula. Introduce the following terms:

$$\mu[P_{\mathbf{Z}}] \doteq \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}}[k((Z^1, \ldots, Z^d), \cdot)],$$
$$\mu[P_{\mathbf{U}}] \doteq \mathbb{E}_{\mathbf{U}\sim P_{\mathbf{U}}}[k((U^1, \ldots, U^d), \cdot)].$$

Thanks to Lemma 3, it is easy to see that $I^2(\mathbf{X}) = \mathcal{M}^2(\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}}) = \|\mu[P_{\mathbf{Z}}] - \mu[P_{\mathbf{U}}]\|^2_{\mathcal{H}} = \mathbb{E}_{\mathbf{Z},\mathbf{Z}'\sim P_{\mathbf{Z}}}[k(\mathbf{Z}, \mathbf{Z}')] - 2\mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}},\mathbf{U}\sim P_{\mathbf{U}}}[k(\mathbf{Z}, \mathbf{U})] + \mathbb{E}_{\mathbf{U},\mathbf{U}'\sim P_{\mathbf{U}}}[k(\mathbf{U}, \mathbf{U}')]$.

Our goal is to estimate $I(\mathbf{X})$ using the $\mathbf{X}_{1:m}$ i.i.d. sample. This expression is the expected value of the kernel k evaluated in random variables drawn from the uniform and the copula distributions. Assume that we already have a $\mathbf{Z}_{1:m}$ i.i.d. sample from the copula distribution. For simple kernel functions, the expectation w.r.t. the uniform distribution has a simple form. For example, when we use the Gaussian kernel, we have the following unbiased estimator for $I^2(\mathbf{X})$:

$$\mathcal{M}^2_u[\mathcal{F}, \mathbf{Z}_{1:m}, P_{\mathbf{U}}] = \frac{1}{m(m-1)} \sum_{i\neq j} k(\mathbf{Z}_i, \mathbf{Z}_j) - \frac{2}{m} \sum_{i=1}^m \prod_{j=1}^d \int_0^1 \exp\left(\frac{-(\mathbf{Z}_i^j - u)^2}{2\sigma^2}\right) \mathrm{d}u + \left(\int_0^1 \int_0^1 \exp\left(\frac{-(u-u')^2}{2\sigma^2}\right) \mathrm{d}u\mathrm{d}u'\right)^d,$$

which can be expressed by the erf Gauss error function. For more complicated kernels, however, these integrals can not be calculated analytically, therefore we need to approximate them by sampling. In what follows we will investigate this case.

Let $\mathbf{U}_{1:n} = \mathbf{U}_1, \ldots, \mathbf{U}_n$ be an i.i.d. sample drawn from the $U[0,1]^d$ distribution, and let $\mathbf{X}$, $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be i.i.d. samples having distribution $P_{\mathbf{X}}$. The $F_1, \ldots, F_d$ distribution functions are unknown, but we can estimate them efficiently using the empirical distribution functions. For $x, x^j \in \mathbb{R}$ and $1 \leq j \leq d$, let

$$\widehat{F}_j(x) \doteq \widehat{F}_j(x; X^j_{1:m}) \doteq \frac{1}{m}|\{i : 1 \leq i \leq m, x \leq X_i^j\}|$$
$$\widehat{\mathbf{F}}(x^1, \ldots, x^d) \doteq (\widehat{F}_1(x^1), \ldots, \widehat{F}_d(x^d)) \in \mathbb{R}^d.$$

We call the maps $\mathbf{F}$, $\hat{\mathbf{F}}$ the *copula transformation*, and the *empirical copula transformation*, respectively. The sample $(\widehat{\mathbf{Z}}_1, \ldots, \widehat{\mathbf{Z}}_m) \doteq (\widehat{\mathbf{F}}(\mathbf{X}_1), \ldots, \widehat{\mathbf{F}}(\mathbf{X}_m)) \in \mathbb{R}^d$ is called the empirical copula (Dedecker et al., 2007). Note that the j-th coordinate of $\widehat{\mathbf{Z}}_i$ $(1 \leq i \leq m)$ equals

$$\widehat{Z}_i^j = \frac{1}{m}\mathrm{rank}(X_i^j, \{X_1^j, X_2^j, \ldots, X_m^j\}),$$

where $\mathrm{rank}(x, A)$ is the number of elements of A less than or equal to x. Also, observe that the random variables $\widehat{\mathbf{Z}}_1, \ldots, \widehat{\mathbf{Z}}_m$ are not even independent. Nonetheless, as we will see from Lemma 7, the empirical copula $(\widehat{\mathbf{Z}}_1, \ldots, \widehat{\mathbf{Z}}_m)$ is a good approximation of an i.i.d. sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_m) \doteq (\mathbf{F}(\mathbf{X}_1), \ldots, \mathbf{F}(\mathbf{X}_m))$ from the copula distribution of $P_{\mathbf{X}}$. Using (2), we have that

$$\mathcal{M}^2_u[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}] = \frac{1}{m(m-1)} \sum_{i\neq j} \Big[k(\mathbf{Z}_i, \mathbf{Z}_j) + k(\mathbf{U}_i, \mathbf{U}_j) - k(\mathbf{Z}_i, \mathbf{U}_j) - k(\mathbf{U}_i, \mathbf{Z}_j)\Big].$$

From (1), we can also see that

$$\mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}] = \Bigg[\frac{1}{m^2} \sum_{i,j=1}^m k(\mathbf{Z}_i, \mathbf{Z}_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(\mathbf{U}_i, \mathbf{U}_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(\mathbf{Z}_i, \mathbf{U}_j)\Bigg]^{1/2}.$$

In these expressions $\mathbf{Z}_{1:m}$ is not available to us. We estimate them using the empirical copula, $\widehat{\mathbf{Z}}_j \doteq \widehat{\mathbf{F}}(\mathbf{X}_j)$, $j = 1, \ldots, m$. An estimator for $I^2(\mathbf{X})$ can be given by $\widehat{I}^2_u(\mathbf{X}_{1:m})$, where

$$m(m-1)\widehat{I}^2_u(\mathbf{X}_{1:m}) \doteq m(m-1)\mathcal{M}^2_u[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] = \sum_{i\neq j} \Big[k(\widehat{\mathbf{Z}}_i, \widehat{\mathbf{Z}}_j) + k(\mathbf{U}_i, \mathbf{U}_j) - k(\widehat{\mathbf{Z}}_i, \mathbf{U}_j) - k(\mathbf{U}_i, \widehat{\mathbf{Z}}_j)\Big].$$

To calculate this quantity, we only need the ranks of the marginal variables in the sample. Note that

$\widehat{I}_u^2(\mathbf{X}_{1:m})$ is not an unbiased estimator of $I(\mathbf{X})$, but we keep the notation $\widehat{I}_u^2$ to denote that it is derived from the estimator $\mathcal{M}_u^2$.

Using the definition of $\mathcal{M}_b$, we can also propose another estimator for $I(\mathbf{X})$:

$$\widehat{I}_b(\mathbf{X}_{1:m}) \doteq \mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] = \Bigg[\frac{1}{m^2}\sum_{i,j=1}^{m} k(\widehat{\mathbf{Z}}_i, \widehat{\mathbf{Z}}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{U}_i, \mathbf{U}_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(\widehat{\mathbf{Z}}_i, \mathbf{U}_j)\Bigg]^{1/2}.$$

Both estimators are extremely simple to implement requiring only kernel evaluations on the transformed data and the uniform variables. One can also see that the estimators are robust assuming k is bounded in $[0,1]^d \times [0,1]^d$ (but can be unbounded outside of this region, e.g. polynomial kernel). Thanks to the empirical copula transformation, we only need rank statistics $(\widehat{\mathbf{Z}}_{1:m})$ in the estimation, but the actual values of $\mathbf{X}_{1:m}$ sample points are not used. The contribution of one single sample point is diminishing in the estimator as we increase the sample size. Therefore, one arbitrarily large outlier sample point cannot ruin the statistics arbitrarily badly. For more discussion on this, see the Appendix.

In what follows we will analyze the theoretical properties of these estimators. Assume that the kernel function $k(\cdot, \mathbf{z})$ is uniformly Lipschitz continuous on $[0,1]^d$, i.e. there exists $L > 0$ such that for all $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2 \in [0,1]^d$ we have that $|k(\mathbf{z}_1, \mathbf{z}) - k(\mathbf{z}_2, \mathbf{z})| \le L\|\mathbf{z}_1 - \mathbf{z}_2\|$. A typical example is the Gaussian kernel, for which it holds that there exists $L > 0$ such that for all $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2 \in [0,1]^d$

$$\left|\exp(-\frac{\|\mathbf{z}_1 - \mathbf{z}\|^2}{2\sigma^2}) - \exp(-\frac{\|\mathbf{z}_2 - \mathbf{z}\|^2}{2\sigma^2})\right| \le L\|\mathbf{z}_1 - \mathbf{z}_2\|.$$

Lemma 6. *For all $\mathbf{z}_i \in [0,1]^d$, $1 \le i \le 4$, we have that*

$$|k(\mathbf{z}_1, \mathbf{z}_2) - k(\mathbf{z}_3, \mathbf{z}_4)| \le L\|\mathbf{z}_1 - \mathbf{z}_3\| + L\|\mathbf{z}_2 - \mathbf{z}_4\|.$$

The effect of the empirical copula transformation can be studied by a version of the classical Kiefer-Dvoretzky-Wolfowitz theorem due to Massart; see e.g. Devroye & Lugosi (2001). As a simple implication of this theorem, one can show that $\widehat{\mathbf{F}}$ is a consistent estimator of $\mathbf{F}$, and the convergence is uniform:

Lemma 7 (Convergence of the empirical copula)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be an i.i.d. sample from a probability distribution over $\mathbb{R}^d$ with marginal cdf's $F_1, \ldots, F_d$. Let $\mathbf{F}(\mathbf{X})$ be the copula defined above, and let $\widehat{\mathbf{F}}(\mathbf{X}_{1:m})$ be the empirical copula transformation. Then, for any $\epsilon \ge 0$,*

$$\Pr\left[\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_2 > \epsilon\right] \le 2d\exp(-\frac{2m\epsilon^2}{d}) .$$

Let $0 \le k(x,y) \le K$ be a bounded kernel function. The following theorems state the almost sure consistency of the dependence estimators, and provide upper bounds on the rate of convergence.

Theorem 8 (Almost sure consistency)**.** *Almost surely we have that*

$$|\widehat{I}_u^2(\mathbf{X}_{1:m}) - I^2(\mathbf{X})| = \mathcal{O}\left(\max\left\{\sqrt{\frac{dL^2}{m}\log(4dm^2)}, \sqrt{\frac{K^2}{m}\log(4m^2)}\right\}\right).$$

From the below theorem it follows that when n grows fast enough, then $\widehat{I}_b$ is almost surely consistent as well.

Theorem 9 (Almost sure consistency)**.** *Let $n = g(m)$ for some function g such that $\lim_{m\to\infty} g(m) = \infty$. Almost surely it holds that*

$$|\widehat{I}_b(\mathbf{X}_{1:m}) - I(\mathbf{X})| = \mathcal{O}\Bigg(\max\Bigg\{\left(\frac{8dL^2}{m}\log(4dm^2)\right)^{1/4}, \left(\frac{2K(m+n)}{mn}\log(4m^2)\right)^{1/2}\Bigg\} + \left(\frac{K}{m}\right)^{1/2} + \left(\frac{K}{n}\right)^{1/2}\Bigg).$$

As these bounds show, the proposed dependence estimators can be used in high-dimensions as well; they do not suffer from the curse of dimensionality. Based on these estimators, one can derive independence tests too. For details, see the Appendix.

5. Feature Selection

The above defined $I(\mathbf{X})$ dependence measure is invariant to strictly increasing transformations of the marginal variables. In this section we discuss the benefits of this property in the feature selection problem.

Let us have d real valued features $\{X^1, \ldots, X^d\}$, and a target value Y. Numerous feature selection methods use dependence estimation for selecting the most relevant features to predict the target value Y. If we want to select h features, then one obvious approach would be to select those h features that together have the highest dependence with Y. This subset selection problem, unfortunately, is very difficult. Therefore, several approximations and heuristics have been proposed. For example, according to the so-called max-relevance criterion (Peng & Ding, 2005), our goal is to select a feature set $S \subseteq \{X^1, \ldots, X^d\}$, which maximizes the average dependence between the features and the target:

$$\widehat{S} = \arg\max_{S} \frac{1}{|S|} \sum_{X^i \in S} I(X^i, Y). \tag{3}$$

This approach might select highly redundant features, i.e. the dependence among these features could be large. This redundancy can be measured by the expression $\sum_{X^i, X^j \in S} I(X^i, X^j)/|S|^2$.

When two features highly depend on each other, then probably we do not lose too much if we remove one of them. Therefore, our goal is to maximize relevance while minimizing the redundancy among the features

$$\widehat{S} = \arg\max_{S} \sum_{X^i \in S} \frac{I(X^i, Y)}{|S|} - \sum_{X^i, X^j \in S} \frac{I(X^i, X^j)}{|S|^2}. \tag{4}$$

All we need is a good estimator for $I(X^i, X^j)$ and $I(X^i, Y)$ dependencies. Equation (3) and (4) objectives are popular tools for feature selection. Here we will not discuss the advantages and disadvantages of them. We, however, would like to point out that when someone uses objectives that involves dependence estimation, then we want these dependencies to be invariant to strictly increasing transformations of the marginal variables.

6. Numerical Illustrations

We illustrate the theoretical contributions of this paper through a series of numerical experiments demonstrating properties of the copula-based kernel dependency measure.

The $\mathcal{M}(\mathcal{F}, P_{\mathbf{X}}, \prod_{i=1}^d P_{X^i})$ measure could also be used directly, without copula transformation, to estimate dependence. In order to use this approach, we need to generate m sample points from the product distributions of the marginals. Let $\tau_i(1:m)$, $(1 \le i \le d)$ denote independent random permutations of $\{1, \ldots, m\}$. Then $\Pi[\mathbf{X}_{1:m}] \doteq (X^1_{\tau_1(1:m)}, X^2_{\tau_2(1:m)}, \ldots, X^d_{\tau_d(1:m)})^T$ can be considered as samples from the $\prod_{i=1}^d P_{X^i}$ distribution. In other words, if $\mathbf{X}_{1:m}$ is stored in a $d \times m$ dimensional sample matrix and we independently permute the elements of each row, then the distributions of the rows (the marginal distributions of $\mathbf{X}$) remain the same, but they become independent from each other. For brevity, we will call the $\mathcal{M}(\mathcal{F}, P_{\mathbf{X}}, \prod_{i=1}^d P_{X^i})$ quantity MMD dependence measure.

6.1. Feature Selection

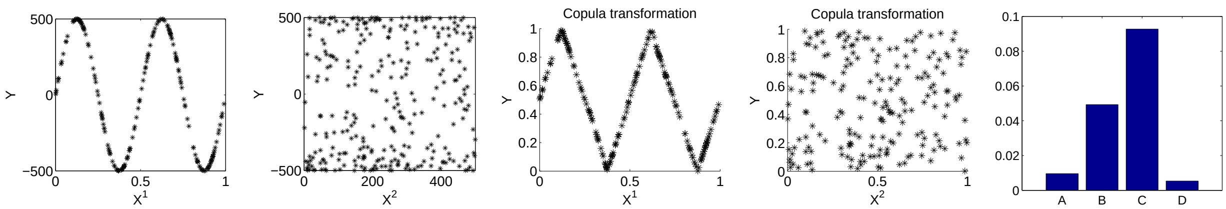
In this experiment we show that $I(\mathbf{X})$ can achieve better performance in feature selection than MMD without copula transformation ($\mathcal{M}(\mathcal{F}, P_{\mathbf{X}}, \prod_{i=1}^d P_{X^i})$).

We constructed the following random variables: $X^1 \sim U[0,1]$, $X^2 \sim U[0,500]$, $Y = 500\sin(4\pi X^1)$. The task in this experiment was to choose the feature between X^1 and X^2 that contains the most information about Y. This feature is of course X^1 since Y is a deterministic function of it, and X^2 is independent of Y; it does not contain any information about Y. 300 sample points from the joint distrbutions of (X^1,Y) and (X^2,Y) are shown in Figure 3(a) and Figure 3(b), respectively. The empirical copula transformed points of (Y, X^1) and (Y, X^2) are displayed in Figure 3(c) and Figure 3(d). When we simply use MMD without copula transformation ($\mathcal{M}(\mathcal{F}, P_{\mathbf{Y},X^i}, P_Y \times P_{X^i})$), then interestingly we got that the estimated dependence between Y and X^1 ($\mathcal{M}_b(\mathcal{F}, (Y, X^1)_{1:m}, \Pi[(Y, X^1)_{1:m}]$), column (A) of Figure 3(e)) was smaller than the estimated dependence between between Y and X^2 ($\mathcal{M}_b(\mathcal{F}, (Y, X^2)_{1:m}, \Pi[(Y, X^2)_{1:m}]$), column (B) of Figure 3(e)). As we can see in this problem, the MMD without copula transformation could not select the right feature. However, when we used copula transformation, then the estimated dependence was larger between Y and X^1 than between Y and X^2. The values of $\widehat{I}^b((Y, X^1)_{1:m})$ and $\widehat{I}^b((Y, X^2)_{1:m})$ are shown in the (C) and (D) columns of Figure 3(e). In this experiment we used Gaussian kernel with $\sigma = 1$.

6.2. Feature Standardization

A frequently used feature preprocessing step is to standardize the features, that is, linearly transform them to have zero mean and unit variance ($\Xi[X]$). One might wonder if this simple transformation can solve the problem of Section 6.1. Below we show an example, where that we have only two zero mean unit variance features, and the MMD feature selection method that is not invariant to the strictly increasing transformations of the features selects a feature that is actually independent from the target value.

Let $U \sim U[0,1]$, $X^1 \doteq \Xi[1/U^2]$, $V \sim U[0,1]$ $X^2 \doteq \Xi[V]$ independent random variables, and let $Y \doteq \Xi[\sin(4\pi X^1)]$. The variables are standardized so they have zero mean and standard variation 1. We sampled 4,000 i.i.d. observations from our observed features X^1 and X^2. The task again was to select the feature that contains the most information about Y. The solution to this problem is X^1 again. The meanings of the columns in Figure 4 are the same as in Figure 3(e). When we simply use MMD without copula transformation, then the estimated dependence between between Y and X^1 was smaller than between Y and X^2 ($\mathcal{M}_b(\mathcal{F}, (Y, X^1)_{1:m}, \Pi[(Y, X^1)_{1:m}])$ and $\mathcal{M}_b(\mathcal{F}, (Y, X^2)_{1:m}, \Pi[(Y, X^2)_{1:m}])$ in column (A) and (B), respectively). The MMD without copula



(a) Y and X^1 (b) Y and X^2 (c) Copula of (Y, X^1) (d) Copula of (Y, X^2) (e) Dependence

Figure 3. (a) Features Y and X^1. (b) Features Y and X^2. (c) Copula distribution of (Y, X^1). (d) Copula distribution of (Y, X^2). Notations of the bar plot in (e): (A) MMD dependence between Y and X^1 $(\mathcal{M}_b(\mathcal{F}, (Y, X^1)_{1:m}, \Pi[(Y, X^1)_{1:m}])$. (B) MMD dependence between Y and X^2 $(\mathcal{M}_b(\mathcal{F}, (Y, X^2)_{1:m}, \Pi[(Y, X^2)_{1:m}])$. (C) copula based dependence between Y and X^1 $(\widehat{I}^b((Y, X^1)_{1:m}))$. (D) copula based dependence between Y and X^2 $(\widehat{I}^b((Y, X^2)_{1:m}))$.

transformation could not select the right feature. However, when we use copula transformation first, then we can see that the estimated dependence between Y and X^1 is larger than between Y and X^2, as expected. (C) and (D) show $\widehat{I}^b((Y, X^1)_{1:m})$ and $\widehat{I}^b((Y, X^2)_{1:m})$, respectively.

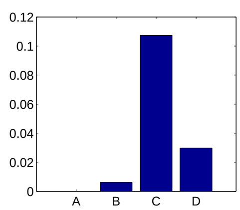


Figure 4. (A-B) Dependence estimation without copula transformation: (A) MMD between Y and X^1, $\mathcal{M}_b(\mathcal{F}, (Y, X^1)_{1:m}, \Pi[(Y, X^1)_{1:m}])$, (B) MMD between Y and X^2, $\mathcal{M}_b(\mathcal{F}, (Y, X^2)_{1:m}, \Pi[(Y, X^2)_{1:m}])$. (C-D) Dependence estimation with copula transformation: (C) $\widehat{I}^b((Y, X^1)_{1:m})$, (D) $\widehat{I}^b((Y, X^2)_{1:m})$.

6.3. Housing Dataset

In the following experiment we study our estimators on the Housing dataset from the UCI repository (Frank & Asuncion, 2010). The dataset contains 506 instances of 14 real valued attributes. The attributes contain various features including per capita crime rate by town, full-value property-tax rate per \$10000, average number of rooms per dwelling, percentage of lower status of the population, median value of owner-occupied homes in \$1000's, etc. Our goal is to predict some of these attributes and select the most important features for this prediction. Since the dataset contains very different features, it is highly nontrivial how to scale them for feature selection when the applied dependence measure is not invariant to strictly increasing transformations of the marginals. This, however, is not an issue for our proposed dependence measure. In this experiment our goal was to predict the "median value of owner-occupied homes in \$1000's" (feature 14) using one single feature. We used $m = n = 300$ instances for training, and the rest of the data for testing. We applied Gaussian kernel ($\sigma^2 = 1/12$) in the estimators. The MMD without copula transformation chose the "average number of rooms per dwelling" (feature 6) as the closest feature. When instead of MMD we used the proposed $\widehat{I}_b$ estimator, it selected the "lower status of the population" (feature 13). To study the prediction errors of the selected features, we trained linear regressors for each feature using them as explanatory variables. The prediction errors on the test data are shown in Figure 5. In this experiment the smallest error was achieved by the feature that $\widehat{I}_b$ selected (feature 13). MMD without the copula transformation selected the feature that gave only the second smallest error (feature 6).

Low-dimensional embedding can help us visualize the pairwise dependence structure of random variables. For each feature X^i, X^j, we estimated the $d(i,j) = \exp(-I(X^i, X^j))$ quantities. This $d(i,j)$ is large when X^i, X^j is independent, and small when the dependence between them is large. We considered them as "distances" (although the triangle inequality does not hold between them), and then applied multidimensional scaling to embed them into a 2d space. The Housing dataset was used in this experiment too using the same set-up as in the previous study. To estimate the dependence between the features, we tested again $\widehat{I}_b$ (Figure 6(a)) and MMD without copula transformation (Figure 6(b)). We can observe that the locations of these embedded points are very different. If we applied any strictly increasing transformations to the marginal variables, it would not affect the embedding with copula transformation, but we would get very different results when we use MMD without copula transformation. For more numerical experiments, see the supplementary material.

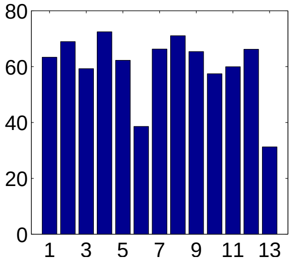
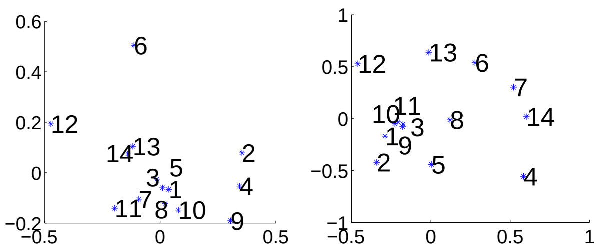


Figure 5. Prediction errors of the features.



(a) Copula-MMD (b) MMD

Figure 6. Low-dimensional embedding of the features using dependence as a proximity measure.

7. Discussion and Conclusion

We introduced a new RKHS-based dependence measure that operates on the copula of continuous distributions. We have shown that the dependence measure is invariant to strictly increasing transformations of the marginal variables, and this property is important in feature selection and low-dimensional embedding of distributions. We also proposed estimators that are almost surely consistent, robust, use rank statistics only, and do not suffer from the curse of dimensionality. We derived upper bounds on the rates of convergence and illustrated the theory through a series of numerical experiments.

References

Bach, Francis R. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H., Schölkopf, B., and Smola, A. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Dedecker, J., Doukhan, P., Lang, G., Leon, J.R., Louhichi, S., and Prieur, C. *Weak Dependence: With Examples and Applications*, volume 190 of *Lecture Notes in Statistics*. Springer, 2007.

Devroye, L. and Lugosi, G. *Combinatorial Methods in Density Estimation*. Springer, 2001.

Fernandes, A. and Gloor, G. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, 26:1135–1139, 2010.

Fortet, R. and Mourier, E. Convergence de laréparation empirique vers la réparation théorique. *Ann. Scient. École Norm*, 70:266–285, 1953.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Gretton, A., Herbrich, R., and Smola, A. The kernel mutual information. In *Proc. ICASSP*, 2003.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pp. 63–77, 2005.

Nelsen, R. *An Introduction to Copulas (Lecture Notes in Statistics)*. Springer, 1998.

Pál, D., Póczos, B., and Szepesvári, Cs. Estimation of renyi entropy and mutual information based on generalized nearest-neighbor graphs. In *NIPS*, 2010.

Peng, H. and Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 27, 2005.

Rényi, A. On measures of dependence. *Acta. Math. Acad. Sci. Hungar*, 10:441–451, 1959.

Rényi, A. On measure of entropy and information. In *4th Berkeley Symposium on Math., Stat., and Prob.*, pp. 547–561, 1961.

Schweizer, B. and Wolff, E. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9, 1981.

Sklar, A. Fonctions de rpartition n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2: 67–93, 2001.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.

Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *J. Statist. Phys.*, 52(1-2):479–487, 1988.

Appendix—Supplementary Material

A. Dependence Applications

Mutual information and other dependence estimators have been used, for example, in causality detection (Hlaváckova-Schindler et al., 2007), feature selection (Peng & Ding, 2005), clustering (Aghagolzadeh et al., 2007), optimal experimental design (Lewi et al., 2007), structure learning (Chow & Liu, 1968), prediction of protein structures (Adami, 2004), fMRI data processing (Chai et al., 2009), boosting and facial expression recognition (Shan et al., 2005). They have also been used for image registration (Kybic, 2006; Hero et al., 2002b;a), independent component and subspace analysis (Learned-Miller & Fisher, 2003; Póczos & Lőrincz, 2005; Hulle, 2008; Szabó et al., 2007). The so-called maximal information coefficient has been used recently for global health, gene expression, major-league baseball, and the human gut microbiota datasets to identify known and novel relationships (Reshef et al., 2011).

B. Dependence Axioms

Originally Rényi (1959) formalized the dependence axiom (iv) differently than did Schweizer & Wolff (1981): Rényi required $I(X_1, \ldots, X_d)$ to be invariant to one-to-one transformation of X_i random variables. It turned out, however, that this axiom is especially difficult to satisfy. Rényi himself showed that that among several other well-known dependence measures the only one which satisfies all his axioms is the maximal correlation coefficient. Nonetheless, Hall (1969) pointed out that this measure has a few serious drawbacks, for example it equals 1 too often, it is too strong for nonparametric measures, and generally is very difficult to estimate. Therefore, Schweizer & Wolff (1981) modified this axiom to require invariance to strictly increasing transformations only.

C. Invariance to Strictly Increasing Transformations

As a sanity check, we demonstrate that $I(\mathbf{X})$ is indeed invariant to strictly increasing transformations of the marginal variables, but MMD is not invariant.

We generated features using the following model: We let $\mathbf{U} \sim U[-1,1]^2$ be a column vector, and then we set $\mathbf{X} \doteq (X^1, X^2)^T \doteq Q\mathbf{U}$, where $Q \in \mathbb{R}^{2\times 2}$ was a randomly chosen invertible matrix such that for the variances we had $\sigma(X^1) = 1$, $\sigma(X^2) = 1$. Therefore (X^1, X^2) had a joint distribution on a rotated 2-dimensional parallelogram. Figure 7(a) shows 400 i.i.d. samples from the features X^1 and X^2. They have zero mean and unit variance. We also constructed another feature set, which consisted of the following nonlinear strictly increasing transformations of X^i: $\widetilde{X}^1 \doteq 1 + (X^1)^3$, $\widetilde{X}^2 \doteq 2 + \tanh(X^2)$. 400 sample points from the joint distribution of $(\widetilde{X}^1, \widetilde{X}^2)$ is shown in Figure 7(b). The empirical copula of (X^1, X^2) and $(\widetilde{X}^1, \widetilde{X}^2)$ are shown in Figure 7(c) and Figure 7(d), respectively. These distributions are the same as expected.

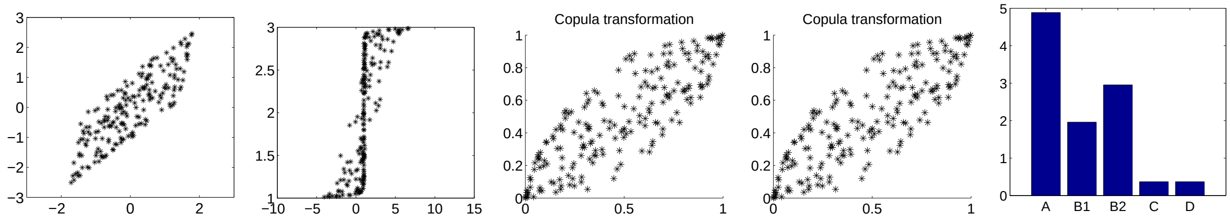
Now we show that when we use $\mathcal{M}(\mathcal{F}, P_{\mathbf{X}}, \prod P_{X^i})$ without copula transformation, then the estimated dependence values can have very different values. For a real valued random variable X, let $\Xi[X]$ denote the standardized variable, that is, $\Xi[X] \doteq (X - \mathbb{E}[X])/\sigma(X)$, which has zero mean and unit variance. We will use the Ξ operator to standardize the $X_{1:m}$ sample too, and in this case $\mathbb{E}[X]$, $\sigma(X)$ is estimated from the empirical mean and empirical standard variation.

Using $m = n = 4000$ sample size and Gaussian kernel[1] with $\sigma = 1$, we calculated the MMD dependence between the marginal variables of the original data (that is, $\mathcal{M}_b(\mathcal{F}, \mathbf{X}_{1:m}, \Pi[\mathbf{X}_{1:m}])$). We also calculated this dependence between the transformed features $(\mathcal{M}_b(\mathcal{F}, \widetilde{\mathbf{X}}_{1:m}, \Pi[\widetilde{\mathbf{X}}_{1:m}]))$, and the transformed then standardized features $(\mathcal{M}_b(\mathcal{F}, \Xi[\widetilde{\mathbf{X}}_{1:m}], \Pi[\Xi[\widetilde{\mathbf{X}}_{1:m}]]))$. These quantities are shown in columns (A), (B1), (B2) of Figure 7(e), respectively. As we can see, these values are very different. Although both X^i and $\Xi[\widetilde{X}^i]$ have zero mean and unit variance, and the marginal variables in both cases contain the same information about each other, still their MMD based dependencies without copula transformation are very different. In the next section we show that this can lead to serious problems in feature selection and low-dimensional embedding. However, when we used copula transformation on the original and transformed features, then the estimated dependencies were the same ($\widehat{I}_b(\mathbf{X}_{1:m})$ and $\widehat{I}_b(\widetilde{\mathbf{X}}_{1:m})$ in the (C) and (D) column of Figure 7(e)). This illustrates that $I(X^1, X^2) = I(f(X^1), g(X^2))$, for f, g strictly increasing functions.

D. Low-dimensional Embedding of Features

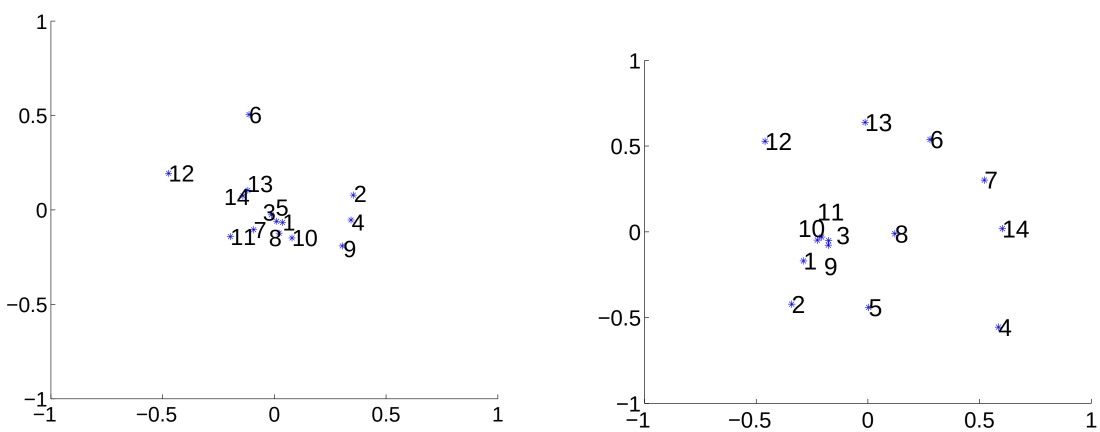
In Figure 8 we show the same results as those of in Figure 6, but here we set the scale to [-1,1] on both axis.

[1] $k(\mathbf{x}, \mathbf{y}) = \exp(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2})$



(a) (X^1, X^2) (b) $(\widetilde{X}^1, \widetilde{X}^2)$ (c) Copula (X^1, X^2) (d) Copula $(\widetilde{X}^1, \widetilde{X}^2)$ (e) Dependence

Figure 7. (a) Original features $(X^1,\ X^2)$. (b) Features transformed by strictly increasing functions $(\widetilde{X}^1, \widetilde{X}^2)$. (c) The copula transformed points of the original features $(X^1,\ X^2)$. (d) The copula transformed points of $(\widetilde{X}^1, \widetilde{X}^2)$. Notations of the bar plot in (e): (A) MMD dependence of the original features $(\mathcal{M}_b(\mathcal{F}, \mathbf{X}_{1:m}, \Pi[\mathbf{X}_{1:m}]))$. (B1) MMD dependence of the transformed features $(\mathcal{M}_b(\mathcal{F}, \widetilde{\mathbf{X}}_{1:m}, \Pi[\widetilde{\mathbf{X}}_{1:m}]))$. (B2) MMD dependence of the transformed and then standardized features $(\mathcal{M}_b(\mathcal{F}, \Xi[\widetilde{\mathbf{X}}_{1:m}], \Pi[\Xi[\widetilde{\mathbf{X}}_{1:m}]]))$ (C) copula based dependence of the original features $(\widehat{I}_b(\mathbf{X}_{1:m}))$. (D) copula based dependence of the transformed features $(\widehat{I}_b(\widetilde{\mathbf{X}}_{1:m}))$.



(a) Copula-MMD (b) MMD

Figure 8. Low-dimensional embedding of the features using dependence as a proximity measure.

E. Almost Sure Consistency

First we derive an upper bound on the error we make due to the application of the empirical copula instead of the true copula in the $\widehat{I}_b$ and $\widehat{I}_u^2$ estimators.

Lemma 10.

$$\Pr(|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| > \epsilon) \leq 2d \exp(-\frac{m\epsilon^4}{8dL^2}).$$

Note that the right hand side in this inequality does not depend on n.

Lemma 11.

$$\Pr(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}]| > \epsilon) \leq 2d \exp(-\frac{m\epsilon^2}{8dL^2}).$$

Proof of Lemma 10 and Lemma 11.

Proof. As we increase the sample size, the empirical copula converges to the true copula and hence $\|\widehat{\mathbf{Z}}_j - \mathbf{Z}_j\|$ error terms decrease. Nonetheless, the number of these error terms in Lemma 10 and in Lemma 11 increases with the sample size m, and thus it is not immediately obvious if we can get a good enough upper bound on this error. In particular, there might exist consistent divergence estimators that will not lead to consistent dependence estimation when applied on the empirical copula. Below we show that for the MMD estimators this is not the case.

Since $|a^{1/2} - b^{1/2}| \leq |a - b|^{1/2}$ and using the Lipschitz continuity, we have that

$$|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]|$$

$$\leq \left(\frac{1}{m^2} \sum_{i,j}^{m} |k(\widehat{\mathbf{Z}}_i, \widehat{\mathbf{Z}}_j) - k(\mathbf{Z}_i, \mathbf{Z}_j)| + \frac{2}{mn} \sum_{i,j}^{m,n} |k(\widehat{\mathbf{Z}}_i, \mathbf{U}_j) - k(\mathbf{Z}_i, \mathbf{U}_j)| \right)^{1/2}$$

$$\leq \left(\frac{1}{m^2} \sum_{i,j}^{m} L\|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\| + L\|\widehat{\mathbf{Z}}_j - \mathbf{Z}_j\| + \frac{2}{mn} \sum_{i,j}^{m,n} L\|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\| \right)^{1/2} \leq \sup_{1\leq i\leq m} (4L\|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\|)^{1/2}$$

$$= 2L^{1/2} \sup_{1\leq i\leq m} \|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\|^{1/2}.$$

Therefore,

$$\begin{aligned}
&\Pr(|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| > \epsilon) \\
&\leq \Pr(2L^{1/2} \sup_{1\leq i\leq m} \|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\|^{1/2} > \epsilon) \\
&= \Pr(\sup_{1\leq i\leq m} \|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\| > \frac{\epsilon^2}{4L}) \leq 2d \exp(-\frac{2m\epsilon^4}{16dL^2}).
\end{aligned}$$

Similarly, for $\mathcal{M}_u^2$ we have that

$$\begin{aligned}
&|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| \\
&\leq \frac{1}{m(m-1)} \Bigg(\sum_{i\neq j}^{m} |k(\widehat{\mathbf{Z}}_i, \widehat{\mathbf{Z}}_j) - k(\mathbf{Z}_i, \mathbf{Z}_j)| \\
&+ |k(\widehat{\mathbf{Z}}_i, \mathbf{U}_j) - k(\mathbf{Z}_i, \mathbf{U}_j)| + |k(\mathbf{U}_i, \widehat{\mathbf{Z}}_j) - k(\mathbf{U}_i, \mathbf{Z}_j)| \Bigg) \\
&\leq \sup_{1\leq i\leq m} 4L\|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Pr(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| > \epsilon) \\
&\leq \Pr(\sup_{1\leq i\leq m} \|\widehat{\mathbf{Z}}_i - \mathbf{Z}_i\| > \frac{\epsilon}{4L}) \leq 2d \exp(-\frac{2m\epsilon^2}{16dL^2}).
\end{aligned}$$

□

Let $0 \leq k(x, y) \leq K$ be a bounded kernel function. It was shown by Gretton et al. (2008) that the following bound holds for the convergence rate of $\mathcal{M}_u^2$, when the *true* copula is known.

$$\begin{aligned}
&\Pr(|\mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}^2[\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}}]| > \epsilon) \\
&\leq 2\exp(-\frac{\epsilon^2 \lfloor m/2 \rfloor}{8K^2}). \qquad (5)
\end{aligned}$$

Similarly, for the rate of $\mathcal{M}_b$ when the *true* copula is known we have that

$$\Pr\left(|\mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}[\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}}]| > \epsilon + 2\left(\frac{K}{m}\right)^{1/2} + 2\left(\frac{K}{n}\right)^{1/2} \right) \leq 2\exp(-\frac{\epsilon^2 mn}{2K(m+n)}).$$

Putting the pieces together, we arrive at the following consistency theorems and convergence rates.

Theorem 12. *$\widehat{I}_u^2(\mathbf{X}_{1:m})$ is a consistent estimator of $I^2(\mathbf{X})$, and*

$$\begin{aligned}
&\Pr(|\widehat{I}_u^2(\mathbf{X}_{1:m}) - I^2(\mathbf{X})| > 2\epsilon) \\
&< 2d \exp(-\frac{m\epsilon^2}{8dL^2}) + 2\exp(-\frac{\epsilon^2 \lfloor m/2 \rfloor}{8K^2}).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\Pr(|\widehat{I}_u^2(\mathbf{X}_{1:m}) - I^2(\mathbf{X})| > 2\epsilon) \\
&= \Pr(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] - I^2(\mathbf{X})| > 2\epsilon) \\
&\leq \Pr(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}]| > \epsilon) \\
&\quad + \Pr(|\mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}^2(\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}})| > \epsilon) \\
&\leq 2d \exp(-\frac{2m\epsilon^2}{16dL^2}) + 2\exp(-\frac{\epsilon^2 \lfloor m/2 \rfloor}{8K^2}).
\end{aligned}$$

□

Similarly,

Theorem 13. *$\widehat{I}_b(\mathbf{X}_{1:m})$ is a consistent estimator of $I(\mathbf{X})$, and*

$$\begin{aligned}
&\Pr\left(|\widehat{I}_b(\mathbf{X}_{1:m}) - I(\mathbf{X})| > 2\epsilon + 2\left(\frac{K}{m}\right)^{1/2} + 2\left(\frac{K}{n}\right)^{1/2} \right) \\
&< 2d \exp(-\frac{m\epsilon^4}{8dL^2}) + 2\exp(-\frac{\epsilon^2 mn}{2K(m+n)}).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\Pr\left(|\widehat{I}_b(\mathbf{X}_{1:m}) - I(\mathbf{X})| > 2\epsilon + 2\left(\frac{K}{m}\right)^{1/2} + 2\left(\frac{K}{n}\right)^{1/2} \right) \\
&= \Pr\left(|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - I(\mathbf{X})| > 2\epsilon + 2\left(\frac{K}{m}\right)^{1/2} + 2\left(\frac{K}{n}\right)^{1/2} \right) \\
&\leq \Pr\Big(|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| \\
&\quad + |\mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}] - I(\mathbf{X})| > 2\epsilon + 2\left(\frac{K}{m}\right)^{1/2} + 2\left(\frac{K}{n}\right)^{1/2} \Big) \\
&\leq \Pr(|\mathcal{M}_b[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}]| > \epsilon) \\
&+ \Pr\Big(|\mathcal{M}_b[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:n}] - \mathcal{M}(\mathcal{F}, P_{\mathbf{Z}}, P_{\mathbf{U}})| > \epsilon
\end{aligned}$$

$$+2\left(\frac{K}{m}\right)^{1/2}+2\left(\frac{K}{n}\right)^{1/2}\Bigg)$$
$$\leq 2d\exp(-\frac{2m\epsilon^4}{16dL^2})+2\exp(-\frac{\epsilon^2 mn}{2K(m+n)}).$$

□

Proof of Theorem 8

As an application of the Borel-Cantelli lemma, we can also show the almost sure consistency of the estimators.

From Theorem (12), we have that

$$\begin{aligned}
&\Pr(|\widehat{I}_u^2(\mathbf{X}_{1:m})-I^2(\mathbf{X})|>2\epsilon)\\
&\quad<2d\exp(-\frac{m\epsilon^2}{8dL^2})+2\exp(-\frac{\epsilon^2\lfloor m/2\rfloor}{8K^2})\\
&\quad<2\max\left(2d\exp(-\frac{m\epsilon^2}{8dL^2}),2\exp(-\frac{\epsilon^2\lfloor m/2\rfloor}{8K^2})\right)\\
&\quad=\tau(m).
\end{aligned}$$

Let

$$\epsilon(m)=\max\left(\sqrt{\frac{8dL^2}{m}\log(4dm^2)},\sqrt{\frac{8K^2}{\lfloor m/2\rfloor}\log(4m^2)}\right)$$

such that $\tau(m)<1/m^2$

This implies that

$$\sum_{m=1}^{\infty}\Pr(|\widehat{I}_u^2(\mathbf{X}_{1:m})-I^2(\mathbf{X})|>2\epsilon(m))<\sum_{m=1}^{\infty}\frac{1}{m^2}<\infty.$$

By the application of the Borel-Cantelli lemma, we can see that the probability that infinitely many of the events $\{|\widehat{I}_u^2(\mathbf{X}_{1:m})-I^2(\mathbf{X})|>2\epsilon(m)\}$ occur is 0. Therefore,

$$\begin{aligned}
&|\widehat{I}_u^2(\mathbf{X}_{1:m})-I^2(\mathbf{X})|=\mathcal{O}(2\epsilon(m))\\
&=\mathcal{O}\left(\max\{\sqrt{\frac{32dL^2}{m}\log(4dm^2)},\sqrt{\frac{32K^2}{\lfloor m/2\rfloor}\log(4m^2)}\}\right)
\end{aligned}$$

Proof of Theorem 9

The proof of this theorem is similar to proof of Theorem 8, and we omit the details. After some calculations, we have that

$$|\widehat{I}_b(\mathbf{X}_{1:m})-I(\mathbf{X})|=\mathcal{O}\Bigg(2\max\Bigg\{\left(\frac{8dL^2}{m}\log(4dm^2)\right)^{1/4},$$
$$\left(\frac{2K(m+n)}{mn}\log(4m^2)\right)^{1/2}\Bigg\}+2\left(\frac{K}{m}\right)^{1/2}+2\left(\frac{K}{n}\right)^{1/2}\Bigg).$$

F. Independence Test

In this section we provide methods for testing hypothesis H_0: $I(\mathbf{X})=0$, i.e. $X_1,\ldots,X_d$ are independent from each other. The alternative hypothesis is H_1: $I(\mathbf{X})>0$, which implies that the random variables are dependent.

Lemma 14. *Let $m=n$, that is, we have the same number of samples from $P_{\mathbf{X}}$ and $U[0,1]^d$. Under H_0 we have*

$$\Pr\left(\widehat{I}_b(\mathbf{X}_{1:m})>2\epsilon+\left(\frac{2K}{m}\right)^{1/2}\right)$$
$$\leq 2d\exp(-\frac{2m\epsilon^4}{16dL^2})+\exp(-\frac{\epsilon^2 m}{4K}).$$

Proof. We already know from (Gretton et al., 2008) that

$$\Pr(\mathcal{M}_b[\mathcal{F},\mathbf{U}_{1:m},\mathbf{U}_{1:m}]>\epsilon+\left(\frac{2K}{m}\right)^{1/2})\leq\exp(-\frac{\epsilon^2 m}{4K}).$$

Since under H_0 the copula distribution $P_{\mathbf{Z}}$ is the uniform distribution (i.e. $P_{\mathbf{U}_{1:m}}=P_{\mathbf{Z}_{1:m}}$), it is easy to see that

$$\begin{aligned}
&\Pr\left(\widehat{I}_b(\mathbf{X}_{1:m})>2\epsilon+\left(\frac{2K}{m}\right)^{1/2}\right)\\
&=\Pr\left(\mathcal{M}_b[\mathcal{F},\widehat{\mathbf{Z}}_{1:m},\mathbf{U}_{1:m}]>2\epsilon+\left(\frac{2K}{m}\right)^{1/2}\right)\\
&\leq\Pr\Bigg(|\mathcal{M}_b[\mathcal{F},\widehat{\mathbf{Z}}_{1:m},\mathbf{U}_{1:m}]-\mathcal{M}_b[\mathcal{F},\mathbf{Z}_{1:m},\mathbf{U}_{1:m}]|\\
&\quad+\mathcal{M}_b[\mathcal{F},\mathbf{Z}_{1:m},\mathbf{U}_{1:m}]>2\epsilon+\left(\frac{2K}{m}\right)^{1/2}\Bigg)\\
&\leq\Pr(|\mathcal{M}_b[\mathcal{F},\widehat{\mathbf{Z}}_{1:m},\mathbf{U}_{1:m}]-\mathcal{M}_b[\mathcal{F},\mathbf{Z}_{1:m},\mathbf{U}_{1:m}]|>\epsilon)\\
&\quad+\Pr(\mathcal{M}_b[\mathcal{F},\mathbf{U}_{1:m},\mathbf{U}_{1:m}]>\epsilon+\left(\frac{2K}{m}\right)^{1/2})\\
&\leq 2d\exp(-\frac{2m\epsilon^4}{16dL^2})+\exp(-\frac{\epsilon^2 m}{4K}).
\end{aligned}$$

□

Similar independence test can be derived based on the estimator $\widehat{I}_u^2(\mathbf{X}_{1:m})$.

Lemma 15. *Let $m=n$. Under H_0 we have that*

$$\Pr\left(\widehat{I}_u^2(\mathbf{X}_{1:m})>2\epsilon\right)\leq\exp(-\frac{\epsilon^2\lfloor m/2\rfloor}{8K^2})$$
$$+2d\exp(-\frac{2m\epsilon^2}{16dL^2}).$$

Proof. From (5), it is immediate that

$$\Pr(\mathcal{M}_u^2[\mathcal{F}, \mathbf{U}_{1:m}, \mathbf{U}_{1:m}] > 2\epsilon) \leq \exp(-\frac{\epsilon^2 \lfloor m/2 \rfloor}{8K^2})$$

Now,

$$\begin{aligned}
&\Pr\left(\widehat{I}_u^2(\mathbf{X}_{1:m}) > 2\epsilon\right) \\
&\quad = \Pr\left(\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] > 2\epsilon\right) \\
&\quad \leq \Pr\Big(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}]| \\
&\qquad + \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}] > 2\epsilon\Big) \\
&\quad \leq \Pr(|\mathcal{M}_u^2[\mathcal{F}, \widehat{\mathbf{Z}}_{1:m}, \mathbf{U}_{1:m}] - \mathcal{M}_u^2[\mathcal{F}, \mathbf{Z}_{1:m}, \mathbf{U}_{1:m}]| > \epsilon) \\
&\qquad + \Pr(\mathcal{M}_u^2[\mathcal{F}, \mathbf{U}_{1:m}, \mathbf{U}_{1:m}] > \epsilon) \\
&\quad \leq \exp(-\frac{\epsilon^2 \lfloor m/2 \rfloor}{8K^2}) + 2d \exp(-\frac{2m\epsilon^2}{16dL^2}).
\end{aligned}$$

□

G. Robustness

Inspired by Tukey's finite-sample influence curve, we can show that the asymptotic effect of one sample point in the proposed dependence estimators is negligible. For each ϵ there is a threshold number $M = M(\epsilon)$ such that when then sample size m is large enough ($m > M$), then the contribution of one sample point in the estimator is less then ϵ. In that sense the estimator is robust: an arbitrarily large outlier cannot ruin the statistics if we have enough sample points. For example, the empirical average (as the estimator of the mean) does not have this property; it is not a robust estimator. Thanks to the copula transformation, $\widehat{I}_b$ and $\widehat{I}_u^2$ estimators are robust when k is bounded in $[0,1]^d \times [0,1]^d$.

When using copula methods the contribution of one sample point $\mathbf{x}$ in the worst case is

$$\max_x \widehat{I}_u^2(\mathbf{X}_{1:m}, \mathbf{x}) - \widehat{I}_u^2(\mathbf{X}_{1:m}) = \mathcal{O}(\max_{\mathbf{u},\mathbf{u}' \in [0,1]^d} \frac{k(\mathbf{u}, \mathbf{u}')}{m}).$$

Without copula transformation, this is

$$\mathcal{O}(\max_{\mathbf{u},\mathbf{u}' \in \mathbb{R}^d} \frac{k(\mathbf{u}, \mathbf{u}')}{m}),$$

which can be much larger. In particular, when k is not bounded, then this term is infinite. Similarly can be proved that $\widehat{I}_b$ is a robust estimator w.r.t the finite-sample influence curve.

H. Other Proofs

H.1. Proof of Lemma 5.

Proof. Clearly, $P_\mathbf{Z} = P_\mathbf{U}$ if and only if $X^1, \ldots, X^d$ are independent from each other. The domain of k, $[0,1]^d \times [0,1]^d$, is a compact metric space, and thus it is known that $\mathcal{M}(\mathcal{F}, P_\mathbf{Z}, P_\mathbf{U}) = 0$ if and only if $P_\mathbf{Z} = P_\mathbf{U}$. (Gretton et al., 2008). □

H.2. Proof of Lemma 6.

Proof. Using the triangle inequality and the symmetry of the kernel k, $|k(\mathbf{z}_1, \mathbf{z}_2) - k(\mathbf{z}_3, \mathbf{z}_4)| \leq |k(\mathbf{z}_1, \mathbf{z}_2) - k(\mathbf{z}_3, \mathbf{z}_2)| + |k(\mathbf{z}_2, \mathbf{z}_3) - k(\mathbf{z}_4, \mathbf{z}_3)| \leq L\|\mathbf{z}_1 - \mathbf{z}_3\| + L\|\mathbf{z}_2 - \mathbf{z}_4\|$. □

H.3. Proof of Lemma 7.

Proof. The Massart version of Kiefer-Dvoretzky-Wolfowitz theorem (Devroye & Lugosi, 2001) is as follows.

Theorem 16. *Let $X_1, \ldots, X_m$ be an i.i.d. sample from a probability distribution over $\mathbb{R}$ with c.d.f. $F : \mathbb{R} \to [0,1]$, and let the empirical c.d.f. be defined as above $\widehat{F}(x) = \frac{1}{m}|\{i : 1 \leq i \leq m, X_i \leq x\}|$ for $x \in \mathbb{R}$. Then, for any $\epsilon \geq 0$,*

$$\Pr\left[\sup_{x \in \mathbb{R}} |F(x) - \widehat{F}(x)| > \epsilon\right] \leq 2\exp(-2m\epsilon^2).$$

Using $\|\cdot\|_2 \leq \sqrt{d}\|\cdot\|_\infty$ in $\mathbb{R}^d$ and the union-bound we have

$$\begin{aligned}
&\Pr\left[\sup_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_2 > \epsilon\right] \\
&\quad \leq \Pr\left[\sup_{\mathbf{x} \in \mathbb{R}^d} \sqrt{d}\|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_\infty > \epsilon\right] \\
&\quad = \Pr\left[\sup_{x \in \mathbb{R}} \max_{1 \leq j \leq d} |F_j(x) - \widehat{F}_j(x)| > \epsilon/\sqrt{d}\right] \\
&\quad \leq \sum_{i=1}^d \Pr\left[\sup_{x \in \mathbb{R}} |F_j(x) - \widehat{F}_j(x)| > \epsilon/\sqrt{d}\right] \\
&\quad \leq 2d \exp(-\frac{2m\epsilon^2}{d}).
\end{aligned}$$

□

References

Adami, C. Information theory in molecular biology. *Physics of Life Reviews*, 1:3–22, 2004.

Aghagolzadeh, M., Soltanian-Zadeh, H., Araabi, B., and Aghagolzadeh, A. A hierarchical clustering based on mutual information maximization. In *in Proc. of IEEE International Conference on Image Processing*, pp. 277–280, 2007.

Chai, B., Walther, D. B., Beck, D. M., and Fei-Fei, L. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

Chow, C. K. and Liu, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14: 462–467, 1968.

Devroye, L. and Lugosi, G. *Combinatorial Methods in Density Estimation*. Springer, 2001.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two sample problem, 2008. MPI for Biological Cybernetics.

Hall, H. On characterizing dependence in joint distributions, 1969. University of North Carolina. Institute of Statistics Mimeo Series No. 507.

Hero, A. O., Ma, B., Michel, O., and Gorman, J. Alpha-divergence for classification, indexing and retrieval, 2002a. Communications and Signal Processing Laboratory Technical Report CSPL-328.

Hero, A. O., Ma, B., Michel, O., and Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002b.

Hlaváckova-Schindler, K., Paluŝb, M., Vejmelkab, M., and Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

Hulle, M. M. Van. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.

Kybic, J. Incremental updating of nearest neighbor-based high-dimensional entropy estimation. In *Proc. Acoustics, Speech and Signal Processing*, 2006.

Learned-Miller, E. and Fisher, J. W. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

Lewi, J., Butera, R., and Paninski, L. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

Peng, H. and Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 27, 2005.

Póczos, B. and Lőrincz, A. Independent subspace analysis using geodesic spanning trees. In *ICML*, pp. 673–680, 2005.

Rényi, A. On measures of dependence. *Acta. Math. Acad. Sci. Hungar*, 10:441–451, 1959.

Reshef, David N., Reshef, Yakir A., Finucane, Hilary K., Grossman, Sharon R., McVean, Gilean, Turnbaugh, Peter J., Lander, Eric S., Mitzenmacher, Michael, and Sabeti, Pardis C. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011. doi: 10.1126/science.1205438.

Schweizer, B. and Wolff, E. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9, 1981.

Shan, C., Gong, S., and Mcowan, P. W. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

Szabó, Z., Póczos, B., and Lőrincz, A. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.